ENAR

INTERNATIONAL BIOMETRIC SOCIETY

Eastern North American Region

Program & Abstracts

2003

SPRING MEETING

with IMS and Sections of ASA

March 30–April 2, 2003

Marriott, Waterside

Tampa, FL

TABLE of CONTENTS

ACKNOWLEDGEMENTS

EXHIBITORS

American Statistical Association
Biostat
The Cambridge Group, LTD
Cambridge University Press
ClinForce, Inc.
CRC Press
Cytel Software Corporation
Duxbury, Thomson
John Wiley & Sons, Inc.
Oxford Univesity Press
PPD
SAS Institute Inc.
Smith Hanley
Society for Industrial and Applied Mathematics
Springer-Verlag New York, Inc.
StatSoft, Inc.

SPONSORS

We gratefully acknowledge the support of:

AstraZeneca
Amgen, Inc.
Biogen, Inc.
Bristol-Myers Squibb Company
Eli Lilly and Company
Cytel Software Corporation
GlaxoSmithKline
ICON Clinical Research
Johnson & Johnson Pharmaceutical Research & Development
Merck Research Laboratories
Pfizer, Inc.
PPD
SAS Institute Inc.
Schering-Plough Research Institute
Smith Hanley Associates
Statistics Collaborative, Inc.
The Emmes Corporation

The 2003 Local Arrangements Committee would like to thank the Tampa Convention & Visitors Bureau for providing support in planning the meeting & Stefanie Baumgarten for her invaluable assistance.

ARRANGEMENTS

ENAR Program Committee
Chair: Oliver Schabenberger, SAS Institute Inc.
Karen J. Bandeen-Roche, Johns Hopkins Bloomberg School of Public Health
Greg Campbell, U.S. Food and Drug Administration
Mary Christman, University of Maryland
Lisa LaVange, Inspire Pharmaceuticals, Inc.
Robert J. Tempelman, Michigan State University

Education Advisory Committee
Ron Brookmeyer, Johns Hopkins Bloomberg School of Public Health
Timothy G. Gregoire (Chair), Yale University
Marcia Gumpertz, North Carolina State University
Thomas A. Louis, Johns Hopkins Bloomberg School of Public Health
Judith Rich O'Fallon, Mayo Clinic
Oliver Schabenberger, SAS Institute Inc.
Linda J. Young, University of Nebraska

ASA Section Representatives
Vincent C. Arena, University of Pittsburgh
Julia L. Bienias, Rush Institute for Healthy Aging
Montserrat Fuentes, North Carolina State University
Mary Beth Landrum, Harvard Medical School
Stacy R. Lindborg, Lilly Corporate Center
Madhuri Mulekar, University of South Alabama
John Neuhaus, University of California, San Francisco
Alicia Y. Toledano, Brown University

IMS Program Chair
Daniel Scharfstein, Johns Hopkins Bloomberg School of Public Health

IMS Contributed Papers Coordinator
Karl W. Broman, Johns Hopkins Bloomberg School of Public Health

ENAR Local Arrangements
Alan Cantor, Moffitt Cancer Center, University of South Florida

ENAR Student Volunteers
Cynthia Wilson Garvan, University of Florida

Diversity Workshop
Scarlett Bellamy
Dionne Price

ENAR Student Award Committee
Linda J. Young, Chair
Dianne Finkelstein
Kyung Mann Kim
Xihong Lin
Walter Piegorsch
Carol Redmond
Naisyin Wang
Dan Zelterman
Dale Zimmerman

OFFICERS AND COMMITTEES

Executive Committee – Officers

President	Timothy G. Gregoire
Past President	Carol K. Redmond
President-Elect	Marie Davidian
Secretary (2003-2004)	Sharon-Lise Normand
Treasurer (2002-2003)	Ralph L. Kodell

Regional Committee – (RECOM)

President (Chair) Timothy G. Gregoire
Six ordinary members (elected to 3-year terms):
+ Gregory Campbell (RAB Chair)

2001-2003	2002-2004	2003-2005
KyungMann Kim	Julie Legler	Stephen George
Wasima N. Rida	Lance Waller	Roderick Little

Regional Members of the Council of the International Biometric Society

Marie Davidian
Peter Macdonald
Walter W. Piegorsch

Appointed Members of Regional Advisory Board (3-year terms)

Chair: Greg Campbell
Chair-Elect: Stacy Lindborg

2001-2003	2002-2004	2003 - 2005
Mary Kathryn Cowles	Mark Conaway	Scarlett Bellamy
Carol A. Gotway Crawford	Jason Fine	Christopher R. Bilder
Ronald McRoberts	Susan Halabi	DuBois Bowman
Beverly G. Mellen	Elizabeth Margosches	James J. Chen
Linda W. Pickle	Sati Mazumdar	Michael Daniels
Daniel Schnell	Rasmus Nielsen	Francesca Dominici
Robert J. Tempelman	Myunghee Cho Paik	Montserrat Fuentes
Wanzhu Tu	Yuko Palesch	Stacy R. Lindborg
Matthew P. Wand	Dionne Price	Brian D. Marx
Naisyin Wang	Joanna Shih	Alicia Y. Toledano

PROGRAMS

2003 Joint Statistical Meeting
Calvin Williams

2003 Spring Meeting – Tampa, FL
Local Arrangements Chair: Alan Cantor
Program Chair: Oliver Schabenberger

2004 Joint Statistical Meeting
Paul Rathouz

2004 Spring Meeting – Pittsburgh, PA
Local Arrangements Chair: Joyce Chang and Ada Youk
Program Chair: Thomas R. Ten Have

Biometrics Editors
Daniel Commenges, Brian Cullis and Marie Davidian

Biometric Bulletin Editor
Thomas R. Ten Have

ENAR Correspondent for the *Biometric Bulletin*
Mary Christman

ENAR Executive Director
Kathy Hoskins

International Biometric Society Business Manager
Claire Shanley

REPRESENTATIVES

Committee of Presidents of Statistical Societies (COPSS)
ENAR Representatives
Timothy G. Gregoire (President) Carol K. Redmond (Past Pres) Marie Davidian (Pres-Elect)

ENAR Standing/Continuing Committee Chairs

Nominating	Carol K. Redmond
Membership	Lisa A. Weissfeld
Sponsorship	Lisa LaVange
Information Technology Oversight (ITOC)	Stewart Anderson

American Association for the Advancement of Science (Joint with WNAR)
Terms through February 22, 2005

Section N. Medical Sciences	Karen Bandeen-Roche
Section G, Biological Sciences	Zhao-Bang Zeng
Section U, Statistics	Aparna Huzurbazar
Section O, Agriculture	Bahman Shafii

National Institute of Statistical Sciences
(ENAR President is also an ex-officio member) Board of Trustees
Member: Timothy G. Gregoire

Visit the ENAR website (www.enar.org) for the most up to date source of information on ENAR activities.

2003 STUDENT AWARD WINNERS

Van Ryzin Award Winner

Shibao Feng, University of Michigan

Award Winners

Tatiyana Apanasovich, Texas A&M University
Li Chen, North Carolina State University
Inna Chervoneva, Temple University
Haitao Chu, Emory University Rollins School of Public Health
Feng Gao, Emory University Rollins School of Public Health
E. Andres Houseman, Harvard School of Public Health
Abigail Jager, University of Chicago
Yuan Ji, University of Wisconsin-Madison
Brent Johnson, North Carolina State University
Yan Li, Yale University
Yongyi Min, University of Florida
Yuhyun Park, Harvard School of Public Health
Nusrat Rabbee, Harvard School of Public Health
Margaret Short, University of Minnesota
Eric Tassone, Emory University
Abdus S. Wahed, North Carolina State University
Xiaofei Wang, University of North Carolina
Kenneth J. Wilkins, Harvard School of Public Health
Jun Yan, University of Wisconsin–Madison
Guosheng Yin, University of North Carolina at Chapel Hill

FUTURE MEETINGS OF THE INTERNATIONAL BIOMETRIC SOCIETY

2004 International Biometric Conference
Cairns, Australia

2004 Spring Meeting
Pittsburgh, PA

2005 Spring Meeting
Austin, TX

SPECIAL THANKS

2003 ENAR Program Committee

Oliver Schabenberger (Chair), SAS Institute Inc.
Karen J. Bandeen-Roche, Johns Hopkins Bloomberg School of Public Health
Greg Campbell, U.S. Food and Drug Administration
Mary Christman, University of Maryland
Lisa LaVange, Inspire Pharmaceuticals, Inc.
Robert J. Tempelman, Michigan State University

ASA Section Representatives

Vincent C. Arena, University of Pittsburgh
Julia L. Bienias, Rush Institute for Healthy Aging
Montserrat Fuentes, North Carolina State University
Mary Beth Landrum, Harvard Medical School
Stacy R. Lindborg, Lilly Corporate Center
Madhuri Mulekar, University of South Alabama
John Neuhaus, University of California, San Francisco
Alicia Y. Toledano, Brown University

ENAR Education Advisory Committee

Ron Brookmeyer, Johns Hopkins Bloomberg School of Public Health
Timothy G. Gregoire (Chair), Yale University
Marcia Gumpertz, North Carolina State University
Thomas A. Louis, Johns Hopkins Bloomberg School of Public Health
Judith Rich O'Fallon, Mayo Clinic
Oliver Schabenberger, SAS Institute Inc.
Linda J. Young, University of Nebraska

ENAR Local Arrangements

Alan Cantor, Moffitt Cancer Center, University of South Florida

IMS Program Chair

Daniel Scharfstein, Johns Hopkins Bloomberg School of Public Health

ENAR Student Awards

Linda J. Young, University of Nebraska

ENAR Student Volunteers

Cynthia Wilson Garvan, University of Florida

ENAR SHORT COURSES

SC1: Statistical Methods and Software for the Analysis of DNA Microarray Experiments (half-day) SALONS G-H

Instructors: Sandrine Dudoit (University of California, C Berkeley) and Rafael Irizarry (Johns Hopkins Bloomberg School of Public Health)

Description:
cDNA microarrays and high-density oligonucleotide chips are novel biotechnologies which allow the monitoring of expression levels in cells for thousands of genes simultaneously. Microarrays are being applied increasingly in biological and medical research to address a wide range of problems (e.g. to study the molecular variations among tumors in cancer research, or to study the gene expression response of model organisms such as yeast to different types of stress conditions). Gene expression experiments generate large and complex multivariate datasets. The application of sound statistical design and analysis principles can greatly improve the efficiency and reliability of microarray experiments throughout the data acquisition and analysis process.

This short course will survey statistical and computational issues arising in the design and analysis of DNA microarray experiments. After a brief introduction to genome biology, the following topics will be addressed: experimental design, image analysis, normalization, the identification of differentially expressed genes, pattern discovery and recognition for genes and mRNA samples. Access to an efficient, portable, and distributed statistical computing environment is an essential aspect of the analysis of gene expression data. Statistical computing resources developed as part of the Bioconductor project will also be discussed. This collaborative effort aims more generally to produce an open source and open development computing environment for biologists and statisticians working in bioinformatics (**http://www.bioconductor.org**).

Prerequisite:

- Familiarity with the S language (S, S-plus, R)

DATE: Sunday, March 30, 2003

Full Day Fee

Members	$190 ($215 after 2/20)
Nonmembers	$215 ($240 after 2/20)

Half Day Fee

Members	$110 ($135 after 2/20)
Nonmembers	$135 ($160 after 2/20)

Short Course Registration

Saturday, March 29	3:00 - 5:00 p.m.
Sunday, March 30	7:00 - 8:30 a.m.

Course

Sunday, March 30 *(lunch on your own)*	8:30 a.m. - 5:00 p.m. (full day)
Sunday, March 30	1:00 - 5:00 p.m. (half day)

SC2: Design and Analysis of Studies of Diagnostic and Screening Tests (full-day) SALON C

Instructors: Alicia Toledano, Jeffrey Blume, and Constantine Gatsonis (all at Brown University)

Description:
This course will present a comprehensive survey of the state of the art in evaluating the performance of diagnostic and screening tests. Course participants will learn the key considerations in designing a study of a diagnostic or screening test and will become familiar with statistical methods used to measure and compare test accuracy, especially receiver operating characteristic (ROC) curve methodology. Methods for performing sample size calculations in simple studies through complex studies involving multiple test modalities and multiple test interpreters will be reviewed. Research synthesis of diagnostic and screening test accuracy evaluations will be discussed. Examples, including studies from the American College of Radiology Imaging Network (ACRIN), will be used to illustrate the methods described. Review of available software will be integrated throughout. The course will conclude with a discussion of screening that moves beyond accuracy to incorporate additional statistical, epidemiological, health policy, and public policy considerations.

Prerequisite:

- Knowledge of methods for analyzing binary data.

SC3: Smoothing for Smarties (full-day) SALON D

Instructors: Paul Eilers (Leiden University, The Netherlands) and Brian Marx (Louisiana State University)

Description:
Smoothing has become an established part of the statistical toolbox and it is gradually finding its way into most of the statistical software packages. There exist a wide choice of methods, like smoothing splines, regression splines, kernel smoothing and local likelihood. It is difficult for the novice to make an informed choice.

The main theme of our course is the use of a combination of local basis functions (B-splines) and roughness penalties: the P-spline approach. This smoother stays as close as possible to established regression methodology, with many advantages:
- compact computations;
- compact results (a small set of coefficients);
- easily computed diagnostics: cross-validation or AIC, standard errors, effective degrees of freedom;
- parametric models are obtained as limiting cases of heavy smoothing;
- straightforward extension to non-normal data, along the lines of the generalized linear model;
- P-splines are ideal building blocks for generalized additive models, varying-coefficient models and signal regression.

We present the basic theory and illustrate it with many practical examples: scatterplot smoothing, trend lines, density estimation, dose-response curves, and smooth high dimensional coefficient vectors.

Software, example data and demonstration scripts will be available for S-Plus (R freeware) and Matlab. Their use will be illustrated during the course. After taking this course one will be able to go home to analyze his/her own data.

Hands-on work on installed computers is not planned, but a possible alternative is to provide the attendants with software and data beforehand, to be installed on their portable PCs.

Prerequisite:

- Multiple regression and familiarity with basic matrix operations.

SC4: Longitudinal Data Analysis (full-day) SALONS I-J

Instructors: Geert Verbeke (Biostatistical Center of the Katholieke Universiteit Leuven, Belgium) and Geert Molenberghs (Limburgs Universitair Centrum, Belgium)

The course consists of three main parts: continuous longitudinal data, discrete longitudinal data, and incomplete data.

Based on Verbeke and Molenberghs (Springer, 1997, 2000), and the forthcoming Molenberghs and Verbeke (2004), a general introduction to longitudinal data and the linear mixed model for continuous responses will be presented. The topic will be approached from the modeller's and practitioner's points of view. Emphasis will be on model formulation, parameter estimation, and hypothesis testing, as well as on the distinction between the random-effects (hierarchical) model and the implied marginal model. Apart from classical model building strategies, many of which have been implemented in standard statistical software, a number of flexible extensions and additional tools for model diagnosis will be indicated. Illustrations will be given based on the SAS procedure MIXED. The course will be explanatory rather than mathematically rigorous. Emphasis is on giving sufficient detail in order for participants to have a general overview of frequently used approaches, with their advantages and disadvantages, while giving reference to other sources where more detailed information is available. Also, it will be explained in detail how the different approaches can be implemented in SAS and how to interpret the SAS output.

When the response of interest is categorical, a number of modeling options are open. For example, the linear mixed model concepts can be extended towards generalized linear mixed models. An important alternative approach is the use of generalized estimating equations (GEE). A lot of emphasis will be put on the fact that the regression parameters in both types of models have different interpretations. Advantages and disadvantages of both procedures will be discussed and compared in detail, and illustrations will be based on the SAS procedures GENMOD and NLMIXED. Some other approaches will be sketched briefly.

Finally, when analyzing longitudinal data, one is often confronted with missing observations, i.e., scheduled measurements have not been made, due to a variety of (known or unknown) reasons. It will be shown that, if no appropriate measures are taken, missing data can seriously bias results, and cause great difficulty to appropriate interpretation.

Session 1:
- Introduction: Examples of longitudinal studies; Cross-sectional versus longitudinal studies; Merits of longitudinal studies
- Exploring longitudinal data: Graphical exploration; Exploring the mean structure; Exploring the covariance structure
- Introduction of linear mixed models: two-stage modeling approach; the general linear mixed model
- Inference for marginal model: ML and REML, robust inference for mean, inference for covariance

Session 2:
- Inference for random effects: Empirical Bayes estimates; advantages and disadvantages; interpretation
- Guidelines for model construction and model diagnosis
- Some model extensions

Session 3:
- Generalized linear mixed models: Model introduction; Inference; Interpretation
- Generalized estimating equations (GEE): Model introduction; Inference; Interpretation

Session 4:
- Missing values in longitudinal data: Examples
- The impact on interpretation, efficiency and inference
- Some simple, naive solutions
- Introduction to selection models, pattern-mixture models, and sensitivity analysis

Target audience: applied statisticians and biomedical researchers in industry, public health organizations, contract research organizations, and academia.

Prerequisite:

- Working knowledge of linear regression and ANOVA is assumed. The course is accessible for both MSc and Ph.D. level statisticians.

Learning outcomes and instructional methods: as a result of the course, participants should be able to perform a basic analysis for a particular longitudinal data set at hand. Based on a selection of exploratory tools, the nature of the data, and the research questions to be answered in the analyses, they should be able to construct an appropriate statistical model, to fit the model within the SAS framework, and to interpret the obtained results. Further, participants should be aware not only of the possibilities and strengths of a particular selected approach, but also of its drawbacks in comparison to other methods.

Course materials will include copies of the transparencies used in the course. Optionally available from the Springer-Verlag exhibitor will be the authors' textbook: Verbeke G. & Molenberghs G. (2000) *Linear mixed models for longitu-*

SC5: Introduction to Geographic Information Systems & Spatial Analysis for Environmental and Public Health Studies

SALON A

(Split-Day: SC5A–Morning SC5B–Afternoon)

Instructors: Lance A. Waller (Emory University) and Carol A. Gotway Crawford (Centers for Disease Control and Prevention)

STRUCTURE: Morning and afternoon sessions are self-contained so individuals can register for either session singly, or for both sessions together.

Introduction to Geographic Information Systems (Morning course – 8:30 a.m-12:00 p.m.)

1. What is GIS?
2. Spatial data (location, attribute)
3. Types of location
   a. Vector (point, line, area) b. Raster (remote sensing)
4. Where to get spatial data?
   a. Geocoding b. Public data (metadata)
5. Basic graphical display in GIS
   a. Types of maps
      i. Point maps ii. Choropleth maps iii. Flow maps
   b. Basic symbolization
      i. Shading ii. Color
6. Basic GIS operations:
   a. Layering
   b. Buffering
   c. Spatial Query
   d. Value-added to existing data
7. Basic statistics-what statistics can a GIS do now?
   a. Point (count) in polygon
   b. Summaries
8. Examples
   a. Environmental Justice
   b. Sea Turtle Nesting (maybe)
   c. Cardiovascular disease mortality (health disparities/ small area analysis)
9. Statistical shortcomings (preview of afternoon)

GIS and Spatial Data Analysis (Afternoon Course – 1:00 p.m. - 5:00 p.m.)

1. Brief review of basic GIS definitions and operations (for afternoon-only attendees).
2. Types of spatial data/analyses (data types and questions of interest)
   a. Geostatistical
   b. Point
   c. Lattice
3. Computing: Statistical computation in GIS, GIS computation instatistical packages:
   a. Geostatistical Analyst in ArcView
   b. SAS, spatial analysis and PROCGIS
   c. S+SpatialStats
   d. S+/ArcView link
   e. SpaceStat
   f. R/GRASS
   g. Others (Omegahat, Orca)
4. Geostatistics/kriging/ArcView + Geostat Analyst
   a. Variogram estimation
   b. Spatial prediction
   c. Data break: raccoon rabies in Connecticut
5. Point processes
   a. Limited functionality in GIS (now)
   b. Exploring .csv data to R/S+
   c. Splancs library d. intensity estimation
   e. K function f. Data breaks: graves, nerve fibers
6. Lattice data
   a. small area estimation/clustering tests
   b. Limited functionality in GIS (now)
   c. port to winBUGS/geoBUGS
   d. S+SpatialStats
   e. Spacestat
7. e-handouts (website with example data and analyses (most from book).

Prerequisite:

- This short course will assume prior training such as that provided by a Master's level course in multiple linear regression and familiarity with the basic concepts of logistic and Poisson regression.

ENAR 2003 TUTORIALS

T1: (Bio)Equivalence Testing

FLORIDA SALON VI

Date: Monday, March 31 **Time: 8:30-10:15 a.m.**
Instructor: Roger L. Berger (North Carolina State University)

Description:
In this tutorial the rationale and implementation of equivalence tests will be discussed. The use of intersection-union methodology to construct valid equivalence tests, even in complicated situations like comparison of dissolution curves, will be explained. On the other hand, the pitfalls in incorrectly generalizing methods that work in simple situations to more complicated situations will be discussed. The broad applicability of equivalence tests will be illustrated with examples from ecology and land reclamation as well as the more common pharmaceutical applications.

Prerequisite:

- The tutorial will assume prior training such as that provided by a Master's level course in applied statistics.

T2: Dimension Reduction and Graphics in Regression

FLORIDA SALON VI

Date: Monday, March 31 **Time: 1:45-3:30 p.m.**
Instructor: Dennis Cook (University of Minnesota)

Description:
In simple regression a 2D plot of the response versus the predictor displays all the sample information, and can be quite helpful for gaining insights about the data and, if appropriate, for guiding the choice of a first model. Analogous displays of all the data are not possible with many predictors, but fully informative displays are possible in situations where we can find a low-dimensional sufficient summary view that contains or is inferred to contain all the relevant sample information. In this context, regression is defined quite broadly as the study of the conditional distribution of the response given the predictors without prespecifying a parametric model. Sufficient summary views can be invaluable for guiding the choice of a first model and for gaining insights about the regression, as 2D plots are in simple regression. Seemingly complicated regressions can often be summarized adequately in a relatively simple summary plot.

Beginning with a series of illustrations, foundations of sufficient summary views and methods for constructing them will be reviewed. The emphasis will be on regression, but the ideas and some of the methods are equally applicable in other areas like classification and discrimination. Background information is available at *www.stat.umn.edu/RegGraph*. Most of the methods to be discussed are available in the computer program Arc that can be obtained at *www.stat.umn.edu/arc.*

T3: MCMC: How It Works; When It Fails

FLORIDA SALON VI

Date: Tuesday, April 1 **Time: 8:30-10:15 a.m.**
Instructors: Mary Kathryn ("Kate") Cowles and Matt Bognar (University of Iowa)

Description:
The target audience for this tutorial is people who use WinBUGS software or other MCMC methods to fit Bayesian models and who would like to know a bit about what is going on "under the hood" of an MCMC sampler. We will review the workings of the Gibbs sampler, Metropolis-Hastings, and slice-sampling algorithms, showing exactly what they do in the context of some simple models for real data. We will also consider some modeling situatians in which MCMC methods can fail, and how a user can avoid these.

Prerequisite:

- An introductory course (or equivalent experience) in applied Bayesian statistics, including the use of WinBUGS or other MCMC-based software for fitting Bayesian models.

T4: Sample Survey Methods for Biostatisticians FLORIDA SALON VI

Date: Tuesday, April 1 **Time: 1:45-3:30 p.m.**
Instructors: Barry Graubard (National Cancer Institute) and Lisa LaVange (Inspire Pharmaceuticals Inc.)

Description:
Biostatisticians today find themselves often confronted with analyzing data from sample surveys while not extensively trained in that area. The availability of databases from a variety of sources, including large-scale national surveys, along with a greater realization in the value of population-based inference has contributed to this phenomenon. This tutorial will provide a broad overview of sample surveys, focusing on how the complexities of the sample design and weighting adjustments affect data analyses. Methods for conducting descriptive and regression analyses will be illustrated, and topics currently being debated in a variety of quarters, including design-based vs. model-based inference or inference about superpopulation parameters, will be touched upon, with references for further reading. Applications of sample survey methods to non-survey settings, such as inference from studies with correlated observations or weighted data where weights may result from missingness or standardization will also be discussed. Examples of some common data analysis problems will be provided using data from several well-known U.S. surveys. Finally, we will discuss the use of software for analyzing sample survey data. This workshop will be targeted to statistical researchers who are interested in using survey data to answer questions in epidemiology and public health. Although applications will not directed to agriculture, wildlife and natural resources, researchers in these areas should also find the methods discussed in this workshop applicable to their problems.

Prerequisite:

- A working knowledge of linear, logistic, and proportional hazards regression is required.

ENAR 2003 ROUNDTABLES

DATE: Monday, March 31, 2003
12:15 – 1:30 p.m.

Fee: $30 Location: Florida Salon IV

Space is limited. Preregistration is advised.

R1: Grant Writing

Discussion Leader: Lynne Billard, University of Georgia (1994-95 IBS President)

Grants are a fundamental part of our careers. The grant writing process will be described. This includes discussion of the types of grants available, how to access grants details, guidelines for writing grants, and pointers for success in securing funding from granting agencies.

R2: Statistical Leadership Within the Pharmaceutical Industry: Current Area of Influence and Beyond

Discussion Leader: Stacy R. Lindborg, Eli Lilly & Co.

The majority of statisticians hired into the pharmaceutical industry are employed in Phase III/IV research. Statisticians are brought in due to their technical expertise but often fill much of their time with tasks, which don't require advanced training (e.g., powering studies, basic analyses). Because of our training and understanding of data, statisticians bring an important voice to many different forums. We will consider areas in which statisticians currently have influence as well as areas in which statisticians should be leaders along with avenues for creating greater impact.

R3: Statistical Journals: Publishing Without Perishing

Discussion Leaders: Marie Davidian, North Carolina State University, and Xihong Lin, University of Michigan, Former and Current Coordinating Editors, Biometrics.

Success in publishing in statistical journals is an important evaluation criterion for both junior- and senior-level researchers. Moreover, publication in statistical journals represents the primary mode of dissemination of new results to our profession. Yet many researchers still feel uncertain about publishing in general and the review and editorial processes in particular. What practices lead to success in publishing? What qualities do editors, associate editors, and reviewers look for in a journal article? How should an author choose an appropriate outlet for his or her work? An issue that affects all authors is the fact that times to review for our journals lag well behind those in other disciplines. What are the reasons for this and what can be done to improve the situation? What do editors and associate editors expect from referees? What should authors expect from the editorial process? This roundtable will serve as a forum to discuss these and other relevant issues.

R4: Issues in Regulatory Statistics

Discussion Leader: Boguang Zhen, FDA Office of Biostatistics and Epidemiology

This roundtable will discuss issues in regulatory statistics including Intention-to-treat, equivalence, lot consistency, submissions requirements and other topics the participants find of interest.

R5: Preparing Students for Statistical Practice

Discussion Leader: Ramon C. Littell, University of Florida

How to prepare students for statistical practice is a well-worn topic, but one for which there is not uniform agreement. Programs at universities differ widely. Participants will discuss experiences with consulting courses, statistical laboratories, and industry internships as training vehicles. Prospective employers are especially invited to provide input on their expectations new employees.

R6: Handling of Incomplete Data in Longitudinal Studies

Discussion Leader: Geert Molenberghs, Limburgs Universitair Centrum, Belgium

Most, if not all, clinical and non-clinical longitudinal studies are prone to incompleteness. In a biopharmaceutical context, a lot of relatively simple methods for dealing with such incomplete data are used. On the other hand, ever more advanced methods for handling incomplete longitudinal studies are proposed from an academic view-point. Is it necessary and possible to think about ways of bridging this gap?

ENAR 2003 ROUNDTABLES

R7: Longitudinal Study, Panel Data or Structural Equation: How Do We Read the Map at the Crossroads of Statistical Traditions?

Discussion Leader: Mari Palta, University of Wisconsin

It is not uncommon to discover that "new" methodological approaches developed in Biostatistics have been well known in Econometrics or in Psychometrics for some time— and vice versa. Differences in terminology, and the huge amount of literature in each of these disciplines make it difficult to assess overlap and be aware of developments outside one's own tradition. We will discuss how important it is to keep track of such developments, including considerations of how differences in subject matter application may influence modeling assumptions. and hence the choice of approach. We will also address whether it is possible to make the task of cross-fertilization easier, for example, by making sure that alternative terminologies are taught in statistics courses.

R8: Current Issues in Environmental Health Research

Discussion Leader: Louise Ryan, Harvard School of Public Health

Recent advances in environmental health research have led to a number of interesting statistical challenges. In this roundtable, we will discuss some of the open statistical research topics motivated by environmental problems, including the use of biomarkers, spatial statistics, multilevel modeling and genomics.

R9: What Should a Modern Statistics Degree Program Look Like?

Discussion Leader: Walter W. Stroup, University of Nebraska

Several talks at JSM (this year & other years) take our profession to task for antique degree curricula. Depending on whose talk you listen to, we need more modeling, more data processing, more this & that (what to give up is often not addressed, at least not very satisfyingly). Others say we will never have an identity as a discipline until we offer undergraduate majors as a rule rather than an exception. Several speakers talked about the need to establish statistics as more than "a branch of mathematics." Several used the analogy to physics — you can't do physics without math, but physics is not math. My question is, as I design a new curriculum from the ground up (a luxury we have at Nebraska this year because of our unique set of circumstances) what should it look like? If one breaks the mold, how does one make the program credible to other institutions, employers, colleagues in other departments that use our service courses, etc.

R10: Data Monitoring Committees: What Happens When Unexpected Harm Emerges?

Discussion Leader: Janet Wittes, Statistics Collaborative

Data monitoring committees (DMC) for clinical trials often have clearly defined monitoring boundaries for efficacy. Safety, however, poses a less well defined problem. Confronted with an unexpected excess of adverse events in the treated group relative to the control group, a DMC will often not have guidelines to help assess the statistical soundness of the finding. At the roundtable, we will discuss specific examples of trials that have experienced important excess in a clinically important adverse event. We will also explore approaches to formalizing the process by which a DMC should operate in these situations.

R11: Issues in Design of Studies Evaluating Diagnostic and Screening Tests

Discussion Leader: Alicia Y. Toledano, Brown University

Statisticians are increasingly asked to help design and analyze studies evaluating the accuracy of diagnostic and screening tests. We will discuss three key differences in these studies *versus* studies of therapies: issues in establishing truth for each study subject; the use of each subject as her/his own control; and the framework of using random effects to model variability in test results that is inevitable when medical personnel interpret tests (e.g., medical images such as ultrasound and PET scan, psychiatric interviews, or tissue specimens).

The NIH Statistical Grant Review and Funding Opportunities

5:20-7:00 pm, Sunday, March 30, 2003

Organizer and Chair: Xihong Lin, University of Michigan

5:20-5:55 pm The Grant Review Process at NIH
Ann Hardy, Center for Scientific Review, NIH

Dr. Hardy will provide a general overview of the process used at NIH to review investigator-initiated grants. The various phases of the process will be described including pre-submission, assignment of grants to a study section, review, and post-review. The composition and function of study sections and what happens during a study section meeting will be detailed. Common problems encountered with applications will be reviewed and human subjects requirements for grant applications will be presented. Some tips on where to find information about programmatic interests, various grant mechanisms, and the grant review process will be provided.

5:55-6:30 pm A Review of the NCI's Statistics Grant Portfolio
Ram Tiwari, National Cancer Institute

Dr. Tiwari will provide an overview of the NCI statistical grant portfolio. A large number of NIH statistical methodological grants are funded by NCI through the Division of Cancer Control and Population Sciences (DCCPS). The DCCPS recently released a web site: www.statfund.cancer.gov to provide information about statistical grant applications and the NCI statistical grant portfolio for the extramural statistical community. It recently also made a decision that all new statistical methodological grants in the Division would be assigned to the Statistical Research and Applications Branch. Dr. Tiwari will review the impact of this decision on the NCI Statistical Grant Portfolio, present the current status of the Portfolio. Dr. Tiwari will describe his role and responsibilities as the NCI program director managing the statistical grant portfolio, and will seek suggestions that could lead in improving the web site and in better management of the Portfolio.

6:30-7:00 pm The NIH Statistical Grant Review and the SNEM-5 Study Section
Louise Ryan, Harvard School of Public Health

Professor Ryan will provide an overview of the SNEM-5 study section, the study section dealing with proposals in statistical methodology. She will describe the nature and structure of the study section, the criteria that the study section uses to evaluate grant applications. Common problems encountered in applications such as budget issues, human subject participation, etc. will be discussed.

There will be ample time to address questions from the audience throughout the Forum.

ENAR

PROGRAM SUMMARY

Saturday, March 29

3:00 p.m.–5:00 p.m. Conference Registration
(Convention Registration Desk)

Sunday, March 30

7:30 a.m.–6:30 p.m. Conference Registration
(Convention Registration Desk)

8:30 a.m.–12:00 p.m. Short Courses
SC5A: Introduction to Geographic Information Systems
(Salon A)

8:30 a.m.–5:00 p.m. Short Courses
SC2: Design and Analysis of Studies of Diagnostic and Screening Tests
(Salon C)
SC3: Smoothing for Smarties
(Salon D)
SC4: Longitudinal Data Analysis
(Salon I-J)
SC5: Introduction to Geographic Information Systems & Spatial Analysis for Environmental and Public Health Studies
(Salon A)
SC5A: 8:30 a.m.–12:00 p.m., SC5B: 1:00–5:00 p.m.

11:00 a.m.–4:00 p.m. Diversity Workshop
(Meeting Room 10))

1:00 p.m.–5:00 p.m. Short Courses
SC1: Statistical Methods and Software for the Analysis of DNA Microarray Experiments
(Salon G-H)
SC5B: GIS and Spatial Data Analysis
(Salon A)

3:00 p.m.–5:00 p.m. Exhibits Open
(Grand Ballroom Prefunction Area)

4:00 p.m.–7:00 p.m. ENAR Executive Committee
(Greco Boardroom) (closed)

4:30 p.m.–6:30 p.m. Placement Service
(Meeting Room 11)

5:20 p.m.–7:05 p.m. Forum Presentation
NIH Statistical Grant Applications, Reviews and Funding Opportunities
(Salon B)

8:00 p.m.–11:00 p.m. Social Mixer and Poster Session
(Florida Salons IV, V, and VI)

Monday, March 31

7:30 a.m.–8:30 a.m. Student Breakfast
(Meeting Room 10)

7:30 a.m.–5:00 p.m. Conference Registration
(Convention Registraion Desk)

9:00 a.m.–5:00 p.m. Placement Service
(Meeting Room 11)

8:30 a.m.–5:30 p.m. Exhibits Open
(Grandball Room Prefunction Area)

8:30 a.m.–10:15 a.m. Tutorial
T1: (Bio)Equivalence Testing
(Florida Salon VI)
Scientific Program
1. Linking Models to Design and Inference in Sample Surveys
(Salon D)
2. Spatio-temporal Modeling of Environmental Processes
(Salon F)
3. IMS Medallion Lectures: Applications of High-Dimensional Data Analyses to Microarray Data
(Salon E)
4. Special Contributed Session: Advances in Growth Mixture Modeling for Preventive and Treatment Trials
(Salon A)
5. Contributed Papers: HMDIN: "How Many Do I Need" or the Secrets of Sample Size Determination
(Salon B)
6. Contributed Papers: Advances in Mixed Models
(Salon J)
7. Contributed Papers: Estimating Equations
(Salon I)
8. Contributed Papers: Recurrent Event Data, Frailty Models, and Survival Analysis
(Salon C)
9. Contributed Papers: Topics in Screening and Diagnostic Tests
(Salon H)

10:15 a.m.–10:30 a.m. Break
(Grandball Room Pre-Function Area)

10:30 a.m.–12:15 p.m. Scientific Program
10. Effects of Technological Advances on Introductory Statistics Courses
(Salon D)
11. Recent Developments in Clustering and Mixtures with Application to Spatial Models and Image Analysis
(Salon G)
12. Statistics in Genetics
(Salon E)

PROGRAM SUMMARY

13. Combining Disparate Environmental Data
(Salon F)
14. Contributed Papers: Bayesian Methods and Applications
(Salon A)
15. IMS Contributed Papers: Topics in Mathematical Statistics
(Salon B)
16. Contributed Papers: Hazard Modeling and Estimation
(Salon J)
17. Contributed Papers: Censored, Incomplete, and Missing Data
(Salon I)
18. Contributed Papers: Statistical Methods in the Study of HIV and AIDS
(Salon C)
19. Contributed Papers: Design Issues in Clinical Trials and Epidemiology
(Salon H)

12:15 p.m.–1:30 p.m. Roundtable Luncheons
(Florida Salon IV)

12:30 p.m.–4:30 p.m. Regional Advisory Board (RAB)
(Florida Salons II &III) Luncheon Meeting
(By Invitation Only)

1:45 p.m.–3:30 p.m. Tutorial
T2: Dimension Reduction and Graphics in Regression
(Florida Salon VI)
Scientific Program
20. Placebo-Controlled Trials: Ethics, Science and Economics
(Salon C)
21. Statistics and Counter-Bioterrorism
(Salon E)
22. Innovative Approaches and Challenges in the Analysis of Recurrent Failure Time Data
(Salon D)
23. Bayesian Hierarchical Spatial Temporal Methods
(Salon G)
24. Statistical Design and Analysis of Large High Throughput Screening Data for Drug Discovery
(Salon F)
25. Contributed Papers: Analysis of Microarrays and Gene Expression I
(Salon A)
26. Contributed Papers: Binary and Binomial Data
(Salon J)
27. Contributed Papers: Bioequivalence and Non-Inferiority
(Salon I)
28. Contributed Papers: Nonparametric and Exact Methods
(Salon H)
29. Contributed Papers: Statistical Inference for Correlated Data
(Salon B)

3:30 p.m.–3:45 p.m. Break
(Grand Ballroom Pre-function Area)

3:45 p.m.–5:30 p.m. Scientific Program
30. Statistical Methods for Studying Placebo Response
(Salon F)
31. Recent Developments in Bayesian Non/Semi-parametrics
(Salon G)
32. Spatial Surveillance for Incident Clusters and Biological Terrorism
(Salon E)
33. Emerging Exact Methods in Biomedical Research
(Salon I)
34. High-dimensional Hypothesis Testing and Model Selection for Functional Data Analysis
(Salon B)
35. Contributed Papers: Analysis of Linkage and Relationship
(Salon A)
36. Contributed Papers: Statistical Methods for Longitudinal Data
(Salon C)
37. Contributed Papers: Survey Sampling: Theory and Applications
(Salon D)
38. Contributed Papers: Semi- and Nonparametric Methods in Survival Analysis
(Salon H)

6:30 p.m.–7:30 p.m. President's Reception
(Florida Salons II & III) (By Invitation Only)

TUESDAY, APRIL 1

7:30 a.m.–5:00 p.m. Conference Registration
(Convention Registraion Desk)

9:00 a.m.–5:00 p.m. Placement Service
(Meeting Room 11)

8:30 a.m.–5:30 p.m. Exhibits Open
(Grand Ballroom Pre-Function Area)

8:30 a.m.–10:15 a.m. Tutorial
T3: MCMC: How it Works, When It Fails
(Florida Salon VI)
Scientific Program
39. Sampling Methods for Selecting Population Controls
(Salon D)

PROGRAM SUMMARY

40. Multi-stage Decisions and Dynamic Treatment Regimes
(Salon C)
41. Random Effects, Priors and Graphs: A 360 Degree View of Hierarchical Models
(Salon G)
42. Health Effects and Environmental Risk Assessment of Air Pollution
(Salon H)
43. Contributed Papers: Applications in Statistical Genetics
(Salon A)
44. Contributed Papers: Clinical Trials: Adaptive, Group Sequential Designs, and Interim Analysis
(Salon I)
45. Contributed Papers: Spatial Data: Statistical Inference and Modeling
(Salon J)
46. Contributed Papers: Applications of Longitudinal Data Analysis
(Salon B)

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
(Salons E & F)

12:30 p.m.–4:30 p.m. Regional Committee (RECOM)
(Meeting Room 12) Luncheon Meeting
(By Invitation Only)

1:45 p.m.–3:30 p.m. Tutorial
T4: Sample Survey Methods for Biostatisticians
(Florida Salon VI)
Scientific Program
47. Network Design
(Salon D)
48. Statistical Issues in Serial Analysis of Gene Expression (SAGE)
(Salon E)
49. Functional Analysis of Longitudinal Data
(Salon B)
50. The Statistics and Politics of Recent Screening Controversies
(Salon F)
51. Analysis of Adverse Events Reports
(Salon A)
52. Contributed Papers: Generalized Linear and Nonlinear Mixed Models
(Salon C)
53. Contributed Papers: Stochastic Processes: Spatial, Temporal, and Spatio-Temporal
(Salon G)
54. Contributed Papers: Latent Variable Modeling
(Salon H)
55. Contributed Papers: Survival Analysis and Clinical Trials
(Salon I)

3:30 p.m.–3:45 p.m. Break
(Grand Ballroom Pre-Function Area)

3:45 p.m.–5:30 p.m. Scientific Program
56. Practical Issues in Nonlinear Mixed Models
(Salon D)
57. The Measurement of Health Disparities
(Salon G)
58. GEE and GMM: A Bridge between Biostatistics & Econometrics
(Salon E)
59. Current Research in Sample Size Re-Estimation
(Salon C)
60. Emerging Study Designs in Intervention Studies and Clinical Trials: Challenges and Controversy
(Salon F)
61. Contributed Papers: Selection Bias, Verification Bias, and Propensity Scores
(Salon H)
62. Contributed Papers: Estimation and Modeling in the Presence of Missing Data and Nonresponse
(Salon B)
63. Contributed Papers: Analysis of Microarrays and Gene Expression II
(Salon A)
64. Contributed Papers: Image Analysis: MRI, PET and Proteomics
(Salon I)
65. Contributed Papers: Inference on Risks and Rates
(Salon J)

5:30 p.m.–6:30 p.m. ENAR Business Meeting
(Salon A) (Open to all members)

6:30 p.m.–9:30 p.m. Tuesday Night Event
(Starship Dining Yacht dinner cruise)

PROGRAM SUMMARY

WEDNESDAY, APRIL 2

7:30 a.m.–9:00 a.m. Planning Committee (closed)
(Greco Boardroom)

8:00 a.m.–12:30 p.m. Conference Registration
(Convention Registraion Desk)

8:00 a.m.–12:00 p.m. Exhibits Open
(Grand Ballroom Pre-Function Area)

8:30 a.m.–10:15 a.m. Scientific Program
66. Causal Inference at the Biostatistics-Econometrics Interface
(Salon E)
67. Multivariate Mixed Effects Modeling of Discrete and Continuous Outcomes
(Salon F)
68. Special Contributed Session: Recent Methods to Reduce the Impact of Nuisance Parameters
(Salon G)
69. Contributed Papers: Survey Research Methods
(Salon D)
70. Contributed Papers: The Study of Health Effects and Health Care
(Salon C)
71. Contributed Papers: Probability, Distribution Theory and Inference
(Salon H)
72. Contributed Papers: Overdispersed Data and Hierarchical Models
(Salon B)
73. Contributed Papers: Analysis of Gene Expression
(Salon A)

10:15 a.m.–10:30 a.m. Break
(Grand Ballroom Pre-Function Area)

10:30 a.m.–12:15 p.m. Scientific Program
74. Genome-Scale Biology: Statistical Challenges and Solutions
(Salon F)
75. Phase 2/3 Combination Designs to Accelerate Drug Development
(Salon D)
76. Competing Risk Analysis
(Salon C)
77. Statistical Aspects of Polygraphy
(Salon E)
78. Contributed Papers: Semi- and Nonparametric Methods for Longitudinal Data
(Salon B)
79. Contributed Papers: Receiver Operating Characteristics and Screening
(Salon G)
80. Contributed Papers: Causal Inference and Missing Data
(Salon H)
81. Contributed Papers: Random Effects and Measurement Error
(Salon I)
82. Contributed Papers: Analysis of Microarrays and Gene Expression III
(Salon A)
83. Contributed Papers: Generalized Linear Models and Regression
(Salon J)

DIVERSITY WORKSHOP

On Sunday, March 30th, there will be a workshop entitled "Fostering Diversity in Biostatistics". The workshop will provide a forum for discussion of important issues related to diversity. Themes of the workshop will include career and training opportunities within biostatistics/statistics/biometry as well as mentoring, recruiting, and retaining students in (bio)statistical programs, with an emphasis on traditionally underrepresented minority groups. Anyone (current and undergraduate institutions as well as statistical representatives from industry and governmental agencies) interested in addressing diversity issues is welcome to participate in the Workshop, which should be uniquely beneficial to all attendees, regardless of their professional experience. For further details, please contact **Scarlett Bellamy** at *sbellamy@cceb.upenn.edu* or **Dionne Price** at *priced@cder.fda.gov.* **Diversity Workshop registration via** *http://cceb.med.upenn.edu/diversity/*
(Meeting Room 10)

SCIENTIFIC PROGRAM

Times may change slightly prior to the meetings. Please check the errata sheet for possible changes to sessions. Asterisks (*) indicate paper presenters. Student Travel Award Winner presentations appear in **boldface.**

Monday, March 31
8:30–10:15 a.m.

1. Linking Models to Design and Inference in Sample Surveys

Sponsors: *ENAR and ASA Survey Research Methods Section*
Organizer: *Timothy G. Gregoire, Yale University*
Chair: *Timothy G. Gregoire, Yale University*
(Salon D)

8:30 Sampling and Experiments
Steven K. Thompson*, Pennsylvania State University

9:00 Nonparametric Model-Assisted Estimation of Distribution Functions from Survey Data
Alicia A. Johnson and Jay Breidt*, Colorado State University; Jean D. Opsomer, Iowa State University

9:30 Using Selection Functions to Combine Data from a Probability Survey and a Non-Probability Survey
Don L. Stevens, Jr.*, Oregon State University

10:00 Floor Discussion

2. Spatio-temporal Modeling of Environmental Processes

Sponsor: *ASA Section on Statistics and the Environment*
Organizer: *Christopher Wikle, University of Missouri, Columbia*
Chair: *Chong Z. He, University of Missouri, Columbia*
(Salon F)

8:30 Model-based Ecological Assessment of Riverine Systems by Combining Information from Multiple Sources
Mark S. Handcock*, University of Washington; Anthony Olsen, EPA-Corvallis

8:55 Space-Time Models Combining Hydrodynamics and Satellite Data
Jonathan R. Stroud*, University of Pennsylvania; Michael L. Stein, University of Chicago; Barry M. Lesht, Argonne National Laboratory; David Schwab, NOAA-GLERL; Dmitry Beletsky, University of Michigan

9:20 Dynamic Spatial Change-of-Resolution Modeling
Gardar Johannesson, Noel Cressie*, The Ohio State University; Hsin-Cheng Huang, Institute of Statistical Science, Academia Sinica

9:45 Efficient Parametrization of High-Dimensional Spatio-Temporal Models
Christopher K. Wikle* and Bill Xu, University of Missouri, Columbia

10:10 Floor Discussion

3. IMS Medallion Lectures: Applications of High-Dimensional Data Analyses to Microarray Data

Sponsor: *IMS*
Organizer: *Rafael Irizarry, Johns Hopkins University*
Chair: *Rafael Irizarry, Johns Hopkins University*
(Salon E)

8:30 Distance Weighted Discrimination
J. S. Marron*, UNC; Michael Todd, Cornell University

9:10 Study of Gene Expression: Statistics, Biology, and Microarrays
Ker-Chau Li*, UCLA

9:50 Discussant: Dennis Cook, University of Minnesota

10:10 Floor Discussion

4. Special Contributed Session: Advances in Growth Mixture Modeling for Preventive and Treatment Trials

Sponsor: *ENAR*
Organizers: *C. Hendricks Brown, University of South Florida and Bengt Muthén, University of California at Los Angeles*
Chair: *Karen J. Bandeen-Roche, Johns Hopkins University Bloomberg School of Public Health*
(Salon A)

8:30 Growth Mixture Modeling in Randomized Trials
Bengt Muthén*, University of California at Los Angeles

8:50 Multiple Imputation and Pseudoimputation for Model Diagnostics in Finite Mixture Growth Modeling
C. Hendricks Brown* and Chen-Pin Wang, University of South Florida

9:10 Bayesian Hierarchical Modeling of Heterogeneity in Multiple Contingency Tables
Getachew A. Dagne*, C. Hendricks Brown, University of South Florida; George W. Howe, The George Washington University

9:30 What's New in Latent Class Enumeration for General Growth Mixture Models
Katherine E. Masyn*, UCLA

9:50 Exact Marginal Likelihood and Conditional Tests of Interactions Involving Latent Class Regression and General Growth Mixture Modeling
Klaus Larsen*, Clinical Research Unit, Hvidovre Hospital, Denmark; C. Hendricks Brown, University of South Florida

10:10 Floor Discussion

5. Contributed Papers: HMDIN: "How Many Do I Need" or the Secrets of Sample Size Determination

Sponsor: *ENAR*
Chair: *Janet Wittes, Statistics Collaborative Inc.*
(Salon B)

8:30 Sample Size Calculation and Optimal Design With Logistic Regression
Eugene Demidenko*, Dartmouth Medical School

8:45 Experimental Design and Sample Size Determination for Testing Synergism
Ming T. Tan*, Hongbin Fang, and Guoliang Tian, University of Maryland Greenebaum Cancer Center; Peter Houghton, St. Jude Children's Research Hospital

9:00 Bayesian Sample Size Calculations for Case-Control Studies
Cyr Emile M'lan*, The Hospital for Sick Children, Toronto, Canada

9:15 Sample Size Determination for Comparing Several Survival Curves Under Proportional Hazards Model and Unequal Allocation
Susan Halabi*, Duke University; Bahadur Singh, Duke University and Lineberger Comprehensive Cancer Center-UNC

9:30 A Sample Size Method for Testing the Ratio of Two Mean Lifetimes: Application to Aging Intervention Study
Chengjie Xiong*, J. P. Miller, Yan Yan, Washington University in St. Louis

9:45 Floor Discussion

6. Contributed Papers: Advances in Mixed Models

Sponsor: *ENAR*
Chair: *Michael A. O'Connell, Insightful Corporation*
(Salon J)

8:30 Joint Modelling of Multivariate Longitudinal Profiles: Pitfalls of the Random-Effects Approach
Steffen Fieuws* and Geert Verbeke, Biostatistical Centre, Katholieke Universiteit Leuven, Belgium

8:45 Predicting Random Effects of Realized Clusters in Finite Populations
Edward J. Stanek III*, University of Massachusetts, Amherst; Julio D. Singer, University of São Paulo, Brazil

9:00 An Assessment of Heteroskedastic T-Error Linear Mixed Models for the Analysis of Field Data Collected from Diverse Environments
Kadir Kizilkaya and Robert J. Tempelman*, Michigan State University

9:15 Random Coefficient Dynamic Models for Longitudinal Data When the Baseline Variable is Correlated with Random Effects
Haihong Li*, Frontier Science and Technology Research Foundation, Inc.; Jianhua Z. Huang, University of Pennsylvania; Hulin Wu, Frontier Science and Technology Research Foundation, Inc.

9:30 Some Useful Diagnostics for Random Effects Models with Unweighted and Weighted Data
Julia L. Bienias*, Rush-Presbyterian-St. Luke's Medical Center; Charles B. Hall, Albert Einstein College of Medicine; Woojeong Bang, Rush-Presbyterian-St. Luke's Medical Center

9:45 Random Effects Selection in Linear Mixed Models
Zhen Chen* and David B. Dunson, National Institute of Environmental Health Sciences

10:00 Testing Dependent Observation Times for Longitudinal Data
Yangjin Kim*, University of Missouri, Columbia

7. Contributed Papers: Estimating Equations

Sponsor: *ENAR*
Chair: *Andreas G. Klein, University of Illinois, Urbana-Champaign*
(Salon I)

8:30 Orthogonalized Residuals for Estimation of Marginally Specified Association Parameters in Multivariate Binary Data
Richard C. Zink* and Bahjat F. Qaqish, UNC

8:45 Weighted Estimating Equations for Semiparametric Transformation Models with Censored Data from a Case-Cohort Design
Lan Kong*, Jiawen Cai, and Pranab Kumar Sen, UNC

9:00 Mean Estimating Equations Approach to Analysing Cluster-Correlated Data with Nonignorable Cluster Sizes
Emmanuel Benhin*, Statistics Canada; J. N. K. Rao, Carleton University, Canada; Alistair J. Scott, University of Auckland, New Zealand

9:15 Modified GEE and Goodness-of-Marginal-Fit Test with Correlated Binary Responses for Contingency Tables
Ji-Hyun Lee* and Bahjat F. Qaqish, UNC

9:30 Extended Estimating Equations for Link and Variance Function Parameters in Generalized Linear Models
Anirban Basu* and Paul J. Rathouz, University of Chicago

9:45 Semiparametric Rank Regression in Stability Analysis
Annpey Pong*, Forest Laboratories, Inc.; Ying Q. Chen and Biao Xing, UC Berkeley

10:00 Generalized Estimating Equations for Mixed-Effects Models
Wei Wang* and Yiliang Zhu, University of South Florida

8. Contributed Papers: Recurrent Event Data, Frailty Models, and Survival Analysis

Sponsor: *ENAR*
Chair: *David V. Glidden, University of California, San Francisco*
(Salon C)

8:30 Models and Bayesian Analysis of Recurrent Events Data
Debajyoti Sinha*, Medical University of South Carolina

8:45 Bayesian Inference for PVF Frailty Models
Madhuja Mallick* and Nalini Ravishanker, University of Connecticut

9:00 **Analysis of Frailty Survival Models using Poisson Variance Structures**
Shibao Feng* and Robert A. Wolfe, University of Michigan

9:15 Adjusting for Clinic Effects when Estimating a Treatment Effect: Stratified, Clustered, and Frailty Models
Lynn E. Eberly*, University of Minnesota; Lingfeng Yang, University of Pennsylvania

9:30 Two-Sample Test for the Distribution of Age at Onset from Life-Time Incidence Case-Control Data
Emilia Bagiella*, Columbia University

9:45 Proportion of Treatment Effect (PTE) Explained by a Surrogate Marker
Cong Chen*, Merck Research Labs; Hongwei Wang, Rutgers University; Steven M. Snapinn, Merck Research Labs

10:00 Confidence Interval Estimation for CHD Risk Prediction by Different Survival Models
Usha S. Govindarajulu* and Ralph B. D'Agostino, Sr., Boston University

9. Contributed Papers: Topics in Screening and Diagnostic Tests

Sponsor: *ENAR*
Chair: *Alicia A. Toledano, Brown University*
(Salon H)

8:30 Chronological Event Modeling for Screening Mammography
Prashni Paliwal*, University of Connecticut.; Alan E. Gelfand, Duke University; Linn Abraham, and William E. Barlow, Group Health Cooperative of Puget Sound; Joann Elmore, Group Health Cooperative and University of Washington

8:45 Understanding the Factors Underlying Disparities in Cancer Screening Rates Between Race/Ethnic Groups: The Peters-Belson Approach
Sowmya R. Rao* and Barry I. Graubard, National Cancer Institute, NIH, DHHS; Joseph L. Gastwirth, The George Washington University

9:00 Estimating Sensitivity and Specificity from Repeated Screening Tests
Michael K. Parides* and Emilia Bagiella, Columbia University

9:15 Bayesian Sequential Design for Multiple Discrete Outcomes with Application to Continuous Drug Screening
Peter Mueller, Gary L. Rosner, and Roberto Carta*, University of Texas M.D. Anderson Cancer Center

9:30 Combining Biomarkers Through Monotone Non-parametric Regression
Yue Wang* and Jeremy M. G. Taylor, University of Michigan

9:45 Floor Discussion

Monday, March 31
10:15–10:30 a.m.

Break – Grand Ballroom Pre-Function Area

Monday, March 31
10:30 a.m.–12:15 p.m.

10. Effects of Technological Advances on Introductory Statistics Courses

Sponsor: *ASA Section on Statistical Education*
Organizer: *Madhuri S. Mulekar, University of South Alabama*
Chair: *Madhuri S. Mulekar, University of South Alabama*
(Salon D)

10:30 How Technology Allows Data to Guide the Storyline of Statistical Studies
Christine A. Franklin*, University of Georgia

10:55 Using Technology to Investigate Statistical Concepts
Allan J. Rossman*, and Beth L. Chance*, Cal Poly-San Luis Obispo

11:20 WebStat 3.0: Web-based Software for Statistics Education
R. Webster West*, University of South Carolina

11:45 An Online Multimedia Textbook in Statistics
David M. Lane*, Rice University

12:10 Floor Discussion

11. Recent Developments in Clustering and Mixtures with Application to Spatial Models and Image Analysis

Sponsor: *IMS*
Organizer: *Ramani S. Pilla, Case Western Reserve University*
Chair: *Ramani S. Pilla, Case Western Reserve University*
(Salon G)

10:30 Partial Mixture Estimation With Application to Clustering David W. Scott*, Rice University; Chad A. Shaw, Baylor College of Medicine

10:55 Bayes Computations for Poisson and Marked Spatial Processes
Hemant Ishwaran*, Cleveland Clinic Foundation

11:20 Robust Clustering Methods and Visualization Based on Data Depth
Rebecka Jornsten,Yehuda Vardi*, Rutgers University and Cun-Hui Zhang, Rutgers University

11:45 Segmenting Magnetic Resonance Images via HierarchicalMixture Modelling
Carey E. Priebe*, Michael I. Miller, and J. Tilak Ratnanather, Center for Imaging Science, Johns Hopkins University

12:10 Floor Discussion

12. Statistics in Genetics

Sponsor: *ASA Biopharmaceutical Section*
Organizer: *Liwen Xi, Merck Co. Inc..*
Chair: *Liwen Xi, Merck Co. Inc.*
(Salon E)

10:30 Robust Singular Value Decomposition Analysis of Microarray Data
Li Liu*, Aventis Pharmaceuticals; Douglas M. Hawkins, University of Minnesota; Sujoy Ghosh, Glaxo Smith Kline; S. Stanley Young, National Institute of Statistical Sciences

11:00 Statistical Methods for Time Course Gene Expression Data
Hongzhe Li*, University of California, Davis

11:30 Statistical and Regulatory Issues in the Evaluation of Genetic and Genomic Tests
Gregory Campbell*, Food and Drug Administration

12:00 Discussant: A. Lawrence Gould, Merck Research Laboratories

13. Combining Disparate Environmental Data

Sponsor: *ENAR*
Organizers: *Carol A. Gotway Crawford, Centers for Disease Control and Prevention and Linda J. Young, University of Nebraska-Lincoln*
Chair: *Carol A. Gotway Crawford, Centers for Disease Control and Prevention*
(Salon F)

10:30 Combining Snow Water Equivalent Data from Multiple Sources to Estimate Spatio-Temporal Trends and Compare Measurement Systems
Mary Kathryn Cowles*, Dale L. Zimmerman, Aaron Christ, and David L. McGinnis, University of Iowa

10:55 Model Validation and Spatial Interpolation by Combining Observations with Outputs from Numerical Models
Montserrat Fuentes*, North Carolina State University

11:20 Combining State and National Survey Data to Assess Wildlife Population Trends
Sarah M. Nusser* and William R. Clark, Iowa State Univ.

11:45 Assessing Environmental Impacts Using Data from Multiple Scales
Linda J.Young*, University of Nebraska; Carol A. Gotway Crawford, Centers for Disease Control and Prevention

12:10 Floor Discussion

14. Contributed Papers: Bayesian Methods and Applications

Sponsor: *ENAR*
Chair: *Bradley P. Carlin, University of Minnesota*
(Salon A)

10:30 A Bayesian Selection Procedure for Rankings
Tom L. Bratcher* and Cody S. Hamilton, Baylor University

10:45 The Behrens-Fisher Problem: A Bayesian Solution
Jianrong Wu*, St. Jude Children's Research Hospital

11:00 Informative Prior Specification for Linear Regression with Interactions
Sally W. Thurston*, University of Rochester; Joseph G. Ibrahim, University of North Carolina; Susan Korrick, Harvard School of Public Health and Harvard Medical School

11:15 Bayesian Methods for Regression Using Surrogate Variables
David H. Manner*, Eli Lilly and Company; John W. Seaman and Dean M. Young, Baylor University

11:30 Improving Models and Methods for the Statistical Analysis of Tg.AC Mouse Bioassays
Wendell D. Jones* and Patrick Crockett, Analytical Sciences, Inc.

11:45 Bayesian Analysis of Multinomial Data
Chong He*, University of Missouri

12:00 On Comparing Poisson Rates when Data is Underreported with Applications to Cervical Cancer
James D. Stamey*, Stephen F. Austin State University; Tom L. Bratcher, Baylor University

15. IMS Contributed Papers: Topics in Mathematical Statistics

Sponsor: *IMS*
Chair: *Francesca Dominici, Johns Hopkins University Bloomberg School of Public Health*
(Salon B)

10:30 A Note on Finding Geodesic Equation of Two Parameter Gamma Distribution
William W. Chen*, Internal Revenue Service

10:45 Unimodal Density Estimation Using Kernel Methods
Peter Hall, Centre for Mathematics and its Applications, Australian National University; Li-Shan Huang*, University of Rochester

11:00 Estimation After Model Selection in a Gaussian Model
Vanja M. Dukic, University of Chicago; Edsel A. Pena*, University of South Carolina

11:15 Statistical Analysis of Single Molecule Experiments
Samuel Kou*, and Jun S. Liu, Harvard University

11:30 Semiparametric Skew-Elliptical Distributions
Yanyuan Ma* and Marc G. Genton, North Carolina State University

11:45 Statistical Analysis of an Interaction Threshold along a Fixed-Ratio Ray
Chris Gennings, Walter H. Carter, Jr., Virginia Commonwealth University; Ken K. Liao and Raymond S. Yang, Colorado State University;

12:00 Floor Discussion

16. Contributed Papers: Hazard Modeling and Estimation

Sponsor: *ENAR*
Chair: *Ronald Gangnon, University of Wisconsin, Madison*
(Salon J)

10:30 Use of a Validation Subset in Discrete Proportional Hazards Models with Mismeasured Outcomes
Amalia S. Meier*, Fred Hutchinson Cancer Research Center

10:45 **Mixed Baseline Additive Hazards Model for Multivariate Failure Time Data**
Guosheng Yin* and Jianwen Cai, UNC

11:00 The Cox Semi-Markov Illness-Death Model: A Large Sample Study
Youyi Shu*, Merck Research Laboratories; John P. Klein, Medical College of Wisconsin

11:15 Proportional Hazards Analysis of Randomized Trials Subject to Competing Causes of Censoring
Daniel O. Scharfstein, David S. Cohen*, Johns Hopkins Univ. Bloomberg School of Public Health; Mark J. van der Laan, University of California, Berkeley; James M. Robins, Harvard School of Public Health

11:30 Divergence of Relative Risk, Hazard Ratio, and Odds Ratio in Prospective Epidemiological Studies
Michael J. Symons and Dominic T. Moore*, University of North Carolina

11:45 Cox Model Based Prosper Function Analysis for Organ Allocation
Jian Yu* and Susan Murray, University of Michigan

12:00 Estimation of Transition Rates in a Multistage Proportional Hazards Model for Longitudinal Data
Rafiqul I. Chowdhury*, Kuwait University, Kuwait; Ataharul M. Islam, Science University, Penang, Malaysia; Makhdoom A. Shah, Kuwait University, Kuwait

17. Contributed Papers: Censored, Incomplete, and Missing Data

Sponsor: *ENAR*
Chair: *Susan Halabi, Duke University*
(Salon I)

10:30 The Effects of Model Misspecification and Unrecognizably Incomplete Data in Flowgraph Models
C. Lillian Yau*, Tulane University; Aparna V. Huzurbazar, University of New Mexico

10:45 Multiple Imputation in Survival Analysis of Interval- and Right-Censored Data
Jaroslaw Harezlak, Harvard School of Public Health; Wanzhu Tu*, Indiana University School of Medicine and Regenstrief Institute

11:00 Analysis of Medical Costs with Censored Data
Yong Zhou* and Jianwen Cai, UNC

11:15 Penalized Spline of Response Propensity Methods in Missing Data Analysis
Hyonggin An* and Roderick J. A. Little, University of Michigan

11:30 Drawing Inference from a Region of Estimates: Sensitivity Analysis for Missing Data
Stijn Vansteelandt*, Els Goetghebeur, Ghent University, Belgium; Michael Kenward, London School of Hygiene and Tropical Medicine, U.K.; Geert Molenberghs, Limburgs Universitair Centrum, Belgium

11:45 Unbiased Estimate and Efficient Tests in Clinical Trials with Dropouts
Peter C. O'Brien, Mayo Clinic; David Zhang*, Genentech; Kent R. Bailey, Mayo Clinic

12:00 Avoiding Bias when Rounding in Multiple Imputation
Nicholas J. Horton*, Boston University School of Medicine; Stuart R. Lipsitz, Medical University of South Carolina; Michael Parzen, University of Chicago

18. Contributed Papers: Statistical Methods in the Study of HIV and AIDS

Sponsor: *ENAR*
Chair: *Emilia Bagiella, Columbia University*
(Salon C)

10:30 Comparison of Linear, Nonlinear and Semiparametric Models for Estimating HIV Dynamic Parameters
Hulin Wu, Frontier Science & Technology Research Foundation; Caixia Zhao, Texas A&M University; Hua Liang*, St. Jude Children's Research Hospital

10:45 Per-Act Rate of HIV Transmission from Infected Thai Men to Their Momogamous Female Partners
Charles M. Heilig*, Centers for Disease Control and Prevention; Steve Shiboski, Univsersity of California, San Francisco

11:00 Estimation of Parameters and Assessment of Treatment Effects on HIV Pathogenesis Under HAART by State Space Models
Wai-Yuan Tan, Ping Zhang*, The University of Memphis; Xiao-Ping Xiong, Pat Flynn, St. Jude Children's Research Hospital

11:15 A Bayesian Approach for Estimating Antiviral Efficacy in HIV Dynamic Models
Yangxin Huang*, Hulin Wu, Frontier Science & Technology Research Foundation

11:30 **Analyzing Multiply Matched Cohort Studies with Two Different Comparison Groups: Application to Pregnancy Rates Among HIV+ Women**
Yan Li*, Daniel Zelterman, and Brian W. Forsyth, Yale University

11:45 Floor Discussion

19. Contributed Papers: Design Issues in Clinical Trials and Epidemiology

Sponsor: *ENAR*
Chair: *Maura Stokes, SAS Institute Inc.*
(Salon H)

10:30 A New Two-Dimensional Design for Phase I Clinical Trials
Kai Wang*, and Anastasia Ivanova, UNC

10:45 Discussion of Issues in Designing Protocols for Stem Cell Transplantation and Practical Variations of the Classical Phase II Study
Anne I. Goldman*, University of Minnesota

11:00 Simulations to Improve Experimental Designs for U-Shaped Dose-Response Modeling
Daniel L. Hunt*, St. Jude Children's Research Hospital

11:15 Optimal Study Design for Settings Where the Exposure of Interest is Subject to Short and Long Term Variation
Sohee Park* and Louise Ryan, Harvard School of Public Health

11:30 Evaluating Observer Agreement for Agreement Study Design with Replicated Readings
Jingli Song*, Huiman X. Barnhart, and Michael Haber, Emory University

11:45 A Cure Model Approach to Bayesian Phase I Trial Designs
Thomas M. Braun*, University of Michigan

12:00 Estimation of Influenza Antiviral Agent Efficacy Based on Household Data
Yang Yang* and Ira M. Longini, Emory University

Monday, March 31
12:15–1:30 p.m.

Roundtable Luncheons – Florida Salon IV

Monday, March 31
1:45–3:30 p.m.

20. Placebo-Controlled Trials: Ethics, Science and Economics

Sponsor: *ASA Biopharmaceutical Section*
Organizers: *Frank Shen, Bristol-Myers Squibb Co. and Stacy R. Lindborg, Eli Lilly Co.*
Chair: *Frank Shen, Bristol-Myers Squibb Co.*
(Salon C)

1:45 Placebo or Active Controls? Scientific and Ethical Considerations
Susan S. Ellenberg*, Center for Biologics Evaluation and Research, FDA

2:15 The Economics of Placebo in Drug Development
David J. DeBrota*, and Teresa S. Williams, Eli Lilly and Company

2:45 Influence of Control Group Design on Magnitude of Treatment Effect in Double-Blind Randomized Clinical Trials: The Case of Antipsychotic Efficacy for Schizophrenia
Scott W. Woods, Yale School of Medicine; Ralitza V. Gueorguieva, Robert W. Makuch*, Yale University

3:15 Discussant: Harry A. Guess, Merck Research Laboratories

21. Statistics and Counter-Bioterrorism

Sponsor: *ASA Section on Risk Analysis*
Organizer: *David Banks, CBER/FDA*
Chair: *Vincent Arena, University of Pittsburgh*
(Salon E)

1:45 Introduction

1:50 Statistical Models for Anthrax Outbreaks
Ron Brookmeyer* and Natalie Blades, Johns Hopkins University

2:20 Biosurveillance for Early Detection of Terrorism
Ken Kleinman*, Harvard Medical School

2:50 Risk Analysis and Game Theory
David Banks*, U.S. Food and Drug Administration

3:20 Floor Discussion

22. Innovative Approaches and Challenges in the Analysis of Recurrent Failure Time Data

Sponsor: *IMS*
Organizer: *Debashis Ghosh*
Chair: *Doug Schaubel, University of Michigan*
(Salon D)

1:45 The Accelerated Gap Times Model
Robert L. Strawderman*, Cornell University; Edsel A. Pena, University of South Carolina

2:10 The Gamma-Frailty Poisson Model for the Nonparametric Estimation of Panel Count Data
Ying Zhang*, University of Central Florida; Mortaza Jamshidian, California State University, Fullerton

2:35 Robust Inference for Non-Markov Event Data
David V. Glidden*, University of California, San Francisco

3:00 Evaluation of Treatments with Recurrent Event Data
Jerald F. Lawless*, University of Waterloo; Richard J. Cook, University of Waterloo

3:25 Floor Discussion

23. Bayesian Hierarchical Spatial Temporal Methods

Sponsor*: ASA Section on Statistics and the Environment*
Organizer: *Alan E. Gelfand, Duke University and Sudipto Banerjee, University of Minnesota*
Chair: *Alan E. Gelfand, Duke University*
(Salon G)

1:45 Nonstationary Spatial Modelling for Multiple Point Sources Sujit K. Ghosh* and Jacqueline M. Hughes-Oliver, North Carolina State University

2:15 Modelling Replicated Spatial Data Using Spatially-Varying Growth Curves
Sudipto Banerjee* and Gregg Johnson, University of Minnesota; Nick Schneider, Thorp Seed Co

2:45 Conditionally Specified Space-Time Models for Multiple Pollutants
Michael J. Daniels*, University of Florida; Zhigang Zhou, Iowa State University; Hui Zou, Stanford University

3:15 Floor Discussion

24. Statistical Design and Analysis of Large High Throughput Screening Data for Drug Discovery

Sponsor: *ASA Biopharmaceutical Section*
Organizer: *William J. Welch, University of Waterloo*
Chair: *William J. Welch, University of Waterloo*
(Salon F)

1:45 Introduction

1:50 High Throughput Drug Target Identification
S. Stanley Young*, National Institute of Statistical Sciences; Douglas M. Hawkins, University of Minnesota; Li Liu, Aventis Pharmaceuticals

2:20 Chemical Interpretation of Classifiers of High Throughput Screening Data
Yan Yuan*, William J. Welch, and Hugh A. Chipman, University of Waterloo

2:50 Design and Analysis of Pooling Experiments Based on Structure Activity Relationship
Lei Zhu*, GlaxoSmithKline; Jacqueline M. Hughes-Oliver, North Carolina State University; S. Stanley Young, National Institute of Statistical Sciences

3:20 Floor Discussion

25. Contributed Papers: Analysis of Microarrays and Gene Expression I

Sponsor: *ENAR*
Chair: *Ingo Ruczinski, Johns Hopkins University*
(Salon A)

1:45 Methods for Analyzing Array DNA Copy Number Data
Adam B. Olshen* and E. S. Venkatraman, Memorial Sloan-Kettering Cancer Center

2:00 Combining the Model Selection and False Detection Rates (FDRs) to Detect Differentially Expressed Genes in Microarray Experiments
Lang Li, Jingjin Li*, Indiana University; Xiaohua Zhou, University of Washington

2:15 Gene Selection for Microarray Classification using Principal Component Analysis
Antai Wang*, Georgetown University; Aiyi Liu, NIH/NICHD; Edmund A. Gehan, Georgetown University

2:30 Nylon Membrane Microarray Spot Deconvolution and Optimization of Array Grid Design
Lisa H. Ying* and Vladimir Svetnik, Merck & Co., Inc.

2:45 Statistical Design of Reverse Dye Microarrays
Kevin K. Dobbin*, Joanna Shih, and Richard Simon, National Cancer Institute

3:00 Differential Expression Analysis via Pooling Errors for Genes with Similar Expression Intensities
Michael A. O'Connell*, Insightful Corporation

3:15 Meta Tests in Microarray Data Analysis
Grier P. Page*, Jode W. Edwards, David B. Allison, University of Alabama at Birmingham

26. Contributed Papers: Binary and Binomial Data

Sponsor: *ENAR*
Chair: *Edward J. Stanek III, University of Massachusetts, Amherst*
(Salon J)

1:45 A Test for Heterogeneity of Effect in Conditional Logistic Regression
Randall H. Rieger*, West Chester University; Clarice R. Weinberg, NIEHS

2:00 A Novel Variable Selection and Significance Testing Technique for Quantile Regression
David T. Redden*, Jose R. Fernandez, and David B. Allison, University of Alabama at Birmingham

2:15 Validity of Using Logistic Regression Analysis in the Assessment of Correlation Between Hospital Volume and Surgical Mortality
Yen-Hong Kuo*, Meridian Health System, Neptune, New Jersey; Yen-Liang Kuo, Tri-Service General Hospital, Taipei, Taiwan, Republic of China

2:30 On a Family of Parametric Models for Analyzing Clustered Binary Data
E. Olusegun George, University of Memphis; Deo Kumar Srivastava*, St. Jude Children's Research Hospital; Roopa Seshadri, Northwestern University

2:45 A Mixture Model for the Analysis of Correlated Binomial Data
N. Rao Chaganty*, Old Dominion University

3:00 Analysis of Clustered Binary Data in the Presence of Varying Clusters
Abigail G. Matthews*, Rebecca A. Betensky, and Dianne M. Finkelstein, Harvard School of Public Health

3:15 Floor Discussion

27. Contributed Papers: Bioequivalence and Non-inferiority

Sponsor: *ENAR*
Chair: *Vance Berger, NIH*
(Salon I)

1:45 Distributional Properties of Ratio of Two Random Variables with Application to Active Control Non inferiority Trials
Yong-Cheng Wang*, Gang Chen, and George Chi, Food and Drug Administration

2:00 Two Approaches to the Behrens-Fisher Problem with Extensions to Replicate Data
Terry Hyslop*, Thomas Jefferson University

2:15 Exact Inference for Bioequivalence Between Two Drugs in Incomplete 2x2 Crossover Experiment
Sang-Gue Park*, Chung-Ang University and Harvard School of Public Health; Nam-Kyu Lim, Chung-Ang University

2:30 A New Asymptotic Test for Testing Non-Inferiority for Paired Binary Data
David Li* and Cong Chen, Merck Research Labs

2:45 Sample Size Re-calculation in Non-inferiority Trials that Include a Placebo Arm
Todd A. Schwartz*, UNC; Jonathan S. Denne, Eli Lilly and Company

3:00 Assessing Equivalence or Noninferiority in Screening Tests with Paired Binomial Endpoints
Boris G. Zaslavsky*, FDA/CBER/DBE

3:15 Floor Discussion

28. Contributed Papers: Nonparametric and Exact Methods

Sponsor: *ENAR*
Chair: *Dan Spitzner, Virginia Tech*
(Salon H)

1:45 Nonparametric Regression for Weakest Link Models
Roger S. Day* and Thomas J. Richards, University of Pittsburgh

2:00 Nonparametric Resampling-Based Tests for Association: With Application to High-Dimensional Unordered Categorical Covariate Sets
Greg DiRienzo*, Victor DeGruttola, and Marcello Pagano, Harvard School of Public Health

2:15 New Discriminant Procedure Applied to Structure Activity/Property
Kimberly S. Crimin*, Pharmacia Corporation

2:30 Variance Inflation in Strongly Nonnormal Bivariate Regression Models. Simulation Study
Simon Rosenfeld* and Victor Kipnis, National Institutes of Health and National Cancer Institute

2:45 Comparison of Unconditional Exact Tests for Testing the Difference of Independent Binomial Proportions
Jimmy A. Doi*, North Carolina State University; Roger L. Berger, National Science Foundation

3:00 Envelope Empirical Likelihood Ratio of Two Sample Hybrid Model with Censoring
Kyoungmi Kim*, University of Kentucky

3:15 Bayesian Isotonic Regression and Trend Analysis for Continous Regression Functions
Brian Neelon*, UNC; David B. Dunson, National Institute of Environmental Health Sciences

29. Contributed Papers: Statistical Inference for Correlated Data

Sponsor: *ENAR*
Chair: *Estelle Russek-Cohen, University of Maryland, College Park*
(Salon B)

1:45 **Distribution Diagnostics for Linear Models with Correlated Outcomes: Applications to Environmental Time Series**
E. Andres Houseman*, Louise M. Ryan, and Brent A. Coull, Harvard School of Public Health

2:00 A Latent Process Model for Joint Analysis of Time-To-Event and Longitudinal Data with Application to the Proscar Long-Term Efficacy and Safety Study
Weili He*, Merck & Co., Inc.; Weichung Joe Shih, Yong Lin, UMDNJ School of Public Health; Hui Quan, Merck & Co., Inc.; George R. Rhoads, UMDNJ School of Public Health

2:15 Analysis of Ten-Year PSA "Trajectories" After Prostate Cancer Surgery
Elizabeth L. Turner*, James A. Hanley, Nandini Dendukuri, McGill University; Peter C. Albertsen, University of Connecticut

2:30 Indexes of Tracking in Longitudinal Data with Individually Varying Times of Observation
Steven E. Hanna*, McMaster University

2:45 Optimal Criterion for Defining a Positive Test Based on Markers
Xuelin Huang*, University of Texas, M.D. Anderson Cancer Center

3:00 A Multi-State Stochastic Model for Longitudinal Data with Discrete Outcomes
Sujuan Gao*, Indiana University School of Medicine

3:15 **Testing for Spatial Correlation in Binary Data with Application to Aberrant Crypt Foci in Colon Carcinogenesis**
Tatiyana V. Apanasovich*, Texas A&M University

Monday, March 31
3:30–3:45 p.m.

Break – Grand Ballroom Pre-Function Area

Monday, March 31
3:45–5:30 p.m.

30. Statistical Methods for Studying Placebo Response

Sponsors: *ASA Biopharmaceutical Sectionand ASA Biometrics Section*
Organizer: *Eva Petkova, Columbia University and New York Psychiatric Institute*
Chair: *Todd Ogden, Columbia University*
(Salon F)

3:45 Identification of Placebo Responders Among Drug Treated Subjects
Thaddeus Tarpey*, Wright State University; Eva Petkova, Todd Ogden, and Ying Chen, Columbia Univ.

4:15 Growth Curve Modelling of Placebo and Drug Responses
Michael R. Elliott* and Thomas R. Ten Have, University of Pennsylvania School of Medicine

4:45 Individual Patient Data Meta-Regression of Placebo Response in Recent Antipsychotic Efficacy Trials
Jeffrey A. Welge* and Paul E. Keck, Jr., University of Cincinnati College of Medicine

5:15 Discussant: Eva Petkova, Columbia University and New York Psychiatric Institute

5:25 Floor Discussion

31. Recent Developments in Bayesian Non/Semi-parametrics

Sponsor: *IMS*
Organizer: *Lancelot James, Hong Kong University of Science and Technology*
Chair: *Hemant Ishwaran, Cleveland Clinic Foundation*
(Salon G)

3:45 Spatial Neutral to the Right Processes
Lancelot James*, Hong Kong University of Science and Technology

4:17 Posterior Consistency for Regression Model
R. V. Ramamoorthi*, Michigan State University

4:49 Statistical Validation of Computer Models
James O. Berger*, SAMSI and Duke University

5:21 Floor Discussion

32. Spatial Surveillance for Incident Clusters and Biological Terrorism

Sponsor: *ASA Section on Health Policy Statistics*
Organizer: *Ken Kleinman, Harvard Medical School and Harvard Pilgrim Health Care*
Chair: *Ken Kleinman, Harvard Medical School and Harvard Pilgrim Health Care*
(Salon E)

3:45 Introduction

3:50 A Poisson Cumulative Sum Approach to Spatial Surveillance
Peter A. Rogerson*, SUNY Buffalo

4:20 A Space-Time Permutation Scan Statistic for Spatial Disease Surveillance
M. Kulldorf*, University of Connecticut Health Center; F. Mostashari; R. Heffernan; J. Hartman

4:50 The Distribution of Interpoint Distances and Cluster Detection
Marco Bonetti, Laura Forsberg, Al Ozonoff and Marcello Pagano*, Harvard School of Public Health

5:20 Floor Discussion

33. Emerging Exact Methods in Biomedical Research

Sponsor: *ASA Biometrics Section*
Organizer: *Craig B. Borkowf, Centers for Disease Control and Prevention*
Chair: *Craig B. Borkowf, Centers for Disease Control and Prevention*
(Salon I)

3:45 Smooth and Accurate Multivariate Confidence Regions
Bo Yang, Shering-Plough Research Institute; John E. Kolassa*, Rutgers University

4:10 Adjusting for Baseline Covariates by Inducing a Partial Ordering
Vance W. Berger*, NIH; Judie Y. Zhou, Merck

4:35 Proving Non-Inferiority/Equivalence With Dichotomous Endpoints Using Exact Methods
Ivan S. F. Chan*, Merck Research Laboratories

5:00 Comparison of Exact Tests and Confidence Intervals for the Relative Risk (The Ratio of Two Binomial Parameters)
Pralay Mukhopadhyay* and Roger L. Berger, North Carolina State University

5:25 Floor Discussion

34. High-dimensional Hypothesis Testing and Model Selection for Functional Data Analysis

Sponsor: *ASA Biometrics Section*
Organizers: *Runze Li, The Pennsylvania State University and Dan J. Spitzner, Virginia Tech*
Chair: *James S. Marron, University of North Carolina, Chapel Hill*
(Salon B)

3:45 P-Splines as a Tool for Semiparametric Longitudinal Modelling
Naomi S. Altman*, Pennsylvania State University

4:10 High-Dimensional Testing in Functional Data Analysis With Applications to Means and Covariances
Dan J. Spitzner*, Virginia Tech

4:35 A Basis Function Approach for the Estimation of Varying Coefficient Models With Longitudinal Data
Jianhua Huang, University of Pennsylvania; Colin O. Wu*, DECA, National Heart Lung and Blood Institute; Lan Zhou, University of Pennsylvania

5:00 Variable Selection and Nonparametric Goodness-of-fit Test in Functional Data Analysis
Jianqing Fan, Chinese University of Hong Kong; Runze Li*, Pennsylvania State University

5:25 Floor Discussion

35. Contributed Papers: Analysis of Linkage and Relationship

Sponsor: *ENAR*
Chair: *Joanna Shih, National Cancer Institute*
(Salon A)

3:45 Multivariate Linkage Analysis in Quantitative Traits
Mariza de Andrade*, Curtis Olswold, and Stephen T. Turner, Mayo Clinic

4:00 Multipoint Linkage Disequilibrium Mapping for Complex Diseases
Kung-Yee Liang, Johns Hopkins University Bloomberg School of Public Health; Fang-Chi Hsu*, Wake Forest University School of Medicine; Terri H. Beaty, Johns Hopkins University Bloomberg School of Public Health

4:15 Weighting of Pedigrees in Genetic Linkage Analysis
Haydar Sengul*, Daniel E. Weeks, and Eleanor Feingold, University of Pittsburgh

4:30 Estimating and Approximating Expected Log-Likelihoods in Relationship Inference
Solveig K. Sieberts* and Elizabeth A. Thompson, University of Washington

4:45 **Power Calculations for Familial Aggregation Studies**
Nusrat Rabbee* and Rebecca Betensky, Harvard School of Public Health

5:00 Bayesian Variable-Selection with Related Predictors and Its Application in Quantitative Trail Mapping
Cheongeun Oh*, SUNY at Stony Brook

5:15 Floor Discussion

36. Contributed Papers: Statistical Methods for Longitudinal Data

Sponsors: *ENAR and IMS*
Chair: *Geert Verbeke, Katholieke Universiteit Leuven, Belgium (Salon C)*

3:45 Test for Longitudinal Data with Skewed Responses
Taesung Park*, University of Pittsburgh and Seoul National University; Jong-Hyeon Jeong, University of Pittsburgh; Sung-Hyun Kang, Seoul National University

4:00 On The Analysis of Longitudinal Data Measured Using Questionnaire Instrument
Shesh N. Rai*, Shelly Lensing, James M. Boyett, and Sean Phipps, St. Jude Children's Research Hospital

4:15 Detecting Disease Onset Using a Latent Class Model for Joint Analysis of Longitudinal and Survival Outcomes
Elizabeth H. Slate* and Liqiu Jiang, Medical University of South Carolina

4:30 The Use of Score Tests for Inference on Variance Components
Geert Verbeke*, Katholieke Universiteit Leuven, Belgium; Geert Molenberghs, Limburgs Universitair Centrum, Belgium

4:45 Latent Trajectory Models for Longitudinal Data: A Bayesian Approach
Sujata M. Patil*, and Trivellore E. Raghunathan and Jean T. Shope, University of Michigan

5:00 A Joint Cure Rate and Longitudinal Models with Applications to Cancer Vaccine Trials
Elizabeth R. Brown*, University of Washington; Joseph G. Ibrahim, UNC

5:15 Floor Discussion

37. Contributed Papers: Survey Sampling: Theory and Applications

Sponsor: *ENAR*
Chair: *Julia L. Bienias, Rush-Pres.-St. Luke's Medical Ctr. (Salon D)*

3:45 A Simple Method for Determining Confidence Intervals for Population Attributable Risk from Complex Surveys
Sundar Natarajan*, Ralph H. Johnson, VAMC & Medical University of South Carolina; Stuart R. Lipsitz, Medical University of South Carolina

4:00 When Is Stratification Detrimental to Design? Part I
Gretchen Marcucci*, Rho, Inc.

4:15 When is Stratification Detrimental to a Design? Part II
Katherine L. Monti*, Rho, Inc.

4:30 Variance-Covariance Functions for Domain Means of Ordinal Survey Items
A. James O'Malley* and Alan M. Zaslavsky, Harvard Medical School

4:45 Compatible Estimators of the Components of Growth for an Annual Forest Inventory Design
Francis A. Roesch*, USDA Forest Service, Asheville, NC

5:00 Sampling Experiments for Teaching Statistics
Michael J. Symons*, and Dana Quade, UNC

5:15 Floor Discussion

38. Contributed Papers: Semi-and Nonparametric Methods in Survival Analysis

Sponsor: *ENAR*
Chair: *Somesh Chattopadhyay, Florida State University (Salon H)*

3:45 Nonparametric Paired Two-Sample Tests for Censored Survival Data with Longitudinal Covariates
Shari Messinger*, University of Miami School of Medicine; Susan Murray, University of Michigan

4:00 Model Assessment and Outlier Detection for a Pseudospline-based Survival Model
Zekarias Berhane*, Joyce Chang, and Lisa Weissfeld, University of Pittsburgh

4:15 Nonparametric Estimation of the Bivariate Survivor Function
Zoe Moodie* and Ross L. Prentice, Fred Hutchison Cancer Research Center

4:30 Semiparametric Bayesian Analysis of the Proportional Odds Model
Nibedita Bandyopadhyay*, and Ananda Sen, Oakland University

4:45 Confidence Bounds for the Relative Risk of Two Binomial Distributions with Application in Survival Analysis
Changyong Feng*, University of Kansas Medical Center

5:00 Survival Analysis Using Auxiliary Variables via Nonparametric Multiple Imputation
Chiu-Hsieh Hsu* and Jeremy M.G. Taylor, University of Michigan

5:15 **Nonparametric Estimation for Baseline Hazard Function and Covariate Effects with Time-Dependent Covariates Under Proportional Hazard Structure**
Feng Gao* and Amita K. Manatunga, Emory University; Shande Chen, University of North Texas

Tuesday, April 1
8:30–10:15 a.m.

39. Sampling Methods for Selecting Population Controls

Sponsors: *ASA Survey Research Methods Section and ENAR*
Organizer: *Julia L. Bienias, Rush-Pres.-St. Luke's Medical Ctr.*
Chair: *Julia L. Bienias, Rush-Pres.-St. Luke's Medical Ctr.*
(Salon D)

8:30 Introduction

8:35 Population Inference in Case-Control Studies Nested within a Population-Based Cohort
Charles B. Hall*, Albert Einstein College of Medicine

9:00 Applying a Survey Sampling Perspective to the Design and Analysis of Population-Based Case-Control Studies
Ralph DiGaetano* and Lou Rizzo, Westat

9:25 Using Sample Surveys to Obtain a Control Group: Two Research Examples
Donna J. Brogan*, Emory University

9:50 Discussant: Barry I. Graubard, National Cancer Institute

10:05 Floor Discussion

40. Multi-stage Decisions and Dynamic Treatment Regimes

Sponsor: *IMS*
Organizer*: Thomas R. Ten Have, University of Pennsylvania*
Chair: *Thomas R. Ten Have, University of Pennsylvania*
(Salon C)

8:30 Estimation of Optimal Treatment Strategies
Jamie Robins*, Harvard University

8:55 Comparison of Designs for Adaptive Treatment Strategies: Baseline vs. Adaptive Randomization
Philip W. Lavori*, VA CSP/Stanford University; Ree Dawson, Frontier Science

9:20 Simulation-Based Sequential Bayesian Design
Peter Mueller*, University of Texas M.D. Anderson Cancer Center

9:45 Discussant: Susan Murphy, University of Michigan

10:05 Floor Discussion

41. Random Effects, Priors and Graphs: A 360 Degree View of Hierarchical Models

Sponsor: *ENAR*
Organizer: *Karen J. Bandeen-Roche, Johns Hopkins University Bloomberg School of Public Health*
Chair: *Qian-Li Xue, Johns Hopkins University*
(Salon G)

8:30 Location-Scale Models for Hierarchical Categorical Data
Donald Hedeker*, University of Illinois at Chicago

9:00 A Marginalized Multilevel Modeling Approach for Clustered Longitudinal Binary Data with Time Dependent Covariates
Diana L. Miglioretti*, Center for Health Studies, Group Health Cooperative; Patrick J. Heagerty, University of Washington

9:30 On Recent Developments in Graphical Models
Nanny E. Wermuth*, University of Mainz, Germany

10:00 Discussant: Karen J. Bandeen-Roche, Johns Hopkins University Bloomberg School of Public Health

10:10 Floor Discussion

42. Health Effects and Environmental Risk Assessment of Air Pollution

Sponsor: *ASA Section on Statistics and the Environment*
Organizer: *Sujit Ghosh, North Carolina State University*
Chair: *Sujit Ghosh, North Carolina State University*
(Salon H)

8:30 Covariate-Adjusted Spatial CDFs for Air Pollutant Data
Margaret Short, Bradley P. Carlin*, University of Minnesota; Alan E. Gelfand, Duke University

9:00 Short-Term Respiratory Health Effects of Air Pollution in Metropolitan Chicago
Vanja Dukic*, Paul Rathouz, Dana Draghicescu, Gidon Eshel, John Frederick, Ted Naureckas, and Alexis Zubrow, University of Chicago

9:30 Case-Crossover Analysis of the Health Effects of Air Pollution
Holly E. Janes, Lianne Sheppard*, Thomas Lumley, University of Washington

10:00 Discussant: Francesca Dominici, Johns Hopkins University

43. Contributed Papers: Applications in Statistical Genetics

Sponsor: *ENAR*
Chair: *Robert J. Tempelman, Michigan State University*
(Salon A)

8:30 Control of the False Discovery Rate for Random Search Procedures
Kenneth M. Boucher*, Aniko Szabo, University of Utah

8:45 Genome Association Studies of Complex Diseases by Case-Control Designs
Ruzong Fan*, Texas A&M University; Michael Knapp, University of Bonn, Germany

9:00 Use of Tree Models to Investigate Gene-Gene and Gene-Environment Interactions in Lung Cancer Risk
Carol J. Etzel*, University of Texas M.D. Anderson Cancer Center; Hui Zhao, Yumei Cao, University of Texas Health Science Center; Xifeng Wu, Qingyi Wei, Margaret R. Spitz, University of Texas M.D. Anderson Cancer Center

9:15 Measuring the Effect of APO-E Genotypes on Cardiovascular Disease Events in the Framingham Heart Study
Mark E. Glickman*, Mei Fang Kao, Boston University

9:30 Multivariate Conditional Analysis for Complex Traits
Jixiang Wu*; Dongfeng Wu, Mississippi State University; Johnie N. Jenkins, and Jack C. McCarty Jr., USDA-ARS, Mississippi State University

9:45 Bayesian Inference of Population Structure From Dominant Markers Using Mixture of Betas
Rongwei Fu*, Dipak K. Dey, and Kent E. Holsinger, University of Connecticut

10:00 Estimating the Number of False Null Hypotheses in a Multiple-Test Situation
Dan Nettleton*, Iowa State University; J. T. Gene Hwang, Cornell University

44. Contributed Papers: Clinical Trials: Adaptive, Group Sequential Designs, and Interim Analysis

Sponsor: *ENAR*
Chair: *Stewart J. Anderson, University of Pittsburgh*
(Salon I)

8:30 Mid-Course Correction to Group Sequential Clinical Trials
Cyrus R. Mehta*, Cytel Software Corporation; Anastasios A. Tsiatis, North Carolina State University

8:45 SCPRT Methods for Clinical Trials
Xiaoping Xiong*, St. Jude Children's Research Hospital; Ming Tan, University of Maryland Greenebaum Cancer Center; James Boyett, St. Jude Children's Research Hospital

9:00 An Interim Monitoring Strategy for a Small Sample Incidence Density Problem
Shane L. Rosanbalm*, Rho, Inc.

9:15 Estimating End of Trial Variance in Interim Analyses of Clinical Trials of Fixed Duration
Grant Izmirlian*, Richard Fagerstrom, Philip C. Prorok, National Cancer Institute

9:30 A Bayesian Sequential Procedure for Multiple Treatment Outcomes
Maria K. Mor* and Stewart J. Anderson, University of Pittsburgh

9:45 **Estimating Mean Response as a Function of Treatment Duration in an Observational Study where Duration May be Informatively Censored**
Brent A. Johnson* and Anastasios A. Tsiatis, North Carolina State University

10:00 Sensitivity Analysis for Longitudinal Clinical Trials
Herbert Thijs*, Geert Molenberghs, Carolien Beunckens, and Ivy Janssen, L.U.C. Center for Statistics; Craig H. Mallinckrodt, Eli Lilly and Company; and Raymond J. Carroll, Texas A&M University

45. Contributed Papers: Spatial Data: Statistical Inference and Modeling

Sponsor: *ENAR*
Chair: *Brian D. Marx, Louisiana State University*
(Salon J)

8:30 Composite Likelihood Bayesian Cluster Modelling of Small Area Health Data
Andrew B. Lawson*, University of South Carolina

8:45 Using the k-Nearest Neighbor Technique to Classify Satellite Imagery
Ronald E. McRoberts*, USDA Forest Service

9:00 A Data Augmentation Procedure for the Analysis of Censored Spatial Data
Brooke L. Fridley* and Philip Dixon, Iowa State University

9:15 A Spatial Analysis of Sea Turtle Nesting Patterns in Palm Beach County, Florida
Traci L. Leong*, Andrew Barclay, and Lance Waller, Emory University

9:30 Prediction of County Cancer Rates by Hierarchical Spatial Models
Linda W. Pickle*, National Cancer Institute

9:45 **Bayesian Hierarchical Spatial-Temporal Models for Wind Prediction**
Li Chen*, Montserrat Fuentes, and Jerry M. Davis, North Carolina State University

10:00 Fixed-Window Surveillance for Bioterrorism
Sylvan Wallenstein*, Mount Sinai School of Medicine; Joseph I. Naus, Rutgers University

46. Contributed Papers: Applications of Longitudinal Data Analysis

Sponsor: *ENAR*
Chair: *Shesh Rai, St. Jude Children's Research Hospital, Memphis (Salon B)*

8:30 Analysis of Longitudinal Data with Missing Values of Two Qualitatively Different Types
Ofer Harel*, Scott M. Hofer, and Joseph L. Schafer, The Pennsylvania State University

8:45 Semiparametric Analysis of Mean Response Model Accounting for Latent Lag and Saturation Times
Jingrong Yang* and Ying Qing Chen, University of California at Berkeley

9:00 Lifetime Event Trees: Descriptive Statistics and Inference
Oscar Loureiro* and Monika Raju, University of Massachusetts, Amherst; Jean J. Schensul, The Institute of Community Research; Edward J. Stanek III, University of Massachusetts, Amherst

9:15 Estimating Subject-Specific Variance Components with Multivariate Repeated Data
Wei-Ting Hwang* and Kathleen J. Propert, University of Pennsylvania School of Medicine

9:30 A Driver-Response Model of Pulsatile Hormone Data with a Test of the Pulsatile Association
Nichole E. Carlson*, Timothy D. Johnson and Morton B. Brown, University of Michigan

9:45 Digital Wound Imaging Enables Timely Identification of Chronic Non-healing Wounds Minimally Responsive to Hyperbaric Oxygen Treatment
Anuradha Roy*, The University of Texas at San Antonio; John Kalns, United States Air Force School of Aerospace Medicine; CleAnn Loeffler, Texas Lutheran University; and James K. Wright, United States Air Force School of Aerospace Medicine

10:00 Floor Discussion

Tuesday, April 1
10:15–10:30 a.m.

Break – Grand Ballroom Pre-Function Area

Tuesday, April 1
10:30 a.m.–12:15 p.m.

Presidential Invited Address

Sponsor: *ENAR*
Organizer/Chair: *Timothy G. Gregoire, Yale University* Grand Ballroom
(Salons E &F)

10:30 Introduction: Timothy G. Gregoire, Yale University

10:35 Student Travel Awards: Linda J. Young, University of Nebraska, Lincoln

10:45 Aids to Statistical Navigation,
Thomas A. Louis, Johns Hopkins University Bloomberg School of Public Health

Tuesday, April 1
1:45–3:30 p.m.

47. Network Design

Sponsor: *ASA Section on Statistics and the Environment*
Organizer: *David M. Holland, U. S. Environmental Protection Agency*
Chair: *Montserrat Fuentes, North Carolina State University (Salon D)*

1:45 Network Design for Semivariogram Estimation and Kriging
Dale L. Zimmerman*, University of Iowa

2:10 Design of Large-Scale Air Monitoring Networks
David M. Holland*, U. S. Environmental Protection Agency; Arin Chaudhuri and Montserrat Fuentes, North Carolina State University

2:35 Designs for Predicting the Extremes of Spatial Processes
James V. Zidek*, University of British Columbia; Nhu D. Le, BC Cancer Agency; Li Sun, Edmunds.com, Inc.

3:00 Discussant: Richard L. Smith, UNC

3:20 Floor Discussion

48. Statistical Issues in Serial Analysis of Gene Expression (SAGE)

Sponsor: *ENAR*
Organizer: *Jeffrey S. Morris, University of Texas M.D. Anderson Cancer Center*
Chair: *Kim Anh Do, M.D. Anderson Cancer Center*
(Salon E)

1:45 SAGE—A Statistician's View of the Underlying Biology
Keith A. Baggerly*, M.D. Anderson Cancer Center

2:10 Noise and Shadow: Statistical Methods for SAGE Data
Natalie J. Blades* and Giovanni Parmigiani, Johns Hopkins University

2:35 Differential Gene Expression Studies Using SAGE
Eleanor Feingold*, Lisa Weissfeld, and Yan Lin, University of Pittsburgh

3:00 Bayesian Shrinkage Estimation of the Relative Abundance
Jeffrey S. Morris*, Keith A. Baggerly, and Kevin R. Coombes, University of Texas M.D. Anderson Cancer Center

3:25 Floor Discussion

49. Functional Analysis of Longitudinal Data

Sponsor: *IMS*
Organizer: *Jane-Ling Wang, University of California, Davis*
Chair: *Xihong Lin, University of Michigan*
(Salon B)

1:45 Introduction

1:55 Medallion Lecture: Borrowing Strength in the Analysis of Longitudinal and Functional Data
John Rice, UC Berkeley

2:40 Marginal Non- and Semi-parametric Kernel Regression for Longitudinal Data
Naisyin Wang*, Raymond J. Carroll, Texas A&M University, Xihong Lin, University of Michigan

3:05 Discussant: Jane-Ling Wang, University of California, Davis

3:20 Floor Discussion

50. The Statistics and Politics of Recent Screening Controversies

Sponsor: *ENAR*
Organizer*: Steven Goodman, Johns Hopkins University School of Medicine*
Chair: *Steven Goodman, Johns Hopkins University School of Medicine*
(Salon F)

1:45 Introduction

1:50 Mammographic Screening for Breast Cancer: Statistics and Politics
Donald A. Berry*, M.D. Anderson Cancer Center

2:10 Hope and Controversy About Lung Cancer Screening with Helical CT
Constantine Gatsonis*, Brown University

2:30 Prostate Cancer Screening: Reconciling Good Intentions with Good Science
Barnett S. Kramer*, National Institutes of Health

2:50 Floor/Panel Discussion

51. Analysis of Adverse Event Reports

Sponsor: *ASA Section on Risk Analysis*
Organizer: *David Banks, Center for Biologics Evaluation and Research, U.S. Food and Drug Administration*
Chair: *David Banks, Center for Biologics Evaluation and Research, U.S. Food and Drug Administration*
(Salon A)

1:45 Introduction

1:50 Postmarketing Drug Adverse Event Surveillance and the Innocent Bystander Effect
William DuMouchel*, AT&T Shannon Laboratory

2:20 Spontaneous Fire Reports
Michael Greene* and Mark Levenson, Consumer Products Safety Commission

2:50 Estimating the Number of Disappeared
Jana L. Asher*, AAAS Science and Human Rights Program and Carnegie Mellon University; Patrick Ball, AAAS Science and Human Rights Program

3:20 Floor Discussion

52. Contributed Papers: Generalized Linear and Nonlinear Mixed Models

Sponsor: *ENAR*
Chair: *James A. Hanley, McGill University*
(Salon C)

1:45 Mixed Models for Estimating the Number of Distinct Species in Replicated Wildlife Surveys
Robert M. Dorazio*, U.S. Geological Survey; Andrew Royle, U.S. Fish and Wildlife Service

2:00 A Mixed Effects Modelling Approach to Clustered Binomial Data with Random Cluster Sizes
Renjun Ma*, University of New Brunswick

2:15 Generalized Linear Mixed Models: The Impact of Non-normal Random Effects for a Poisson Count Model
Kerrie P. Nelson*, University of South Carolina; Brian G. Leroux, University of Washington

2:30 **A General Approach for Two-Stage Analysis of Multi-Level Clustered Non-Gaussian Data**
Inna Chervoneva*, Thomas Jefferson University; Boris Iglewicz, Temple University

2:45 Analysis of Rater Agreement for Binary Responses
Svend Kreiner, Jorgen H. Petersen*, University of Copenhagen, Denmark; Klaus Larsen, Hvidovre University Hospital, Denmark

3:00 Multivariate Logistic Analyses Using Scale Mixtures of Normals
Sean M. O'Brien* and David B. Dunson, National Institute of Environmental Health Sciences

3:15 Floor Discussion

53. Contributed Papers: Stochastic Processes: Spatial, Temporal, and Spatio-Temporal

Sponsor: *ENAR*
Chair: *Sudipto Banerjee, University of Minnesota*
(Salon G)

1:45 A Convolution Kernel Approach to Modelling Diffusive Proporgation in Spatio-Temporal Processes
Bill Xu* and Christopher K. Wikle, University of Missouri, Columbia

2:00 On a Time Deformation Reducing Nonstationary Stochastic Processes to Local Stationarity
Marc G. Genton*, North Carolina State University; Olivier Perrin, University of Toulouse

2:15 Bayesian Inference for Pairwise Interaction Point Processes
Matthew A. Bognar* and Mary K. Cowles, University of Iowa

2:30 **Covariate-Adjusted and Bivariate Spatial CDF Modeling**
Margaret B. Short*, Bradley P. Carlin, University of Minnesota; Alan E. Gelfand, Duke University

2:45 Bayesian Factor Analysis for Spatially Correlated Data, with Application to Summarizing Area-Level Material Deprivation
Joseph W. Hogan, Brown University; Rusty Tchernis*, Harvard University

3:00 Prediction Performance of Correlated Error Procedures for Spatial Counts
Robert Downer*, Louisiana State University

3:15 Floor Discussion

54. Contributed Papers: Latent Variable Modeling

Sponsor: *ENAR*
Chair: *Elizabeth H. Slate, Medical University of South Carolina*
(Salon H)

1:45 Latent Predictors of Latent Class Outcomes
Melanie M. Wall*, University of Minnesota

2:00 Latent Variable Modeling for Misclassified Polytomous Outcome Variables
Jens C. Eickhoff*, University of Wisconsin, Madison; Yasuo Amemiya, Iowa State University

2:15 Misspecification Error in Latent Variable Models
Yun Ju Sung*, University of Minnesota

2:30 Longitudinal Measurement Error Model for Sleep Disordered Breathing
Liang Li*, Mari Palta, and Jun Shao, University of Wisconsin, Madison

2:45 The Quasi ML Estimation Method for Modeling Nonlinear Effects in a General Latent Variable Modeling Framework
Andreas G. Klein*, University of Illinois, Urbana-Champaign; Bengt O. Muthén, UCLA

3:00 Application of Growth Mixture Modeling to Assess the Development of Childhood Obesity
Chaoyang Li*, University of Kansas Medical Center

3:15 Floor Discussion

55. Contributed Papers: Survival Analysis and Clinical Trials

Sponsor: *ENAR*
Chair: *Lynn E. Eberly, University of Minnesota*
(Salon I)

1:45 **Optimal Estimator for the Survival Distribution and Related Quantities for Treatment Policies in Two-Stage Randomization Designs in Clinical Trials**
Abdus S. Wahed* and Anastasios A. Tsiatis, North Carolina State University

2:00 Partial Least Squares for the Accelerated Failure Time Model with Right Censored Data
Jie Huang* and David Harrington, Harvard School of Public Health

2:15 **Estimating Subject-Specific Survival Function under Accelerated Failure Time Model**
Yuhyun Park* and L. J. Wei, Harvard School of Public Health

2:30 Simultaneous Estimate of Treatment Effects on Two Binary Endpoints and Their Association in Randomized Clinical Trials
Wenquan Wang*, University of Iowa College of Public Health; Robert F. Woolson, Medical University of South Carolina; William R. Clarke, University of Iowa College of Public Health

2:45 An Evidential Alternative to the Logrank Test
Wesley D. Eddings*, Johns Hopkins University

3:00 Validation and Calibration: an Empirical Study of Prognostic Survival Models
Somesh Chattopadhyay* and Daniel L. McGee, Florida State University

3:15 Floor Discussion

Tuesday, April 1
3:30–3:45 p.m

Break – Grand Ballroom Pre-Function Area

Tuesday, April 1
3:45–5:30 p.m

56. Practical Issues in Nonlinear Mixed Models

Sponsor: *ENAR*
Organizers: *Walter W. Stroup, University of Nebraska, Lincoln and Estelle Russek-Cohen, University of Maryland, College Park*
Chair: *Larry Douglass, University of Maryland, College Park*
(Salon D)

3:45 Case Studies Illustrating Unanswered Questions in Inference With Nonlinear Mixed Models
Walter W. Stroup*, University of Nebraska, Lincoln

4:10 Approximation Techniques in Nonlinear Mixed Models: Strengths, Weaknesses and Implementation
Edward F. Vonesh*, Baxter Healthcare Corporation

4:35 Inference From Complex Experimental Designs Using Nonlinear Models: A Simulation Study
Erin E. Blankenship* and Walter W. Stroup, University of Nebraska, Lincoln

5:00 Small Samples and the Use of Nonlinear Mixed Effects Models: A Comparative Simulation Study
Mohamed Ould-Moustapha* and Estelle Russek-Cohen*, University of Maryland, College Park

5:25 Floor Discussion

57. The Measurement of Health Disparities

Sponsors: *ENAR and ASA Section on Health Policy Statistics*
Organizers: *Mark S. Handcock, University of Washington and Karen J. Bandeen-Roche, Johns Hopkins University Bloomberg School of Public Health*
Chair: *Mark S. Handcock, University of Washington*
(Salon G)

3:45 Studying Neighborhood Effects on Health: An Overview of Methodological Challenges
Stephen Raudenbush, University of Michigan

4:15 Ecological Inference in Epidemiology
Jon Wakefield, University of Washington

4:45 Discussant: Thomas A. Louis, Johns Hopkins University Bloomberg School of Public Health

5:10 Discussion

58. GEE and GMM: A Bridge between Biostatistics & Econometrics

Sponsors: *ENAR and IMS*
Organizer: *Sastry Pantula, North Carolina State University*
Chair: *Randy Carter, University of Florida*
(Salon E)

3:45 Introduction

3:50 A Biostatistical Perspective on the Use of Generalized Estimating Equations
John S. Preisser*, UNC

4:20 The Generalized Method of Moments Approach to Econometric Analysis
Alastair R. Hall*, North Carolina State University

4:50 Generalized Method of Moments (GMM) and Generalized Estimating Equations (GEE): Similarities and Differences
L A. Stefanski*, North Carolina State University

5:20 Floor Discussion

59. Current Research in Sample Size Re-Estimation

Sponsor: *ENAR and ASA Biopharmaceutical Section*
Organizer: *Christopher Coffey, University of Alabama at Birmingham*
Chair: *Keith E. Muller, University of North Carolina, Chapel Hill*
(Salon C)

3:45 Exact and Approximate Methods for Internal Pilots: Beyond Two Group Comparisons
Christopher S. Coffey*, University of Alabama, Birmingham; Keith E. Muller, UNC

4:10 Sample Size Redetermination for Repeated Measures Studies
David M. Zucker, Hebrew University and Jon S. Denne*, Eli Lilly and Company

4:35 Internal Pilot Designs for Confidence Intervals with Good Properties
Michael R. Jiroutek*, Bristol-Myers Squibb Pharmaceutical Research Institute; Keith E. Muller, UNC

5:00 Adaptive Designs in Clinical Trials – Are They Ready for Prime Time?
Janet Wittes*, Statistics Collaborative, Inc.

5:25 Floor Discussion

60. Emerging Study Designs in Intervention Studies and Clinical Trials: Challenges and Controversy

Sponsors: *ASA Biopharmaceutical Section and ASA Biometrics Section*
Organizer: *Xihong Lin, University of Michigan*
Chair: *Xihong Lin, University of Michigan*
(Salon F)

3:45 Causal Inference in Choice-Based Intervention Studies
Roderick J. Little*, Xihong Lin and Qi Long, University of Michigan

4:15 On the Inefficiency of the Adaptive Design for Monitoring Clinical Trials
Anastasios A. Tsiatis*, North Carolina State University

4:45 Evaluating Therapeutic Strategies in Multi-Course Clinical Trials
Peter F. Thall*, M.D. Anderson Cancer Center; Hsi-Guang Sung, Rice University; Elihu H. Estey, M.D. Anderson Cancer Center

5:15 Floor Discussion

61. Contributed Papers: Selection Bias, Verification Bias, and Propensity Scores

Sponsor: *ENAR*
Chair: *Mari Palta, University of Wisconsin, Madison*
(Salon H)

3:45 Explaining Post-propensity Score Covariate Balance
Thomas E. Love*, Case Western Reserve University

4:00 Reducing Bias in Observational Studies Using Propensity Score: A Simulation Study
Daniel C. Park*, Kwan R. Lee, Sam S. Joo, and Xiwu Lin, GlaxoSmithKline; Timothy W. Victor, AstroZeneca

4:15 A Generalized Imputation Method Based on Propensity Score
Hsiao-Lan Wei* and Howard E. Rockette, University of Pittsburgh

4:30 Uncertainty Envelope for Diagnostic Test Ignorance Region
Ying Chen* and Andrzej Kosinski, Emory University

4:45 Floor Discussion

62. Contributed Papers: Estimation and Modeling in the Presence of Missing Data and Nonresponse

Sponsor: *ENAR*
Chair: *Jong-Hyeon Jeong, University of Pittsburgh*
(Salon B)

3:45 Estimation of Parameters for Bivariate Survival Data in the Presence of Missing Covariate Data
Gina M. D'Angelo* and Lisa Weissfeld, University of Pittsburgh

4:00 Statistical Modeling in Tumor Xenograft Experiments with Incomplete Data
Ming Tan, Hong-Bin Fang*, and Guo-Liang Tian, University of Maryland Greenebaum Cancer Center; and Peter J. Houghton, St. Jude Children's Research Hospital

4:15 **A Hybrid Model for Longitudinal Binary Responses Subject to Non-Ignorable Dropout**
Kenneth J. Wilkins* and Garrett M. Fitzmaurice, Harvard School of Public Health

4:30 **Estimating Vaccine Efficacy Using Auxiliary Outcome Data and a Small Validation Sample**
Haitao Chu* and M. Elizabeth Halloran, Emory University

4:45 Estimation Based on Response-biased Data in the Presence of Supplementary Data
Jason Schroeder*, University of Memphis

5:15 Floor Discussion

63. Contributed Papers: Analysis of Microarrays and Gene Expression II

Sponsors: *ENAR and IMS*
Chair: *Kevin K. Dobbin, National Cancer Institute*
(Salon A)

3:45 Implications of Pooling RNA Samples for Microarray Experiments
Xuejun Peng*, Constance L. Wood, and Arnold J. Stromberg, University of Kentucky

4:00 Statistical Modeling of Microarray Data: Viable Alternatives to the Difference Model
Janet K. Holt*, Northern Illinois University; T. Mark Beasley, David B. Allison, The University of Alabama at Birmingham

4:15 Assessing Differential Expression with Few Oligonucleotide Arrays
Jianhua Hu*, Fred A. Wright, UNC

4:30 On the Problems of Sporadic Points and Local Minimum in Unsupervised Learning with Applications to Microarray Analysis
George C. Tseng*, Harvard School of Public Health; Wing H. Wong, Harvard University

4:45 Incorporating Estimation Variability into Detection of Differential Gene Expression in Microarray Experiments
Lei Shen*, The Ohio State University

5:00 Identifying Differentially Expressed Genes in Unreplicated Multiple-Treatment Microarray Timecourse Experiments
Rhonda R. DeCook*, Dan Nettleton, Iowa State University

3:15 Floor Discussion

64. Contributed Papers: Image Analysis: MRI, PET and Proteomics

Sponsors: *ENAR and IMS*
Chair: *Diana L. Miglioretti, Group Health Cooperative*
(Salon I)

3:45 Smooth Background Estimation in Spectra and Images
Paul H.C. Eilers*, Department of Medical Statistics, Leiden University Medical Centre

4:00 Modeling Drift of fMRI Data with Penalized Spline
Wen-Lin Luo* and Thomas E. Nichols, University of Michigan

4:15 Sensitivity of FWE-Corrected Inferences on t Statistic Images
Thomas E. Nichols* and Satoru Hayasaka, University of Michigan

4:30 Semi-local Wavelet Denoising of PET Images
Richard Charnigo*, Raymond Muzic, and Jiayang Sun, Case Western Reserve University

4:45 Multivariate Selection of Biomarkers for Metabolic Syndrome form 2D Gel Proteomics Data
Kwan R. Lee*, Xiwu Lin, Daniel C. Park, and Sergio Eslava, GlaxoSmithKline Pharmaceuticals R&D

5:00 Statistical Techniques in Proteomics–Analysis of 2D Gel Data
Sreelatha Meleth*, Jessy Deshane and Helen Kim, University of Alabama, Birmingham

5:15 Estimation Problems from Data with Change Points and Images
Stephen J. Ganocy* and Jiayang Sun, Case Western Reserve University

65. Contributed Papers: Inference on Risks and Rates

Sponsor: *ENAR*
Chair: *James J. Chen, National Center for Toxicological Research*
(Salon J)

3:45 Efficient Estimation of the Risk of a Disease by Quartile-Categories of a Predictor Variable Using Generalized Additive Models (GAMS)
Craig B. Borkowf*, Centers for Disease Control and Prevention and Paul S. Albert, National Cancer Institute

4:00 Joint Analysis of Incidence, Progression, Regression and Disappearance Rates
Guan-Hua Huang*, Ronald Klein and Barbara E.K. Klein, University of Wisconsin-Madison

4:15 Data Depth and Infant Mortality Differentials
Ann F. Von Holle* and Chirayath M. Suchindran, UNC

4:30 Effects of Data Limitations When Modeling Fatal Occupational Injury Rates
James F. Bena*, National Institute for Occupational Safety and Health; A. John Bailer, Miami University andNational Institute for Occupational Safety and Health

4:45 A Comparison of Variable Dimension Models and Overparameterized Models for Spatially Clustered Disease Rates
Ronald E. Gangnon* and Murray K. Clayton, University of Wisconsin, Madison

5:00 Estimating and Summarizing Interactions in Multiple Strata
Daniel L. McGee and Xu-Feng Niu*, Florida State University

5:15 A Parametric Model for Studying Organism Fitness using Step-Stress Experiments
Sonya Greven, UNC; A. J. Bailer*, Miami University; Lawrence L. Kupper and Keith E. Muller, UNC

Wednesday, April 2
8:30–10:15 a.m.

66. Causal Inference at the Biostatistics-Econometrics Interface

Sponsors: *IMS and ENAR*
Organizer: *Joe Hogan, Brown University*
Chair: *Joe Hogan, Brown University*
(Salon E)

8:30 Methodology for Evaluating a Partially Controlled Longitudinal Treatment using Principal Stratification, with Application to a Needle Exchange Program
Constantine E. Frangakis*, Ronald S. Brookmeyer, Ravi Varadhan, Mahboobeh Safaeian, Johns Hopkins University Bloomberg School of Public Health; David Vlahov, The New York Academy of Medicine; Steffanie A. Strathdee, Johns Hopkins University Bloomberg School of Public Health

8:55 Dynamic Treatment Effects
Karsten T. Hansen*, Nortwestern University; James J. Heckman, The University of Chicago

9:20 Discussant 1: Rod Little, University of Michigan

9:40 Discussant 2: Tony Lancaster, Brown University

10:00 Floor Discussion

67. Multivariate Mixed Effects Modeling of Discrete and Continuous Outcomes

Sponsor: *ENAR*
Organizer: *David B. Dunson, National Institute of Environmental Health Sciences*
Chair: *Zhen Chen, National Institute of Environmental Health Sciences*
(Salon F)

8:30 Multiple Outcomes with Applications to Risk Assessment in Toxicology Studies
Paul Catalano*, Harvard School of Public Health

8:55 Latent variable models for multivariate discrete and continuous outcomes
Louise Ryan*, Harvard School of Public Health

9:20 Bayesian Joint Models of Cluster Size and Subunit-Specific Outcomes
David B. Dunson* and Zhen Chen, National Institute of Environmental Health Sciences

9:45 Joint Mixed Models for Discrete and Continuous Outcomes
Ralitza V. Gueorguieva*, Yale University

10:10 Discussant: Mary Sammel, University of Pennsylvania

68. Special Contributed Session: Recent Methods to Reduce the Impact of Nuisance Parameters

Sponsors: *ENAR and IMS*
Organizer: *Amita Manatunga, Emory University*
Chair: *Amita Manatunga, Emory University*
(Salon G)

8:30 Composite Conditional Likelihood for Sparse Dependent Data
John J. Hanfelt*, Emory University

8:50 Adjusted Profile Estimating Function
Molin Wang*, Harvard School of Public Health; John J. Hanfelt, Emory University

9:10 A Type of Restricted Maximum Likelihood Estimator of Variance Components
Jiangang Liao*, UMDNJ

9:30 Elimination of Nuisance Parameters Via Locally-Ancillary Quasi-Scores
Paul J. Rathouz*, University of Chicago

9:50 Discussant: Bruce Lindsay, University of Pennsylvania

10:05 Floor Discussion

69. Contributed Papers: Survey Research Methods

Sponsor: *ENAR*
Chair: *Jay Breidt, Colorado State University*
(Salon D)

8:30 Semiparametric Model-Assisted Estimation of Distribution Functions in Surveys with Auxiliary Information
Alicia A. Johnson*, Jay Breidt, Colorado State University; Jean D. Opsomer, Iowa State University

8:45 Nonparametric Survey Regression Estimation in Two-Stage Spatial Sampling
Siobhan P. Everson-Stewart*, Jay Breidt, Colorado State University; Jean D. Opsomer, Iowa State University

9:00 Use of Random Permutation Model in Rate Estimation and Standardization
Wenjun Li*, Tufts-New England Medical Center; Edward J. Stanek III, University of Massachusetts, Amherst

9:15 A History of Data Mining Accidents in the 20th Century
Russel M. Upsumgrub, University of Pittsburgh; Ilene Over* and Eileen Forward, San Hill Institute of Technology;

9:30 Estimation Problems from Biased Data
Bin Wang* and Jiayang Sun, Case Western Reserve University

10:00 Floor Discussion

70. Contributed Papers: The Study of Health Effects and Health Care

Sponsor: *ENAR*
Chair: *Vanja Dukic, University of Chicago*
(Salon C)

8:30 Mutagenic Potency Estimation in the Ames/Salmonella Assay
Susan J. Simmons*, University of North Carolina, Wilmington; Walter W. Piegorsch, University of South Carolina; Errol Zeiger, Errol Zeiger Consulting

8:45 Modeling the Effects of a Bi-Directional Latent Predictor from Multivariate Questionnaire Data
Amy H. Herring*, UNC; David B. Dunson, National Institute of Environmental Health Sciences; Nancy Dole, Carolina Population Center

9:00 Effects of Medical Marijuana Laws on Drug Use among Hospital Emergency Room Admissions

Li Zhu* and Dennis Gorman, Texas A&M University

9:15 Estimation Concerns in the Measurement of Transitions in the Concentration of Medical Expenditures
Steven B. Cohen* and Trena Ezzati-Rice, Center for Cost and Financing Studies, Agency for Healthcare Research and Quality

9:30 The Use of the Kaplan-Meier Estimator of the Censoring Distribution in the Estimation of Medical Costs
Corina M. Sirbu* and Joseph C. Gardiner, Michigan State University

9:45 Incorporating Death into SF-36 Physical Component Score Analyses
Laura L. Johnson*, National Cancer Institute; Paula Diehr, University of Washington

10:00 Comparison of Standard Versus Modified Expiratory Techniques During Spirometry
Madhuri S. Mulekar*, University of South Alabama; Tony Cowan, Providence Hospital, Mobile; William Pruitt, University of South Alabama

71. Contributed Papers: Probability, Distribution Theory, and Inference

Sponsor: *ENAR*
Chair: *Calvin L. Williams, National Science Foundation and Clemson University*
(Salon H)

8:30 Smooth and Iterative Estimators for Copulas
Kouros Owzar*, Duke University Medical Center; and Pranab K. Sen, UNC

8:45 A Copula Model for Mixed Discrete and Continuous Outcomes
Wei Lang*, Wake Forest University School of Medicine; Lisa A. Weissfeld, University of Pittsburgh; Ralph B. D'Agostino, Jr., Wake Forest University School of Medicine

9:00 Confidence Intervals of the Ratio (or Difference) of Two Inverse Gaussian Means
Lili Tian*, University of Florida; Gregory Wilding, University of Buffalo

9:15 Urn Sampling and Interval Censoring
Lu Zheng* and Marvin Zelen, Harvard School of Public Health

9:30 A Comparison of Power of Log Rank and Wilcoxon Tests An Application to Hematologic Disorders Study
Pingfu Fu*, Case Western Reserve University; Mary J. Laughlin, University Hospitals of Cleveland, Case Western Reserve University

9:45 Changepoint Diagnostics for Competing Risks Models: A Conditional Posterior Predictive P-value Approach
Chen-Pin Wang*, University of South Florida; Malay Ghosh, University of Florida

10:00 Floor Discussion

72. Contributed Papers: Overdispersed Data and Hierarchical Models

Sponsor: *ENAR*
Chair: *Carol A. Gotway Crawford, Centers for Disease Control and Prevention*
(Salon B)

8:30 **Random Effects Models for Repeated Measures of Zero-Inflated Count Data**
Yongyi Min* and Alan Agresti, University of Florida

8:45 Hierarchical Models for Repeated Binary Data Using IBF Sampler
Ming Tan and Guoliang Tian*, University of Maryland, Greenebaum Cancer Center

9:00 Dynamic Linear Model of Mortality Data
Amitabha Bhaumik*, Dipak K. Dey, and Nalini Ravishanker, University of Connecticut

9:15 **Bayesian Approaches to Measuring and Mapping Small Area Disparities in Heart Disease Mortality in South Carolina**
Eric C. Tassone*, Lance A. Waller, Emory University; Michele Casper, Ishmael Williams, Kurt Greenlund, and Linda Neff, Centers for Disease Control

9:30 Ascent Based Monte Carlo EM
Brian S. Caffo*, Johns Hopkins University; Wolfgang S. Jank, University of Maryland; Galin L. Jones, University of Minnesota

9:45 Bayesian Inferences on Shape Constrained Hormone Trajectories in the Menstrual Cycle
Laura H. Gunn*, Duke University; David B. Dunson, National Institute of Environmental Health Sciences

10:00 Floor Discussion

73. Contributed Papers: Analysis of Gene Expression

Sponsor: *ENAR*
Chair: *Daniel J. Schnell, The Procter & Gamble Company*
(Salon A)

8:30 On Clustering Methods for Exploring HSC Gene Expression Data
Jie Chen*, University of Missouri-Kansas City; Xi He, and Linheng Li, Stowers Insititute for Medical Research

8:45 Exact Power under Independence for the False Discovery Rate in Gene Expression Array Experiments
Deborah H. Glueck*, University of Colorado Health Sciences Center; Keith E. Muller, UNC ; Larry Hunter, University of Colorado Health Sciences Center

9:00 Mantel Statistics to Correlate Gene Expression Levels from Microarrays with Clinical Covariates
Bill Shannon*, Washington University in St. Louis School of Medicine

9:15 Bayesian Mixture Model for Differential Gene Expression
Feng Tang*, Peter Muller and Kim-Anh Do, M.D. Anderson Cancer Center

9:30 Badge: Bayesian Differential Analysis of Gene Expression Data
Paola Sebastiani*, University of Massachusetts; Marco Ramoni, Harvard Medical School

9:45 Least Squares and Mixed Effects Models to Adjust for Batch Effects in Single-Channel CDNA Microarrays
Mary E. Putt*, Lan Zhou, Heping Hu, and Thomas R. Ten Have, University of Pennsylvania

10:00 Exploratory Survival Analysis with Applications to Gene Expression Microarray Data
Cheng Cheng*, St. Jude Children's Research Hospital

Wednesday, April 2
10:15–10:30 a.m.

Break – Grand Ballroom Pre-Function Area

Wednesday, April 2
10:30 a.m.–12:15 p.m.

74. Genome-Scale Biology: Statistical Challenges and Solutions

Sponsor: *IMS*
Organizer: *Jennifer F. Bryan, University of British Columbia*
Chair: *Natalie Blades, The Jackson Laboratory*
(Salon F)

10:30 Prediction of Survival with Gene Expression
Mark J. van der Laan*, University of California, Berkeley

11:00 Exploring Interactions in Genomic Data
Ingo Ruczinski*, Johns Hopkins University

11:30 Resolving Gene Expression Profiles with Tag-based Technologies
Jennifer F. Bryan*, University of British Columbia

12:00 Floor Discussion

75. Phase 2/3 Combination Designs to Accelerate Drug Development

Sponsors: *ASA Biopharmaceutical Section and ENAR*
Organizer: *Qing Liu, J&J Pharmaceutical Research and Development*
Chair: *Gordon Pledger, J&J Pharmaceutical Research and Development*
(Salon D)

10:30 Introduction

10:35 Background, Concept, and Regulatory Considerations for Multi-Stage Designs in the Implementation of the Accelerated Approval Mechanism
George Y.H. Chi*, Pei L.Yang, Rajeshwari Sridhara, and Gang Chen, Food and Drug Administration

10:55 On Effectiveness of Phase 2/3 Combination Designs for Clinical Development
Qing Liu* and Gordon Pledger, J&J Pharmaceutical

11:15 Estimation of Parameter and its Exact Confidence Interval Following Sequential Sample Size Re-Estimation Trials
Yu Shen*, M.D. Anderson Cancer Center; Yi Cheng, Indiana University at South Bend

11:35 Statistical and Practical Aspects of a Non-Stop Drug Development Strategy
Ronald W. Helms and Karen L. Kesler*, Rho, Inc.

12:00 Floor Discussion

76. Competing Risk Analysis

Sponsors: *ENAR and ASA Biometrics Section*
Organizer: *George W. Carides, Merck Research Laboratories*
Chair: *George W. Carides, Merck Research Laboratories*
(Salon C)

10:30 On the Use of the Competing Risks Proportional Hazards Model in the Choice of Optimal Treatment
Richard Kay*, PAREXEL International

11:00 Cumulative Incidence Regression
Jason P. Fine*, University of Wisconsin, Madison

11:30 Bounds on Joint Survival Probabilities with Positively Dependent Competing Risks
Kalyan Ghosh*, Merck & Co.; Sanat Sarkar, Temple University

12:00 Discussant: Per Kragh Andersen, University of Copenhagen

77. Statistical Aspects of Polygraphy

Sponsor: *ENAR*
Organizer: *Peter B. Imrey, The Cleveland Clinic Foundation*
Chair: *Peter B. Imrey, The Cleveland Clinic Foundation*
(Salon E)

10:30 Lies, Damned Lies: The Statistics of National Security Polygraph Screening
Stephen E. Fienberg, Carnegie-Mellon University; Patricia L. Grambsch, University of Minnesota; Peter B. Imrey*, The Cleveland Clinic Foundation; Aleksandra Slavkovic, Carnegie-Mellon University; For The Committee to Review the Scientific Evidence on the Polygraph, National Research Council

11:05 Meta-analysis of Coarsely Classified Diagnostic Data on Polygraph Testing
Xiao-Hua Andrew Zhou*, VA Puget Sound Health Care System and University of Washington

11:35 TBA

12:05 Floor Discussion

78. Contributed Papers: Semi- and Nonparametric Methods for Longitudinal Data

Sponsor: *ENAR*
Chair: *DuBois Bowman, The Rollins School of Public Health, Emory University*
(Salon B)

10:30 Kernel Estimation in Longitudinal Data with Outcome-Related Observation Times
Donglin Zeng*, UNC; Daohai Yu, Duke University

10:45 Semi-parametric Longitudinal Models Subject to Missing Data by Design
Jaroslaw Harezlak* and Louise Ryan, Harvard School of Public Health; Nicholas Lange, Harvard University Schools of Medicine and Public Health

11:00 A Semiparametric Likelihood Approach to Joint Modeling of Longitudinal and Time-To-Event Data
Xiao Song*, University of Washington; Marie Davidian and Anastasios A. Tsiatis, North Carolina State University

11:15 Local Polynomial Regression Analysis of Clustered Data
Kani Chen, The Hong Kong University of Science and Technology; Zhezhen Jin*, Columbia University

11:30 Backfitting and Local Likelihood Methods for Nonparametric Mixed-effects Models with Longitudinal Data
Jeong-gun Park* and Hulin Wu, Frontier Science & Technology Research Foundation

11:45 Semiparametric Estimation in Transition Measurement Error Models
Wenqin Pan*, Rho, Inc.; Donglin Zeng, UNC; Xihong Lin, University of Michigan

12:00 A Technique for Implementing Fixed-Knot Regression Splines in the General Linear Mixed Model
Lloyd J. Edwards* and Paul W. Stewart, UNC; James E. MacDougall, Genzyme; Ronald W. Helms, Rho, Inc.

79. Contributed Papers: Receiver Operating Characteristics and Screening

Sponsor: *ENAR*
Chair: *Brian Wiens, Amgen Inc.*
(Salon G)

10:30 A Comparison of the Dorfman-Berbaum-Metz and Obuchowski Methods for Receiver Operating Characteristic (ROC) Data
Stephen L. Hillis*, Iowa City VA Medical Center

10:45 Properties of the Partial Area Under the Curve (AUC) for Summary ROC Curves
Stephen D. Walter*, McMaster University

11:00 Assessing the Efficiency of CT Imaging for Detecting and Classifying Liver Lesions in Candidates for Hepatic Surgery
Richard E. Thompson*, Johns Hopkins University Bloomberg School of Public Health; Ihab R. Kamel, Johns Hopkins School of Medicine

11:15 Design and Analysis of Comparative Diagnostic Accuracy Studies with Multiple Correlated Test Results
Aiyi Liu*, National Institute of Child Health and Human Development (NICHD), National Institutes of Health; Enrique F. Schisterman, National Institute of Child Health and Human Development (NICHD), National Institutes of Health; Madhu Mazumdar, Sloan-Kettering Cancer Center

11:30 Decoding Pools of Chemical Compounds in the Presence of Compound Interaction and Dilution
Bingming Yi*, Jacqueline M. Hughes-Oliver, North Carolina State University; Stanley S. Young, GlaxoSmithKline

11:45 Floor Discussion

80. Contributed Papers: Causal Inference and Missing Data

Sponsor: *ENAR*
Chair: *Paul J. Rathouz, University of Chicago*
(Salon H)

10:30 Semiparametric Marginal Regression Models for Repeated Outcomes in the Presence of Informative Follow-up
Haiqun Lin*, Yale University; Daniel O. Scharfstein, Johns Hopkins University Bloomberg School of Public Heallth; Robert A. Rosenheck, Yale University

10:45 **Likelihood Methods for Counterfactual Inference with Confounding Variables**
Abigail L. Jager* and Paul J. Rathouz, University of Chicago

11:00 Causal Linear Models for Non-compliance under Randomized Treatment
Thomas R. Ten Have*, Michael Elliott, Marshall Joffe, and Elaine Zanutto, University of Pennsylvania

11:15 Semiparametric Efficient Estimation of Treatment Effect in a Pretest-Posttest Study
Selene Leon*, Anastasios A. Tsiatis, and Marie Davidian, North Carolina State University

11:30 Multiple Imputation for Nonrandom Missing Outcomes in Trials with Ancillary Reliability or Validity Measures
James B. Kampert*, The Cooper Institute

11:45 Fitting a Mixed Effects Markov Model for Repeated Binary Outcomes with Nonignorable Dropout
Robert J. Gallop*, West Chester University; Thomas R. Ten Have, University of Pennsylvania

12:00 On the Construction of Bounds in Prospective Studies with Missing Ordinal Outcomes: Application to the Good Behavior Game Trial
Daniel O. Scharfstein*, Johns Hopkins University Bloomberg School of Public Health; Charles F. Manski, Northwestern University; James C. Anthony, Johns Hopkins University Bloomberg School of Public Health

81. Contributed Papers: Random Effects and Measurement Error

Sponsor: *ENAR*
Chair: *Ramon C. Littell, University of Florida*
(Salon I)

10:30 Statistical Analysis of PSA Patterns Following Radiation Therapy for Prostate Cancer
James A. Hanley*, Elizabeth L. Turner, Nandini Dendukuri, McGill University; Peter C. Albertsen, University of Connecticut

10:45 Neurotoxicity Risk Assessment: Modeling of Behavioral Changes in Multiple Endpoints
Yiliang Zhu*, University of South Florida

11:00 Performance of Shrinkage Estimators for Prediction in Multiple Regression
Xue Xin*, Tulane University School of Public Health and Tropical Medicine; Leann Myers

11:15 Measurement Error Adjustment in Occupational Health Studies while Accounting for Nondetectable Exposures
Kathleen A. Wannemuehler* and Robert H. Lyles, Emory University

11:30 An Alternative Noniterative Estimator for the Heterogeneity Variance in Meta-Analysis
Kepher H. Makambi*, Howard University Cancer Center

11:45 Floor Discussion

82. Contributed Papers: Analysis of Microarrays and Gene Expression III

Sponsor: *ENAR*
Chair: *Michael O'Connell, Insightful Corporation*
(Salon A)

10:30 Statistical Significance Analysis of Longitudinal Gene Expression Data
Xu Guo*, Huilin Qi, Catherine M. Verfaillie, Wei Pan, University of Minnesota

10:45 Detecting DNA Regulatory Motifs by Incorporating Positional Trends in Information Content
Katherina J. Kechris*, Eric van Zwet, Peter Bickel, and Michael Eisen, UC Berkeley

11:00 A Generalized Additive Model for Microarray Gene Expression Data Analysis
Chenan Tsai*, Food and Drug Administration; Hueymiin Hsueh, National Chengchi University, Taipei, Taiwan; James J. Chen, Food and Drug Administration

11:15 **A Bayesian Classification Method for Treatments Using Microarray Gene Expression Data**
Yuan Ji*, Kam-Wah Tsui, and Kyungmann Kim, University of Wisconsin, Madison

11:30 Bayesian Error-in-Variables Survival Model for the Analysis of Genechip Arrays
Mahlet G. Tadesse*, Texas A&M University; Joseph G. Ibrahim, UNC; Robert Gentleman, Harvard School of Public Health and Dana-Farber Cancer Institute; Sabina Chiaretti, Dana-Farber Cancer Institute

11:45 Linear Regression and Two-Class Classification with Gene Expression Data
Xiaohong Huang*, and Wei Pan, University of Minnesota

12:00 Floor Discussion

83. Contributed Papers: Generalized Linear Models and Regression

Sponsor: *ENAR*
Chair: *Linda W. Pickle, National Cancer Institute*
(Salon J)

10:30 **A Semiparametric Empirical Likelihood Method for Biased Sampling Schemes in Epidemiologic Studies with Auxiliary Covariates**
Xiaofei Wang* and Haibo Zhou, UNC

10:45 **Functional Regression Models and Temporal Processes**
Jun Yan*, University of Wisconsin

11:00 Censored Regression Models for Predicting Suicide Attempt Lethality
Hanga C. Galfalvy*, Maria A. Oquendo, Michael F. Grunebaum, and John J. Mann, New York State Psychiatric Institute

11:15 Regressions with Singular Design
Wenjiang Fu*, Michigan State University

11:30 Methods of Learning in Statistical Education: Design of a Randomized Trial and Analysis using Marginal Structural Models
Felicity T. Boyd* and Marie Diener-West, Johns Hopkins University Bloomberg School of Public Health

11:45 Modeling Excess Zeros in Ordinal Data
Mary E. Kelley*, VA Pittsburgh Healthcare System; Stewart J. Anderson, University of Pittsburgh Graduate School of Public Health

12:00 Floor Discussion

POSTER SESSION SUMMARY

1. Tests for Effects on Tumor Frequency and Latency in Multiple Dosing Photocarcinogenicity Experiments
   Daniel F. Molefe*, Ralph L. Kodell, and James J. Chen, National Center for Toxicological Research

2. Lifetime Risk Of Diabetes Mellitus in the United States
   Kabayam M. Venkat Narayan, James P. Boyle*, Theodore J. Thompson, Stephen W. Sorensen, David W. Williamson, Centers for Disease Control and Prevention

3. A Bayesian Approach to Cost-Effectiveness when the Outcome Depends on a Predictive Model
   Raymond G. Hoffmann*, Laud W. Purushottam, Medical College of Wisconsin
   Hayat Matthew, HIH/NIDCD
   Ana Johnson-Masotti, McMaster University

4. Approaches for Haplotype-Phenotype Association Studies in Unrelated Subjects
   Serkalem Demissie*, Boston University School of Public Health
   Jose M. Ordovas, Jean Mayer USDA Human Nutrition Research Center on Aging at Tufts University
   L. Adrienne Cupples, Boston University School of Public Health

5. Comparing Nonparametric Correlation Coefficients
   Leann Myers* and Maria J. Sirois, Tulane University

6. Geometry of the Bivariate Normal Distribution
   Timothy D. Curtsinger* and Linda J. Goldsmith, University of Louisville

7. The Association of Estrogen Receptor Polymorphism and Change in Osteocalcin, a Bone Turnover Marker
   Wei Liang*, Mary L. Jannausch and Maryfran R. Sowers, University of Michigan

8. Comparison of Results from Two Methods for Examining the Reproducibility of Breast Fluid Biomarkers
   Peter H. Gann, Robert T. Chatterton, Jr., Borko D. Jovanovic, Irene B. Helenowski*, and Angela S. Geiger, Northwestern University

9. Review and Evaluation of Methods for Computing Confidence Intervals for the Ratio of Two Proportions and Implications to Non-Inferiority Clinical Trials
   Rebekkah S. Dann*, and Gary G. Koch, University of North Carolina at Chapel Hill

10. An Evaluation of Methods for Missing Data in a Large-Scale Epidemiological Study
    Jamie L. Bigelow*, Amy Herring, Nancy Dole, Diane Kaczor and David Savitz, University of North Carolina at Chapel Hill

11. Spot Quality Measures in Microarray Experiment
    Taesung Park, University of Pittsburgh and Seoul National University
    Ki Woong Kim, Seoul National University
    Seung Yeoun Lee*, Sejong University
    Jin Hyuk Kim, Hanyang University

12. Estimating risk factors under clustered survival data with time dependent covariates
    Leila D. Amorim*, Shrikant Bangdiwala and Joanne S. Harrell, University of North Carolina at Chapel Hill

13. Group Level Effects in Hierarchical Poisson Models
    Leslie H. Morgan*, Leann Myers, and Frances J. Mather, Tulane University

14. Modeling Influenza Epidemics
    Paola Sebastiani, John Tsimikas, Ling Wang*, University of Massachusetts Amherst;
    Kenneth Mandl, Children's Hospital Informatics Program Harvard Medical School

15. Use of the Kappa With Asymmetric Weights When Evaluating Agreement of Cervical Cytology
    Todd G. Nick*, Zelma Cason, University of Mississippi Medical Center

16. The Effect of Missing Data in the Study of SNPS-to-Phenotype Association
    Min Qin*, Covance Laboratories Inc.
    Paola Sebastiani, University of Massachusetts, Amherst

17. Special Numbers in Sample Size Analysis
    Linda J. Goldsmith*, University of Louisville

18. Sample Size Calculations in Phase II Clinical Trials Using Informative Conugate Priors
    Matthew S. Mayo and Byron J. Gajewski*, University of Kansas Medical Center

19. LINMOD 4: A Program for General Linear Multivariate Models With Missing Data
    Diane J. Catellier* and Keith E. Muller, University of North Carolina

20. Linkage Testing of Affected Subjects When Pedigree Information is Uncertain or Ignored: A Mixture Model Approach
    Bobby L. Jones* and Daniel E. Weeks, University of Pittsburgh Department of Human Genetics

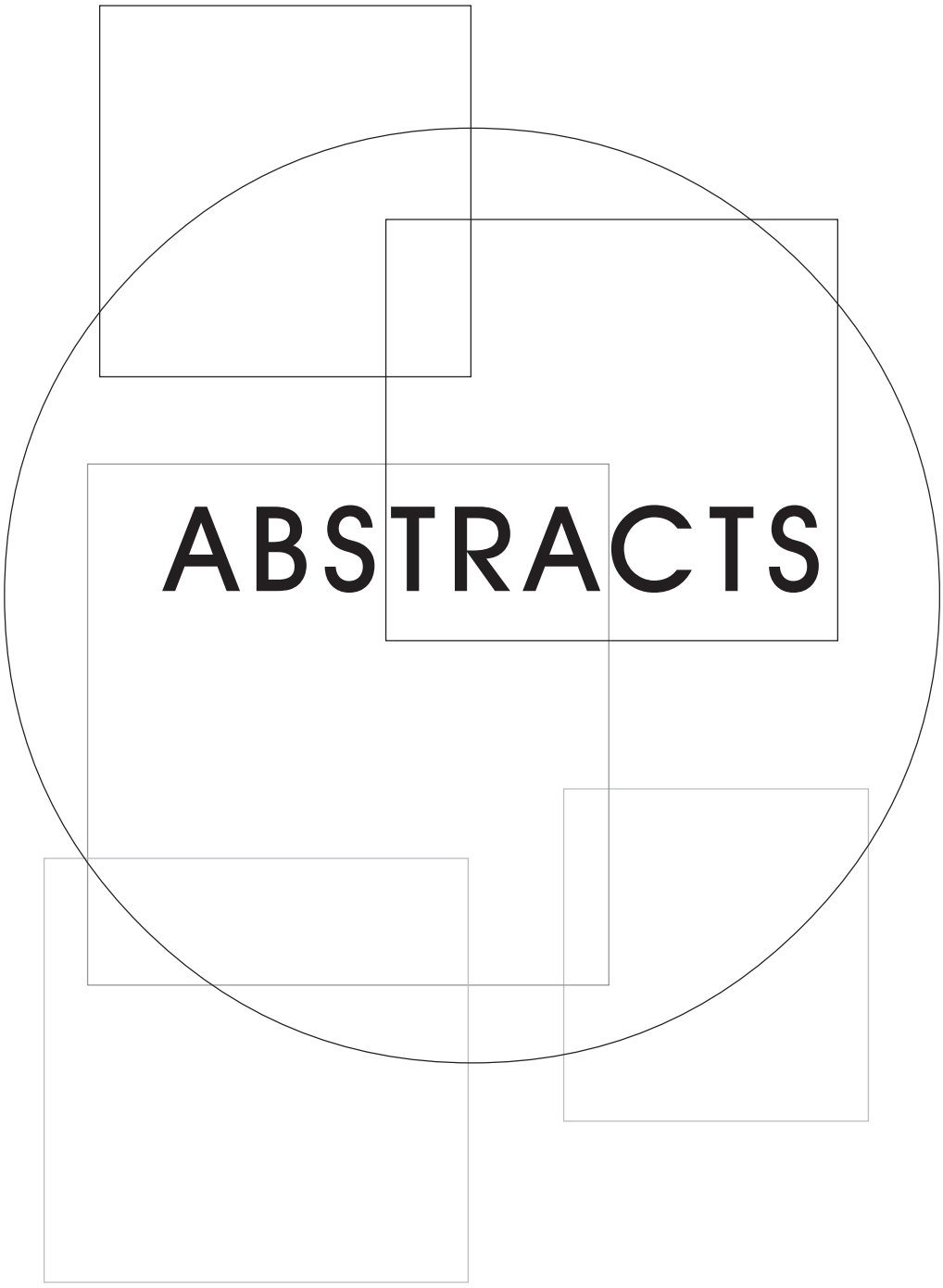
21. Selecting the Optimal Transformation of a Continuous Covariate in Cox's Regression: Implications for Hypothesis Testing
    Mamun Mahmud*, McGill University;
    Michal Abrahamowicz, McGill University and Montreal General Hospital
    Karen Leffondré, Montreal General Hospital

22. Modeling Chronic Stress in Longitudinal Data and the Relation to Heart Disease Development
    Imke Janssen*, Kelly Karavolos and Lynda H. Powell, Rush University

NOTES

ENAR WEBSITE
www.enar.org


I
B
S
ENAR


ABSTRACTS

1. Linking Models to Design and Inference in Sample Surveys

SAMPLING AND EXPERIMENTS

Steven K. Thompson*, Pennsylvania State University

Many studies of natural and human populations involve not only observing a sample from the population but making some type of intervention in the population or performing an experiment to determine the effect of an intervention. Examples include environmental intervention and remediation studies, investigations into methods of forestry or fisheries management, ecological perturbation studies, and environmental health studies. Experimental design concernshow the potential treatments are assigned to units in the study while sampling concerns how the units are selected to be in the study in the first place. In this talk the interrelationship of each of these aspects and implications for study design and inference will be discussed.

e-mail: skt@stat.psu.edu

NONPARAMETRIC MODEL-ASSISTED ESTIMATION OF DISTRIBUTION FUNCTIONS FROM SURVEY DATA

Alicia A. Johnson, Colorado State University
Jay Breidt*, Colorado State University
Jean D. Opsomer, Iowa State University

Estimation of a finite population distribution function under a general sampling design, using auxiliary population information, is considered. Model-based and model-assisted/design-based estimators are compared, where the models in each case are either parametric or non-parametric. Performance of the non-parametric model-assisted estimator relative to the other estimators is evaluated analytically and via simulation under varying degrees of model misspecification.

e-mail: jbreidt@stat.colostate.edu

USING SELECTION FUNCTIONS TO COMBINE DATA FROM A PROBABILITY SURVEY AND A NON-PROBABILITY SURVEY

Don L. Stevens,Jr*, Statistics Department, Oregon State University

Naïve combination of probability and non-probability survey data can result in selection bias and consequent estimation bias. We investigate the use of selection functions to the question of estimating an appropriate weight for non-probability samples. If X has pdf f1(x), a selection function w(x)is a function such that if individuals with X = x are selected with probability w(x) from the population, then the pdf of the resulting population is f2(x) = cw(x)f1(x). We can view the non-probability sample as being filtered through a selection function. If we have a complete auxiliary variable, the ratio of its known pdf to the pdf estimated from the non-probability sample is proportional to the selection function, expressed as a function of the auxiliary variable. In order to apply this to our response variable, we need the selection function expressed as a function of the response variable. Thus, we need to be able to build a plausible model of the response variable i.e., y = h(x1, x2,..., xk), where the xi are complete auxiliary variables. We can then use the distribution of h(×) on the population and the distribution of h(×) on the non-probability sample to estimate w.

e-mail: stevens@stat.orst.edu

---

2. Spatio-Temporal Modeling of Environmental Processes

MODEL-BASED ECOLOGICAL ASSESSMENT OF RIVERINE SYSTEMS BY COMBINING INFORMATION FROM MULTIPLE SOURCES

Mark S. Handcock*, University of Washington
Anthony Olsen, EPA-Corvallis

We describe an approach to improve understanding of the biological integrity of stream and river systems in the United States Mid-Atlantic Region by combining information from separate spatial-temporal monitoring surveys, available contextual information on hydrologic units and remote sensing information. We develop hierarchical spatial statistical models for environmental indicators on the streams and rivers that capture the spatial variation in the measures. These models have been used to estimate the indicators through the riverine system based on the information from multiple sources and aggregate scales. We also quantify the uncertainty in the estimates and develop methods to visualize the resulting estimates and uncertainties.

This research illustrates how statistical methodology can be used to leverage the information in scattered monitoring surveys.

e-mail:

SPACE-TIME MODELS COMBINING HYDRODYNAMICS AND SATELLITE DATA

Jonathan R. Stroud*, Department of Statistics, Wharton School, University of Pennsylvania
Michael L.Stein, Department of Statistics, University of Chicago
Barry M. Lesht, Environmental Research Division, Argonne National Laboratory
David Schwab, NOAA-GLERL
Dmitry Beletsky, University of Michigan

In this talk, we develop a space-time model combining hydrodynamic models with satellite data. We illustrateour approach with an application to suspended sediment modeling in Lake Michigan.

e-mail: stroud@wharton.upenn.edu

DYNAMIC SPATIAL CHANGE-OF-RESOLUTION MODELING

Gardar Johannesson, Department of Statistics, The Ohio State University
Noel Cressie*, Department of Statistics, The Ohio State University
Hsin-Cheng Huang, Institute of Statistical Science, Academia Sinica

In this talk, we consider the problem of spatial-temporal prediction of global processes using a model that recognizes multiple resolutions in the spatial domain. By combining several small regions into a larger region and several larger regions into an even larger region, and soforth, one can build up a scheme for changing resolutions. This can be represented as a tree, and a spatial model is obtained by assuming an autoregressive model in level of resolution (Chou et al., 1994). Here, optimal spatial-prediction procedures can be shown to be extremely fast. Huang et al. (2002) present a mass-balanced, change-of-resolution Kalman filter that is statistically optimal and aggregation consistent. Similar ideas can be used in the spatial-temporal domain; a vector autoregressive model isassumed at the coarsest resolution and the spatial strucureat each time point is as before. We show that the resulting(spatial-temporal) graph is still a tree with its associated fast Kalman-filter-type prediction algorithm. The benefit of adding the time dimension will be addressed based on data from the Total Ozone Mapping Spectrometer (TOMS) instrument, on the Nimbus-7 satellite.

e-mail: ncressie@stat.ohio-state.edu

EFFICIENT PARAMETERIZATION OF HIGH-DIMENSIONAL SPATIO-TEMPORAL MODELS

Christopher K. Wikle*, Department of Statistics, University of Missouri-Columbia
Bill Xu, Department of Statistics, University of Missouri-Columbia

Many processes in the environmental and biological sciencesexhibit complicated spatio-temporal variability, includingwave propagation and diffusive behavior. These processes are often nonstationary in time and/or space as well asnonseparable. We demonstrate how a hierarchical framework based on integro-difference equations withspatially-varying convolution parameters can lead to efficient parameterzations for such processes. We demonstrate the methodology on the problem of now casting radar reflectivities in convective weather environments and the problem of predicting the spread of an invasive species.

e-mail: wiklec@missouri.edu

3. Applications of High-Dimensional Data Analyses to Microarray Data

DISTANCE WEIGHTED DISCRIMINATION

J. S. Marron*, Department of Statistics, University of North Carolina -Chapel Hill
Michael Todd, School of Operations Research and Industrial Engineering, Cornell University

High Dimension Low Sample Size statistical analysis is becoming increasingly important in a wide range of applied functional data contexts. In such situations, it is seen that the appealing discrimination method called the Support Vector Machine can be improved. The revealing concept is ‘data piling’ at the margin. This leads naturally to the development of ‘Distance Weighted Discrimination, which also is based on modern computationally intensive optimization methods, and seems to give improved ‘generalizability.’

e-mail: marron@email.unc.edu

STUDY OF GENE EXPRESSION: STATISTICS, BIOLOGY, AND MICROARRAYS

Ker-Chau Li*, Department of Statistics, UCLA

Microarrays enable mRNA measurement at the full genome scale. They have been successfully applied to monitor gene activities under various physiological or environmental conditions. Investigations on differential expression between normal and disease tissues, or cell-lines, have led to the identification of genes with great diagnostic and clinic potential. Unlike DNA or protein sequence data, microarray outputs are extremely noisy. Broadly speaking, microarray analysis can be carried out at two levels. At the first level, methods are developed to convert an image feature into a single number that reflects the amount of expression of the gene. This step already requires a great deal of statistical and computer expertise.

The second level of analysis begins with the expression data represented by a matrix of n rows and p columns. Each column represents one condition and each row represents one gene. The ij-th element in the matrix shows the level of expression for gene i under condition j. Typically, n is in the order of 1000-10000 and p can be in the order of ten to several hundreds. So this presents a great challenge for large-scale data analysis. In fact, many multivariate statistical methods have been applied, including those from clustering, classification and dimension reduction. In addition to the various in-house research usage, many sets of massive gene expression data are public-accessible. We demonstrate that the opportunity is great for formulating meaningful statistical problems to guide further biological research. Several examples will be given.

e-mail:

4. Advances in Growth Mixture Modeling for Preventive and Treatment Trials

GROWTH MIXTURE MODELING IN RANDOMIZED TRIALS

Bengt O. Muthen*, UCLA

This talk discusses the use of growth mixture modeling to assess treatment effects in randomized trials. The motivation is a study of depression medication in a double-blind placebo-controlled clinical trial. Growth mixture modeling represents heterogeneity among subjects using a finite mixture random effects model. The methodology allows one to examine different impact of a treatment on subject classes characterized by different types of growth trajectories. The analysis of the depression trial finds two classes, where unlike the non-responder class, the responder class shows a drop in the depression rating. The growth mixture analysis demonstrates that medication effects can be assessed in the presence of placebo-response effects. Furthermore, the results can be used to predict from baseline covariates who is likely to be in the non-responder class. This has design implications in that baseline characteristics can be used to predict who will need a different kind of treatment. Estimation is carried out using maximum-likelihood estimation via the EM algorithm as implemented in the Mplus program. Monte Carlo simulations are used to demonstrate stability of the solution at low sample sizes.

e-mail: bmuthen@ucla.edu

MULTIPLE IMPUTATION AND PSEUDOIMPUTATION FOR MODEL DIAGNOSTICS IN FINITE MIXTURE GROWTH MODELING

C. Hendricks Brown*, University of South Florida, Department of Epidemiology and Biostatistics
College of Public Health
Chen-Pin Wang, University of South Florida, Department of Epidemiology and Biostatistics
College of Public Health

Multiple imputations allow for a general inferential approach to handling incomplete data. This technique is useful when a program exists for handling complete datasets but the corresponding procedure is not readily computable when data are incomplete. Recently, many programs have implemented procedures to maximize the full likelihood, treating the data as Missing at Random. These enhancements now allow us to revisit an earlier, once rejected, version of imputation for model checking. In this method, once the observed data MLE is computed, missing categorical or latent class variables are imputed multiple times assuming that the true parameter values are fixed at the marginal MLE. Model diagnostics can be computed based on these completed datasets. We call this general procedure pseudoimputation since it disregards the additional step of randomly sampling the parameter from the posterior distribution.
Pseudoimputation is advantageous when marginal MLE's can be computed easily. Their asymptotic properties and their use for model diagnostics are the subject of this presentation.

e-mail: hbrown@hsc.usf.edu

BAYESIAN HIERARCHICAL MODELING OF HETEROGENEITY IN MULTIPLE CONTINGENCY TABLES

Getachew A. Dagne*, University of South Florida
Hendricks Brown, University of South Florida
George W. Howe, The George Washington University

Etiologic and intervention research in prevention often rely on microcoded data of dyadic or group interactions to provide evidence of change due to development or treatment. Traditionally these data have been collapsed into small or modest size two-way contingency tables, followed by residual analyses. Such an approach not only limits one's ability to fit models, but also can introduce spurious findings. Instead of treating each dyad's or group's two-way contingency table independently, or collapsing the tables into single aggregate table, it is more efficient to analyze associations in all groups simultaneously using hierarchical models. This article presents a Bayesianhierarchical model to analyze several two-way sequential categorical data with random effects that allow different levels of variation across several events. To illustrate this approach, the authors present an analysis of couples' interaction data from a recent clinical study investigating how couples cope when one partner has become unemployed.

e-mail: gdagne@hsc.usf.edu

WHAT'S NEW IN LATENT CLASS ENUMERATION FOR GENERAL GROWTH MIXTURE MODELS

Katherine E. Masyn*, UCLA

This presentation contributes to the ongoing discussion of latent class enumeration in the area of finite mixture modeling and, more specifically, in growth mixture modeling. There are many theories in social research that postulate the existence of qualitatively different developmental trajectories; consider, for example, the oft-cited "life-course persistent" and "adolescent-limited" antisocial behavior trajectories of Moffit (1993). Growth mixture models offer a method for exploring and substantiating such theories in a longitudinal data analysis. This talk will give a survey of currently used information heuristic methods for assessing mixture order, such as the Bayesian Information Criterion (BIC) and normalized entropy. These heuristics will be compared to inferential methods such as the recently proposed exact parametric likelihood ratio test of Lo, Mendell, and Rubin (2001) and the multivariate skewness and kurtosis tests of Muthén and Asparouhov (2002). The performances of the aforementioned approaches for correctly determining the number of latent classes are evaluated and compared through a series of growth mixture model simulations as well as real data examples.

e-mail: kmasyn@ucla.edu

EXACT MARGINAL LIKELIHOOD AND CONDITIONAL TESTS OF INTERACTIONS INVOLVING LATENT CLASS REGRESSION AND GENERAL GROWTH MIXTURE MODELING.

Klaus Larsen*, Clinical Research Unit, Hvidovre Hospital, Denamrk
C. Hendricks Brown, University of South Florida

Latent class regression and general growth mixture models provide a flexible family of finite mixture models for longitudinal data with multiple types of outcomes. Such models permit (1) covariates to influence latent class membership, (2) classes to affect growth trajectories, and (3) classes to affect a distal outcome such as a diagnosis. These methods are now being used, for example, to identify differential patterns of growth in aggressive behavior in childhood and their long-term relationship with antisocial personality disorder. Such methods are extremely useful in examining short-term and long-term effects of a preventive intervention. In analyzing data from a randomized preventive trial with 15 years of follow-up, we noticed a beneficial intervention impact among the highest risk group. Because the sample sizes in the latent classes were small, we examined two exact methods for testing treatment by latent class interactions. One method is based on a marginal likelihood test while the other modifies the Fisher Exact Test to latent variables. We examine the behavior of both these tests and describe their use with both continuous and discrete outcomes.

e-mail: klaus.larsen@hh.hosp.dk

SAMPLE SIZE CALCULATION AND OPTIMAL DESIGN WITH LOGISTIC REGRESSION

Eugene Demidenko*, Dartmouth Medical School

The Wald test is used to compute the required sample size for studies with logistic regression. The sample size formulae are obtained for the following combinations: binary and continuous (normally distributed) single exposure; four possible combinations of exposure and confounder binary/continuous; binary exposure, confounder and their interaction. Unlike existing formulae, and particularly formulae used in the current commercial software packages, our formulae provide the exact nominal power in large sample. We determine the optimal proportion of cases in a case-control study that minimizes the total sample size under given nominal power. We prove that if the exposure is a risk factor than there should be less number of cases than controls, otherwise there should be more cases than controls. The 50/50 design is optimal only when there is no association between exposure and disease. The optimal design is illustrated for the case-control study where the interaction term is of interest.

e-mail: eugene.demidenko@dartmouth.edu

EXPERIMENTAL DESIGN AND SAMPLE SIZE DETERMINATION FOR TESTING SYNERGISM

Ming T. Tan*, University of Maryland Greenebaum Cancer Center
Hongbin Fang, University of Maryland Greenebaum Cancer Center
Guoliang Tian, University of Maryland Greenebaum Cancer Center
Peter Houghton, St Jude Children’s Research Hospital

In anti-cancer drug development, the combined use of two drugs is an important strategy to achieve greater therapeutic success. Often combination studies are performed in animal (mostly mice) models before clinical trials are conducted. These experiments on mice are costly. However, experimental designs and sample size derivations for the joint action of drugs are not currently available except for a few cases where strong model assumptions are made (e.g., Abdelbasit and Plackett, 1982, Tallarida et al., 1999). We propose a novel nonparametric model that does not impose such strong assumptions on the joint action. We then propose an experimental design for the joint action using uniform measure in this nonparametric model. This design is optimal in the sense that it reduces the variability in modeling synergy while allocates the doses to minimize the number of experimental units and to extract maximum information on the joint action of the compounds. Based on this design, we propose a robust F-test to detect the simple similar action of two compounds and a method to determine sample sizes that are economically feasible. We illustrate the method with a study of the joint action of two new anti-cancer agents: temozolomide and irinotecan.

e-mail: mtan@umm.edu

BAYESIAN SAMPLE SIZE CALCULATIONS FOR CASE-CONTROL STUDIES

Cyr Emile M'lan*, Hospital for Sick Children

One of the most important statistical issues at the planning stage of a case-control study is the choice of sample size. Sample size determination for the odds ratio has been investigated from a frequentist viewpoint. While most of the proposed methods have been based on power, it is well known that high power does not necessarily guarantee accurate estimation of important parameters. Therefore, if one chooses to analyze a study by interval estimation rather than p-values, the design of the study should reflect this choice. In addition, there is an increasing literature on the advantages of Bayesian sample size determination, which better incorporates prior information into the calculations, and more fully accounts for the uncertainty of the eventual data which will be collected.

In this talk, we show how sample size determination for estimating the odds ratio can be addressed within the Bayesian paradigm. The criteria investigated is the average length criterion (ALC) by Joseph et al. 1995. Basically, this criteria proposes that the sample size be selected that guarantees a pre-specified length for a marginal posterior credible interval of predetermined coverage, averaged over the predictive distribution of the data. The solution, while easy to define, is technically challenging to carry out in practice. We discuss three different methods for finding the optimal sample size, including exact, approximate, and Monte Carlo. We compare the sample sizes derived from this criterion to those from frequentist power and confidence interval methods.

e-mail: mlan@bioinfo.sickkids.on.ca

SAMPLE SIZE DETERMINATION FOR COMPARING SEVERAL SURVIVAL CURVES UNDER PROPORTIONAL HAZARDS MODEL AND UNEQUAL ALLOCATION

Susan Halabi*, Duke University
Bahadur Singh, Duke University and and Lineberger Comprehensive Cancer Center-UNC

The large sample properties of the Logrank test under a sequence of alternative hypotheses converging towards the null hypothesis values are examined when three or more survival curves are compared assuming the proportional hazards and unequal sample sizes allocation. The results presented here generalizes the result of Sang and Anderson (1995). The sample size formula is presented assuming a
prespecified 80% power and unequal allocation of sample sizes in several treatment groups. Under the proportional hazards setting, three cases are considered: 1) exponential failures- exponential censoring, 2) exponential failures-uniform censoring and 3) Weibull failures-uniform censoring. In all the cases, it is assumed that the censoring distribution is exactly the same across all of the treatment groups. There is very close agreement between the exact and simulated powers. The powers are also computed using two-moment and four-moment power series approximations. There is close agreement with the approximate powers with the exact power. In addition, the sample size formula for the stratified Logrank is also presented for unequal allocation.

e-mail: susan.halabi@duke.edu

A SAMPLE SIZE METHOD FOR TESTING THE RATIO OF TWO MEAN LIFETIMES: APPLICATION TO AGING INTERVENTION STUDY

Chengjie Xiong*, Division of Biostatistics, Washington University in St. Louis
J. P.Miller, Division of Biostatistics, Washington University in St. Louis
Yan Yan, Division of Biostatistics, Washington University in St. Louis

Sample size determination is a very important part of planning for clinical trials. We present a method of computing sample sizes required to achieve adequate statistical power for testing the ratio of two mean lifetimes when both samples are subject to type II censoring. Our approach is based on the location-scale family of log-transformed lifetime distributions as compared to that based on the log-rank test and the family of proportional hazards. The effect of thresholds, or guarantee times' on sample size determination is also discussed. The Aging Intervention Testing Program from National Institute on Aging is used to demonstrate our results.

e-mail: chengjie@wubios.wustl.edu

---

6. Advances in Mixed Models

JOINT MODELLING OF MULTIVARIATE LONGITUDINAL PROFILES: PITFALLS OF THE RANDOM-EFFECTS APPROACH

Steffen Fieuws*, Biostatistical Centre, K.U.Leuven (Belgium)
Geert Verbeke, Biostatistical Centre, K.U.Leuven (Belgium)

Multivariate longitudinal data arise when a set of different responses on the same unit are measured repeatedly over time. An example of a research question for such data is how the evolution of one response is related to the evolution of another response ('association of the evolutions'). A seemingly related, but different question is how the association between responses evolves over time ('evolution of the association'). To answer such research questions a joint modelling strategy is needed.

A flexible strategy, which is used frequently in recent literature, is to model the association between the different responses using random effects. This approach has many advantages and is applicable in a wide variety of situations. More important, the method allows to answer both research questions simultaneously. However, in this presentation it will be shown that this approach implies an association structure which isn't necessarily valid for the observed data. The problem will be illustrated using hearing treshold measurements from the Baltimore Longitudinal Study of Aging.

e-mail: steffen.fieuws@med.kuleuven.ac.be

PREDICTING RANDOM EFFECTS OF REALIZED CLUSTERS IN FINITE POPULATIONS

Edward J. Stanek III*, Department of Biostatistics and Epidemiology, UMASS
Julio D.Singer, Department of Statistics, University of São Paulo, Brazil

Characteristics of individuals, such as level of physical activity or dietary intake, are of interest. The characteristics are often defined as an average over a finite time period, (month or year), with observations made repeatedly over shorter time periods (days). Typically, study subjects are selected via random sampling from a population, so the design corresponds to 2-stage cluster sampling. Mixed models may be fit to such data, treating subjects as clusters, with the characteristics of a realized individual predicted by the best linear unbiased predictor (BLUP). The BLUP does not account for the finite time period that defines the parameter. We develop predictors in this context that account for the sampling fraction. The predictor is the weighted average of the realized subject's sample mean and the BLUP, with some modification in the definition of the variance parameters; the weight is the second stage sampling fraction. The development uses a prediction based approach in the context of a two stage random permutation model. We illustrate these results, and extend them to include measurement error in balanced designs.

e-mail: stanek@schoolph.umass.edu

AN ASSESSMENT OF HETEROSKEDASTIC T-ERROR LINEAR MIXED MODELS FOR THE ANALYSIS OF FIELD DATA COLLECTED FROM DIVERSE ENVIRONMENTS

Kadir Kizilkaya, Michigan State University
Robert J.Tempelman*, Michigan State University

Animal performance data (e.g. weights, milk yields) is typically collected from diverse livestock production systems with data quality often compromised by the occurrence of recording error, preferential treatment and/or the effect of injury or disease. Heteroskedastic error linear mixed effects models have been increasingly advocated for the analysis of such data, based on a structural log link model specification for residual variances as a function of fixed (e.g. age) and random effects (e.g. herds). However, the extent of this residual heteroskedasticity may be partially influenced by outliers. A heteroskedastic t-error linear mixed effects model is proposed for the analysis of such data, using a Bayesian approach for inference. A simulation study was used to validate the use of the deviance information criterion to choose between different specifications with respect to heteroskedasticity and heaviness of tails of the residual density. An application to birth weights in beef cattle indicated that the heteroskedastic t-error linear mixed effects model fitted the data much better than either a heteroskedastic normal error or homoskedastic t-error model.

e-mail: tempelma@msu.edu

RANDOM COEFFICIENT DYNAMIC MODELS FOR LONGITUDINAL DATA WHEN THE BASELINE VARIABLE IS CORRELATED WITH RANDOM EFFECTS

Haihong Li*, Frontier Science and Technology Research Foundation, Inc.
Jianhua Z.Huang, University of Pennsylvania
Hulin Wu, Frontier Science and Technology Research Foundation, Inc.

We consider autoregressive models for longitudinal data where the coefficients are random. Rahiala (1999, Biometrika) showed that the point estimates can be obtained by treating the lagged variable as if it is an ordinary covariate variable.
However, his results were based on the assumption that the initial response variable values are independent of the random effects. In applications this assumption may not always hold. We discuss different cases when the initial observations are correlated with the random coefficients. Estimation procedures and asymptotic results are established. The procedure is applied on an AIDS clinical trial data
for the purpose of dynamic prediction.

e-mail: hli@fstrf.dfci.harvard.edu

SOME USEFUL DIAGNOSTICS FOR RANDOM EFFECTS MODELS WITH UNWEIGHTED AND WEIGHTED DATA

Julia L. Bienias*, Rush-Presbyterian-St. Luke's Medical Center
Charles B.Hall, Albert Einstein College of Medicine
Woojeong Bang, Rush-Presbyterian-St. Luke's Medical Center

Mixed models (fixed plus random effects) are becoming increasingly popular in many application areas. We present a set of useful diagnostics for checking model assumptions and model fit for these models, with a particular emphasis on the growth curve-type models of Laird & Ware (1982). We illustrate these diagnostics with examples. In addition, because of the growing use of complex sampling designs in longitudinal studies, there is a need for diagnostics that can be used with data with unequal probabilities of selection. In particular, the usual diagnostics, both graphical and quantitative, might give misleading results with weighted data because of the unequal sampling weights. Thus, we also discuss our diagnostics in the context of weighted data.

e-mail: jbienias@rush.edu

RANDOM EFFECTS SELECTION IN LINEAR MIXED MODELS

Zhen Chen*, Biostatistics Branch, National Institute of Environmental Health Sciences
David B.Dunson, Biostatistics Branch, National Institute of Environmental Health Sciences

We address the important practical problem of how to select the random effects component in a linear mixed model. A hierarchical Bayesian model is used to identify any random effect with 0 variance. The proposed approach reparameterizes the mixed model so that functions of the covariance parameters of the random effects distribution are incorporated as regression coefficients on standard normal latent variables. We allow random effects to effectively drop out of the model by choosing mixture priors with point mass at zero for the random effects variances. Due to the reparameterization, the model enjoys a conditionally linear structure that facilitates the use of normal conjugate priors. We demonstrate that posterior computation can proceed via a simple and efficient Markov chain Monte Carlo algorithm. The methods are illustrated using simulated data and real data from a study relating prenatal exposure to polychlorinated biphenyls and psychomotor development of children

e-mail: chen14@niehs.nih.gov

TESTING DEPENDENT OBSERVATION TIMES FOR LONGITUDINAL DATA

Yangjin Kim, University of Missouri, Columbia

We consider the situation when observation time process depends on response variables. Most of existing methods for longitudinal data assume that observation time process and response variable process operate independently. We develop a test for this assumption. To construct the test, we consider the joint modeling of response variable and observation times and the derivation of the likelihood function. Specifically, for the response variable, random effects models (Laird and Ware, 1982) will be applied as in most of longitudinal studies and we will model the observation times by naturally treating them as realizations of counting processes with the proportional intensity functions (Andersen et al., 1993). The dependence of the observation process on the response variable is modeled through some random effects shared by both, which naturally results in a dependence test. To estimate unknown parameters, we propose a two step estimation procedure, which can be easily implemented using many existing statistical packages. Then we apply the proposed method to two real exams.

e-mail: yjkim@stat.missouri.edu

ORTHOGONALIZED RESIDUALS FOR ESTIMATION OF MARGINALLY SPECIFIED ASSOCIATION PARAMETERS IN MULTIVARIATE BINARY DATA

Richard C. Zink*, University of North Carolina at Chapel Hill, Department of Biostatistics
Bahjat F.Qaqish, University of North Carolina at Chapel Hill, Department of Biostatistics

We explore marginal regression models for correlated binary responses where estimation of the association structure is the primary focus. A new estimating function approach based on orthogonalized residuals is proposed. A special case of the proposed procedure allows a new representation of the alternating logistic regressions method of Carey et al. (1993) through marginal residuals. The connections between second-order generalized estimating equations, alternating logistic regressions, pseudo-likelihood and other methods are explored. Efficiency comparisons are presented, with emphasis on variable cluster size and on the role of higher-order assumptions. The new method is illustrated with an analysis of data on health-maintenance visits.

e-mail: rzink@bios.unc.edu

WEIGHTED ESTIMATING EQUATIONS FOR SEMIPARAMETRIC TRANSFORMATION MODELS WITH CENSORED DATA FROM A CASE-COHORT DESIGN

Lan Kong*, Department of Biostatistics, University of North Carolina, Chapel Hill
Jianwen Cai, Department of Biostatistics, University of North Carolina, Chapel Hill
Pranab Kumar Sen, Department of Biostatistics, University of North Carolina, Chapel Hill

The case-cohort design was originally introduced by Prentice (1986) to reduce the cost in large cohort studies of a rare disease. Under the case-cohort design, covariates are assembled only for a random sample of the entire cohort and all the cases. In this paper, semiparametric transformation models are considered for failure time data from a case-cohort design. Weighted estimating equations are proposed for the estimation of regression parameters. Asymptotic properties are derived for the parameter estimators using U-statistics theory, asymptotic results for simple random sampling without replacement from a finite population, and martingale convergence results. The finite sample properties of the proposed estimators, as well as the efficiency relative to the full cohort estimators, are assessed via simulation studies. A case-cohort data set from the Atherosclerosis Risk in Communities study is used to illustrate the estimating procedure.

e-mail: lkong@bios.unc.edu

MEAN ESTIMATING EQUATIONS APPROACH TO ANALYSING CLUSTER-CORRELATED DATA WITH NONIGNORABLE CLUSTER SIZES

Emmanuel Benhin*, Statistics Canada
Jon N. K.Rao, Carleton University, Canada
Alistair J. Scott, Department of Statistics, University of Auckland, New Zealand

Most Classical methods for analyzing cluster-correlated biological data implicitly assume that the cluster sizes are ignorable. When this assumption fails, these methods may be invalid. Within-cluster resampling (Hoffman et al. 2001) provides a simple but computationally intensive method which is valid whether cluster sizes are ignorable or nonignorable. This method, however, can lead to unstable parameter estimates or variance estimates. We propose two methods to rectify these shortcomings: combined estimating equations and mean estimating equations. We present some theory of the proposed methods and explore potential areas of application. Simulation studies are presented to assess the performance of the proposed methods.

e-mail: emmanuel.benhin@statcan.ca

MODIFIED GEE AND GOODNESS-OF-MARGINAL-FIT TEST WITH CORRELATED BINARY RESPONSES FOR CONTINGENCY TABLES

Ji-Hyun Lee*, Department of Biostistics, University of North Carolina at Chapel Hill
Bahjat F.Qaqish, Department of Biostistics, University of North Carolina at Chapel Hill

Suppose that in 2x2 contingency tables, the two variables are correlated and analysis only centers on the marginal cells. The relevant model is for the marginal mean, population average, of the response. In this paper a modified generalized estimating equation (GEE) is developed. The modified estimating equation is obtained by using a multinomial covariance structure as a "working covariance matrix" termed by Liang and Zeger (1986). As a global test statistic for detecting model deviation, the methods for uncorrelated binary regression models may be inappropriate for models fitted with a GEE approach. Here, we propose a Goodness-of-Marginal-Fit (GOMF) test statistic based on the modified GEE for the correlated binary data. The GOMF is evaluated theoretically. We illustrate the modified GEEs for the parameters and the proposed GOMF test with two examples. The power of the proposed test statistic is assessed by a variety of simulation.

e-mail: jhlee@bios.unc.edu

EXTENDED ESTIMATING EQUATIONS FOR LINK AND VARIANCE FUNCTION PARAMETERS IN GENERALIZED LINEAR MODELS

Anirban Basu*, Harris School of Public Policy Studies, University of Chicago
Paul J.Rathouz, Department of Health Studies, University of Chicago

We propose an extension to the estimating equations in generalized linear models to estimate parameters in the link function and variance structure simultaneously with regression coefficients. Rather than focusing on the regression coefficients, the purpose of these models is consistent estimation of (i) the mean of the outcome as a function of a set of covariates, and (ii) the partial derivative of the mean function with respect to any covariate. This second parameter is often referred to as the marginal effect by econometricians. The proposed estimation algorithm not only helps to identify a correct link function and to suggest an underlying distribution for a specific application but also serves as a robust estimator when no specific distribution for the outcome measure can be identified. Using Monte-Carlo simulations, we show that the resulting parameter estimators are consistent. The method is illustrated with an analysis of inpatient expenditure data for the Hospitalist study.

e-mail: abasu@uchicago.edu

SEMIPARAMETRIC RANK REGRESSION IN STABILITY ANALYSIS

Annpey Pong*, Forest Laboratories, Inc.
Ying Q.Chen, University of California at Berkeley
Biao Xing, University of California at Berkeley

Stability data are often collected for analysis to determine the shelf-life of certain characteristics of a pharmaceutical product, e.g., a drug's potency over time. Statistical approaches such as the linear regression models are considered as appropriate to analyze stability data. However, most of the previous regression models in research and practice rely heavily on the parametric assumptions, such as normality of the quantitative characteristics. In this article, we propose rank regression procedures when the linear regression models are semiparametric with unspecified error structure. The new procedure will be studied with or without batch-to-batch variation. Numerical studies including Monte Carlo simulations and practical examples are demonstrated with the proposed procedures as well.

e-mail: annpey.pong@frx.com

GENERALIZED ESTIMATING EQUATIONS FOR MIXED-EFFECTS MODELS

Wei Wang*, Department of Epidemiology and Biostatistics, University of South Florida
Yiliang Zhu, Department of Epidemiology and Biostatistics, University of South Florida

Mixed-effects models are increasingly popular in analyzing spatially or temporally clustered data arising from diverse research fields. When the models are nonlinear conditioning on the unobserved random effects, the marginal distribution of the outcome is typically unavailable in closed-form. Consequently numerical algorithms such as Laplace approximation and the EM or its variation are used iteratively between approximation to and maximization of the marginal likelihood. However, it is often difficult or uncertain to specify a distribution for the outcomes or random effects. We propose generalized estimating equations that require only the specification of moments for the data and random effects. The method unifies conditional and marginal GEEs; includes as a special case the score equations for quadratic exponential family under the first order Laplace approximation to the marginal likelihood; and is applicable to generalized linear or nonlinear mixed-effects models. We show that the conditional and marginal GEEs are asymptotically equivalent in terms of both bias and efficiency to the first order. We also consider the small sample performance of the GEEs with respect to the order of the moments utilized and penalty adjustment to random effects estimation.

e-mail: wwang@hsc.usf.edu

8 Recurrent Event Data, Frailty Models, and Survival Analysis

MODELS AND BAYESIAN ANALYSIS OF RECURRENT EVENTS DATA

Debajyoti Sinha*, Biometry & Epidemiology, Medical University of SC

We consider models and Bayesian methods for recurrent events data when the termination time for each subject may depend on the point process of the recurrent events via a frailty variable. This article develops a class of fully specified stochastic models which allow negative association between the risk of termination and the rate of recurrent events. We derive several properties of our model and other existing models to provide a comparison of our model with competing models. We also explore the relationship of our model with existing models for medical costs data. We develop efficient Markov chain Monte Carlo algorithms for sampling from the posterior distribution of the parameters when the data from each subject is recorded via scheduled clinic visits and is subject to right censoring. We demonstrate the usefulness of our new models and methodologies through the reanalysis of a dataset from a clinical trial.

e-mail: sinhad@musc.edu

BAYESIAN INFERENCE FOR PVF FRAILTY MODELS

Madhuja Mallick*, University of Connecticut
Nalini Ravishanker, University of Connecticut

This talk describes inference for multivariate lifetimes data using a conditional proportional hazards model with a power variance family (PVF) shared frailty distribution and a Weibull baseline hazard. Although the PVF distribution is conceptually simple, inference for the PVF frailty model is complicated due to the lack of a closed form expression for the density function of the PVF random variable. Deriving the density function of the PVF variable as a tilted positive stable density, we employ the Bayesian approach using Markov chain Monte Carlo methods for carrying out inference. We also describe useful dependence measures for the model and compare them to the positive stable frailty model. The approach is illustrated using data involving recurrent infections due to insertion of a catheter in patients on portable dialysis machines.

e-mail: madhuja@stat.uconn.edu

ANALYSIS OF FRAILTY SURVIVAL MODELS USING POISSON VARIANCE STRUCTURES

Shibao Feng*, Department of Biostatistics, University of Michigan at Ann Arbor
Robert A.Wolfe, Department of Biostatistics, University of Michigan at Ann Arbor

The likelihood functions of both parametric (e.g., with piecewise constant baseline hazard) and semi-parametric multivariate frailty models are shown to be proportional to the likelihood functions of a class of mixed Poisson regression models. For multivariate lognormal frailty models, the penalized quasi-likelihood (PQL) method can be applied as the inference procedure for mixed Poisson regression models. Thus, a rich variety of random effect structures can be modeled for survival analysis. Simulation studies show that mixed Poisson regression models using PQL methods for inference perform well for both the fixed and random effect parameters of multivariate lognormal frailty models. The procedure is illustrated with a national kidney transplantation dataset.

e-mail: shibao@umich.edu

ADJUSTING FOR CLINIC EFFECTS WHEN ESTIMATING A TREATMENT EFFECT: STRATIFIED, CLUSTERED, AND FRAILTY MODELS

Lynn E. Eberly*, University of Minnesota, School of Public Health,Division of Biostatistics
Lingfeng Yang, University of Pennsylvania, School of Medicine, Division of Biostatistics and Epidemiology

The proportional hazards model, which is used extensively inclinical trials and elsewhere, assumes a common baseline hazard for the event of interest for all observations. However, in a multi-center clinical trial, this is likely an unreasonable assumption, for example when different clinics serve different types of study participant populations. We compare four approaches used to address this problem: no adjustment, and stratified, clustered, and frailty proportional hazards models. Comparisons are based on simulations with varying numbers of clinics, numbers of participants per clinic, extent of censoring, and magnitudes of a randomized two-arm treatment effect. We examine type I error rate, power, and bias for the treatment coefficient. Results show that variability in the clinic specific treatment effects can increase the estimated treatment effect. We also see that the trade-off between number of clinics and number of participants per clinic (for a fixed total sample size) impacts the treatment effect to a similar extent as the clinic-to-clinic variability. The results are further explored using data from several completed clinicaltrials.

e-mail: lynn@biostat.umn.edu

TWO-SAMPLE TEST FOR THE DISTRIBUTION OF AGE AT ONSET FROM LIFE-TIME INCIDENCE CASE-CONTROL DATA

Emilia Bagiella*, Mailman School of Public Health, Columbia University

Case-control studies are a powerful tool for discovering the association between risk factors and lifetime probability of disease, especially if the disease is rare or occurs late in life. The usual survival techniques for estimating the marginal probability of disease onset, like the Kaplan-Meier estimator, are well suited for prospective studies but give biased estimates when applied to case-control data. We present an estimator of the marginal distribution of age at onset for lifetime incidence case-control data. At the core of the estimator are the Lynden-Bell estimator for truncated data and a generalization of the usual approach for case-control data of treating the data as if it came cross-sectionally from a multiplicative intercept model.

We also present a two-sample test to compare two distributions of age at onset for two risk groups. The test is based on the integrated difference between the two survival distributions and we estimate its variance using a Jackknife approach. We apply the methods to data on colorectal polyps obtained from a case-control study of patients undergoing colonoscopy.

e-mail: bagiella@biostat.columbia.edu

PROPORTION OF TREATMENT EFFECT (PTE) EXPLAINED BY A SURROGATE MARKER

Cong chen*, Merck Research Labs
Hongwei Wang, Rutgers Univesity
Steven M. Snapinn, Merck Research Labs

In time-varying covariate analysis of clinical survival data, it is often of interest to estimate the proportion of treatment effect (PTE), along with its confidence intervals, explained by a surrogate marker. The conventional procedure for such an analysis fits data into two working models separately to estimate the treatment effects before and after adjustment of the covariate. The construction of confidence intervals for the PTE under the conventional procedure is very computationally demanding. To overcome this problem, we propose a new procedure to simplify the computation. Under the new procedure, the treatment effects before and after adjustment of the covariate are simultaneously estimated from a single model. The new procedure not only helps simplify the construction of confidence intervals, but also can be effectively applied to multiple-covariates models for the decomposition of overall treatment effect and for the comparison of PTE among several surrogate markers. The new procedure is applied to the motivating data example from the LIFE study, and demonstrates flexibility that the conventional procedure currently lacks.

e-mail: cong_chen@merck.com

CONFIDENCE INTERVAL ESTIMATION FOR CHD RISK PREDICTION BY DIFFERENT SURVIVAL MODELS

Usha S. Govindarajulu*, Boston University
Ralph B.D'Agostino, Boston University

Keaven Andersen (1990, American Heart Journal, 121, 293-298) described an accelerated failure time model to develop prediction equations and also confidence intervals around these predictions for different cardiovascular disease endpoints. The model that Andersen presented was parametric but contained a non-proportionality component and therefore was superior to a Cox proportional hazards regression if the baseline proportional hazards assumption was not met. Using Andersen's ideas for the confidence interval estimation around the predicted probability, two other macros were developed using the Cox model and the Weibull model. This article reviews these three methods, describes an implementation in SAS to estimate these models, and applies these models to data from the Framingham Heart Study.

e-mail: usha@math.bu.edu

CHRONOLOGICAL EVENT MODELING FOR SCREENING MAMMOGRAPHY

Prashni Paliwal*, Department of Statistics, University of Connecticut.
Alan E.Gelfand, Institute of Statistics and Decision Sciences at Duke University.
Linn Abraham, Center of Health Studies, Group Health Cooperative of Puget Sound
William E. BArlow, Center of Health Studies, Group Health Cooperative of Puget Sound
Joann Elmore, Group Health Cooperative and School of Medicine at the University of Washington

Screening mammography is a widely used method for breast cancer detection. We propose to model the process in a chronological fashion. That is, we envision an initial assessment, a follow up assessment if the initial one is positive and, eventually, a determination of whether cancer was present or not. A model can be built at each stage reflecting effects due to patient characteristics, to the facility where mammogram was performed and to the radiologist reading the mammogram. Since assessment is not perfectly associated with outcome, familiar rates of agreement and disagreement are of interest. These rates can be investigated for risk factors we wish to control and their different levels. A motivating dataset is extracted from the records of the Group Health Cooperative in Seattle, WA. A Bayesian framework is adopted for inference and an analysis of the dataset is presented.

e-mail: prashni@mailcity.com

UNDERSTANDING THE FACTORS UNDERLYING DISPARITIES IN CANCER SCREENING RATES BETWEEN RACE/ETHNIC GROUPS: THE PETERS-BELSON APPROACH

Sowmya R. Rao*, National Cancer Institute, NIH, DHHS
Barry I.Graubard, National Cancer Institute, NIH, DHHS
Joseph L. Gastwirth, The George Washington University

Cancer screening rates vary substantially by race/ethnicity and identifying factors that contribute to this disparity between the minority groups and the white majority should aid in designing successful programs. The traditional approach for examining the role of race/ethnicity is to include a categorical variable, indicating minority status, in a regression-type model, whose coefficient estimates this effect. We applied the Peters-Belson (PB) approach, used in wage discrimination studies, to analyze disparities in cancer screening rates between genders as well as different race/ethnic groups from the 1998 National Health Interview Survey (NHIS), and to decompose the difference into a component due to differences in the covariate values in the two groups and a residual difference. Regression model was estimated accounting for the complex sample design. Variances were estimated by the jackknife method where a single primary sampling unit was considered as the deleted group and compared to analytic variances derived from Taylor linearization. Thirteen percent of the 7% observed disparity for colorectal screening, and less than one percent of the 8% observed disparity for digital rectal examination between men and women, was explained by covariate differences. We also present results for cancer screening differences among the different race/ethnic groups by sex.

e-mail: raos@mail.nih.gov

ESTIMATING SENSITIVITY AND SPECIFICTY FROM REPEATED SCREENING TESTS

Michael K. Parides*, Columbia University, Department of Biostatistics
Emilia Bagiella, Columbia University, Department of Biostatistics

When patients are screened more than once to ascertain their disease status calculating sensitivity and specificity by assuming that all measurements are independent yields unbiased estimates of the screening parameters but standard errors of the estimates will, in general, be overestimated. We introduce a method for estimating sensitivity and specificity in this setting considering the screening parameters to be random variables with arbitrary distributions. Our method bases estimation on an adjacent logits polytomous logistic regression model. This is a smoothed linear model for a generalized logit with the observed frequency of positive (or negative) test results and the number of tests as independent variables. This method does not require specification of the prior distribution of the screening parameters. Simulation results indicate that the proposed method produces unbiased estimates and substantially improves efficiency compared to naive estimates.

e-mail: parides@columbia.edu

BAYESIAN SEQUENTIAL DESIGN FOR MULTIPLE DISCRETE OUTCOMES WITH APPLICATION TO CONTINUOUS DRUG SCREENING

Peter Mueller, Dept. of Biostatistics, University of Texas M.D. Anderson Cancer Center
Gary L.Rosner, Dept. of Biostatistics, University of Texas M.D. Anderson Cancer Center
Roberto Carta*, Dept. of Biostatistics, University of Texas M.D. Anderson Cancer Center

At large institutions dedicated to clinical research in cancer, such as the University of Texas M.D. Anderson Cancer Center, a large number of new agents or new combinations of anticancer agents undergo concurrent evaluation for activity. The process is typically carried out through separate phase II studies with only informal learning carried out between studies—even if the studies draw patients with similar disease characteristics.

By sharing as much information as possible between new agents, our model generalizes previous statistical approaches in three important ways. First, we generalize the decision space by allowing for randomly many competing treatments at each time, and by allowing a full sequential design. Second, the model accommodate the diversity of study endpoints and sampling models encountered in cancer trials, as well as combinations of endpoints arising when combining phase I and II clinical trials. Third, we formally incorporate learning rather than combining the information across several related studies in an informal manner.

We used a simulation based approach to optimal sequential design using a dual strategy of constrained action space and forward simulation.

e-mail: roberto@odin.mdacc.tmc.edu

COMBINING BIOMARKERS THROUGH MONOTONE NON-PARAMETRIC REGRESSION

Yue Wang*, Department of Biostatistics, University of Michigan
Jeremy M. G.Taylor, Department of Biostatistics, University of Michigan

Combinations of biomarkers are likely to be more effective for the detection of disease than a single biomarker. In addition, it is often reasonable to assume that the risk of disease changes smoothly as the marker values change and the change in risk is monotone with respect to each marker. In this talk, we present two non-parametric regression methods to estimate the probability of disease as a smooth monotone function of the biomarkers. In the first approach, a set of linear inequality constraints are introduced to mimic the monotonicity. The risk function is estimated by maximizing the penalized likelihood subject to the constraints. In the second approach, tensor-product B-splines are used to approximate the risk function. Through re-parameterization, monotonicity is imposed by introducing simple non-negative bounds on the coefficients. Data from a pancreatic cancer study with two serum markers is used for illustration.

e-mail: wangyue@umich.edu

10. Effects of Technological Advances on Introductory Statistics Courses

HOW TECHNOLOGY ALLOWS DATA TO GUIDE THE STORYLINE OF STATISTICAL STUDIES

Christine A. Franklin*, University of Georgia

Technology has changed the way we teach introductory statistics. The traditional ‘plug and chug’ course is outdated. Computers and statistical calculators easily perform these calculations and in turn, allow visual exploration of data. Simulation on computers/calculators allows early exposure to concepts of inference without the coverage of formal probability. Technology provides the tools to go beyond finding numerical summaries and lets the data guide the storyline or the context of the statistical study for making meaningful interpretations about the question(s) of interest. Computers/calculators mediate between the data and the set of tools at a statistician’s disposal. The type of software used influences the storyline of the analysis. Students of statistics need to practice using computers to analyze real data in realistic situations. They need practice choosing and applying the best tools to obtain information for using sound statistical reasoning.

e-mail: chris@stat.uga.edu

USING TECHNOLOGY TO INVESTIGATE STATISTICAL CONCEPTS

Allan J. Rossman, Cal Poly- San Luis Obispo
Beth L.Chance, Cal Poly- San Luis Obispo

We present an overview of how technology can facilitate the exploration of statistical concepts by students in introductory statistics courses. Some of the concepts that we discuss include resistance, influence, randomization, sampling, sampling distribution, confidence, significance, and robustness. Software tools that we illustrate include Java applets and the educational software package Fathom, as well as mainstream statistics packages such as Minitab.

e-mail: arossman@calpoly.edu

WEBSTAT 3.0: WEB-BASED SOFTWARE FOR STATISTICS EDUCATION

R. Webster West*, University of South Carolina

WebStat 3.0 is the latest version of a Web-based data analysis package that has been developed for application in statistical education. The software reads a variety of data formats and features interactive graphics, simulation capabilities as well as variety of techniques for partitioning and categorizing data. The new features of this software will be demonstrated and its implementation into introductory statistics courses will be discussed.

e-mail:

AN ONLINE MULTIMEDIA TEXTBOOK IN STATISTICS

David M. Lane*, Rice University, Departments of Psychology and Statistics

Students differ in their preferences for textbook styles. Some prefer textbooks which are concise and to the point, not caring much for motivational material or examples of practical applications. Other students find the motivational material fascinating. An author of a textbook is therefore faced with a dilemma since it is impossible to please both types of students. Since there is no practical limit to the number of pages available to a web-based textbook, it is possible to allow students to choose whether they want to view a concise presentation or one containing motivational material. In the 'Online Multimedia Textbook in Statistics,' students can view sections of the book in three ways: standard mode, multimedia mode, and condensed mode. The content in the standard mode contains many examples; it also contains interactive simulations designed to make abstract concepts more concrete. The multimedia mode contains an auditorally-presented lecture accompanied by visual 'slides.' The condensed mode contains just the essential material without examples. This project is still under development and can be found at http://psych.rice.edu/online_stat/
.

e-mail: lane@rice.edu

11. Recent Developments in Clustering and Mixtures with Application to Spatial Models and Image Analysis

PARTIAL MIXTURE ESTIMATION WITH APPLICATION TO CLUSTERING

David W. Scott*, Rice University
Chad A.Shaw, Baylor College of Medicine

The use of density estimation to find clusters in multivariate data can take several forms. Nonparametric approaches may form high-density regions or simply locate sample modes, as in the mode tree. A semiparametric approach is to fit a mixture model and associate each component with a different cluster. Here we describe a hybrid approach, in which we fit a mixture model using a nonparametric criterion. Use of the nonparametric criterion permits local estimation of individual mixture locations. From these estimates, a tree of modes may be formed and tested graphically for weight of evidence.

We use this approach to examine transcriptional patterns in cDNA microarray experiments using laboratory preparations of Dictyostelium discoideum, which is a simple eukaryote that undergoes aggregation and celltype differentiation when starved. This organism is a model system for studying chemotactic aggregation and multicellular development. The search for modes ranges from 5 to 13 dimensions.

e-mail: scottdw@rice.edu

BAYES COMPUTATIONS FOR POISSON AND MARKED SPATIAL PROCESSES

Hemant Ishwaran*, Cleveland Clinic Foundation,
Department of Biostatistics

There now exists a wide range of simple and effective Monte Carlo computational procedures that can be used to fit Dirichlet process mixture (DPM) models. Such models are used in many kinds of nonparametric and semiparametric problems but this typically excludes Poisson spatial process problems which are thought to be unrelated. As a consequence, the development of computational algorithms for Bayesian Poisson spatial processes has gone on in isolation of the rich computational area for DPM's when it could benefit greatly from the simplicity and power of these methods. In this talk I will discuss how to connect Poisson spatial processes to DPM's through the use of weighted gamma process priors. Examples motivating Poisson processes and marked processes will be given. Some details outlining computational algorithms will be presented.

e-mail: ishwaran@bio.ri.ccf.org

ROBUST CLUSTERING METHODS AND VISUALIZATION BASED ON DATA DEPTH

Rebecka Jornsten, Rutgers University
Yehuda Vardi*, Rutgers University
Cun-hui Zhang, Rutgers University

Robust vector clustering methods are developed based on a modified Weiszfeld algorithm for the L1 multivariate median and an associated new concept for data-depth. Multivariate medians are used to represent clusters, while data depth are used to identify nuclei of clusters and outliers. Model selection and visualization tools are developed based on within-cluster data-depth and data-depth with respect to competing clusters. The methods are applied to real-life and simulated data sets to illustrate their performance.

e-mail: vardi@stat.rutgers.edu

SEGMENTING MAGNETIC RESONANCE IMAGES VIA HIERARCHICAL MIXTURE MODELLING

Carey E. Priebe*, Center for Imaging Science, Johns Hopkins University
Michael I.Miller, Center for Imaging Science, Johns Hopkins University
J. Tilak Ratnanather, Center for Imaging Science, Johns Hopkins University

We present a statistically innovative as well as scientifically and practically relevant method for automatically segmenting magnetic resonance images using hierarchical mixture models. Our method is a general tool for automated cortical analysis which promises to contribute substantially to the science of neuropsychiatry. We demonstrate that our method out-performs competing approaches on various magnetic resonance brain imagery segmentation tasks.

e-mail: cep@jhu.edu

12. Statistics in Genetics

ROBUST SINGULAR VALUE DECOMPOSITION ANALYSIS OF MICROARRAY DATA

Li Liu*, Aventis Pharmaceuticals
Douglas M.Hawkins, School of Statistics, University of Minnesota
Sujoy Ghosh, GlaxoSmithKline
S. Stanley Young, National Institute of Statistical Sciences

In microarray data there are a number of biological samples, each assessed for the level of gene expression for a typically large number of genes. There is a need to examine this data with statistical techniques to help discern possible patterns in the data. Our method applies a combination of mathematical and statistical methods to progressively take the data set apart so that different aspects can be examined for both general patterns and for very specific effects. Unfortunately, these data tables are often corrupted with extreme values (outliers), missing values, and non-normal distributions that preclude standard analysis. We develop a robust analysis method to address these problems. The benefits of this robust analysis will be both the understanding of large-scale shifts in gene effects and the isolation of particular sample-by-gene effects that might either be unusual interactions or the result of experimental flaws. Our method makes use of all data gathered, requires a single pass, and does not resort to complex “cleaning” or imputation of the data table before analysis, yet it is impervious to outlying readings. We illustrate the method with a data set from the literature, where missing values, extreme values and non-normal distribution hamper standard methods.

e-mail: doug@stat.umn.edu

STATISTICAL METHODS FOR TIME COURSE GENE EXPRESSION DATA

Hongzhe Li*, UC Davis School of Medicine

Time-course gene expression data are often measured to study dynamic biological systems and gene regulatory networks. To account for time dependency of the measurements over time and the noisy nature of the microarray data, we propose several mode-based statistical methods for analyzing such data using B-splines, including mixed-effects models for clustering genes, shape-invariant models for identifying periodically expressed genes, and mixture-regression models for identifying genes with different expression profiles over times. Both simulated and real data sets of yeast cell cyle and rat circadian rhythmic data will be used for illustrations.

e-mail: hli@ucdavis.edu

STATISTICAL AND REGULATORY ISSUES IN THE EVALUATION OF GENETIC AND GENOMIC TESTS

Gregory Campbell*, Center for Devices and Radiological Health,
Food and Drug Administration

The genomics revolution is reverberating throughout the worlds of pharmaceutical drugs, genetic testing and statistical science. This revolution, which uses single nucleotide polymorphisms (SNPs) and gene expression technology, including cDNA and oligonucleotide microarrays, for a range of tests from home-brews to high-complexity lab kits, can allow the selection or exclusion of patients for therapy (responders or poor metabolizers). The wide variety of US regulatory mechanisms for these tests is discussed. Clinical studies to evaluate the performance of such tests need to follow statistical principles for sound diagnostic test design. Statistical methodology to evaluate such studies can be wide-ranging, including receiver operating characteristic (ROC) methodology, logistic regression, discriminant analysis, multiple comparison procedures resampling, Bayesian hierarchical modeling, recursive partitioning, as well as exploratory techniques such as data mining. Recent examples of approved genetic tests are discussed.

e-mail: gxc@cdrh.fda.gov

COMBINING SNOW WATER EQUIVALENT DATA FROM MULTIPLE SOURCES TO ESTIMATE SPATIO-TEMPORAL TRENDS AND COMPARE MEASUREMENT SYSTEMS

Mary Kathryn Cowles*, University of Iowa
Dale L.Zimmerman, University of Iowa
Aaron Christ, University of Iowa
David L. McGinnis,

Because snowfall is critical to water supplies in the western U.S., government agencies regularly collect data on snow water equivalent (the amount of water in snow) over this region. Four different measurement systems, of possibly different levels of accuracy and reliability, are in operation: snow courses, snow telemetry, aerial markers, and airbornegamma radiation. We fit a multi-stage Bayesian hierarchical model to data collected over a 90-year period with the goals(1) to estimate thel ong-term temporal trend in SWE over the western U.S. and characterize how this trend varies spatially, and (2) to investigate whether there are systematic differences in the accuracy and reliability of the four measurement systems.

We find substantial evidence of a decreasing temporal trend in SWE in the Pacific Northwest and northern Rockies, but no evidence of a trend in the intermountain region and southern Rockies. Our analysis also indicates that some of the systems differ significantly with respect to their accuracy and reliability.

e-mail: kcowles@stat.uiowa.edu

MODEL VALIDATION AND SPATIAL INTERPOLATION BYCOMBINING OBSERVATIONS WITH OUTPUTS FROM NUMERICAL MODELS

Montserrat Fuentes*, North Carolina State University

In many applications, incorporating all the relevant information can be very difficult for various reasons. For instance, the data could be collected or constructed at different spatial scales, and the bias and measurement error of the available data might depend on the source of information.

We present a Bayesian methodology for spatial prediction with combined data. We model the observed data in terms of an underlying unobservable spatial process Z, and we obtain the posterior predictive values of Z given the available data from the different sources. In this work we take into account the lack of stationarity, change of support, potential bias and measurement error of the data.

We apply these methods to the prediction of air pollution concentrations by combining monitoring data with areal pollutant concentrations from the air quality models ran by EPA.

e-mail: fuentes@stat.ncsu.edu

COMBINING STATE AND NATIONAL SURVEY DATA TO ASSESS WILDLIFE POPULATION TRENDS

Sarah M. Nusser*, Department of Statistics, Iowa State University
William R.Clark, Ecology and Evolutionary Biology Graduate Program, Iowa State Univesrity

Many state agencies conduct annual surveys to monitor animal populations. To assess reasons for shifts over time, it is desirable to combine count data from the state survey with environmental context from a national survey. Combining these data sources can be challenging, and for many state agencies, a reasonably simple solution is needed. We examine pheasant population trends in Iowa in relation to a national program to create wildlife habitat while idling erosion-prone agricultural land (CRP). Land cover/use data from the National Resources Inventory and pheasant count data from an annual pheasant population survey in Iowa are used. Our approach involves identifying a common spatial polygon for linking summaries from the two datasets, then estimating parameters to describe temporal trends in land cover and in pheasant populations over a common time period within each polygon. Estimated pheasant parameters are regressed on land cover summaries to generate interpretable summaries of pheasant population responses in relation to regional differences in the physiography and agricultural use of the land.

e-mail: nusser@iastate.edu

ASSESSING ENVIRONMENTAL IMPACTS USING DATA FROM MULTIPLE SCALES

Linda J. Young*, University of Nebraska
Carol A.Gotway Crawford, National Center for Environmental Health, Centers for Disease Control and Prevention

Data from disparate sources are often used to assess environmental impacts. More often than not, these data have been collected at different scales, and each of the scales may be different from the one of interest. Here an overview of the methods that have been proposed will be given. A geostatistical approach will be investigated. Strengths and weaknesses of each method will be discussed.

e-mail: LJYoung@unl.edu

A BAYESIAN SELECTION PROCEDURE FOR RANKINGS

Tom L. Bratcher*, Baylor University
Cody S.Hamilton, Baylor University

In the setting of multiple treatment levels, the all pairwise comparisons approach is very common. Such procedures actually generate sets of possible rankings of the means.

Here a Bayesian procedure is presented that builds such a set of rankings directly. This Bayes decision rule minimizes a risk that is a linear combination of the expected number of selected rankings and the probability of selecting the true ranking. Application is for the normal model with equal variances.

e-mail: tom_bratcher@baylor.edu

THE BEHRENS-FISHER PROBLEM: A BAYESIAN SOLUTION

Jianrong Wu*, Department of Biostatistics, Jude Children's Research Hospital

The Behrens-Fisher problem concerns the inference for the difference between the means of two normal population without assuming equality of the variances. A Bayesian solution is proposed to obtain the posterior distribution of the difference of means. Applying a second order probability matching prior, the proposed soluation of credible interval has nearly exact frequentist coverage probability even for extremely small samples sizes.

e-mail: jianrong.wu@stjude.org

INFORMATIVE PRIOR SPECIFICATION FOR LINEAR REGRESSION WITH INTERACTION

SSally W. Thurston*, Department of Biostatistics and Computational Biology, University of Rochester
Joseph G.Ibrahim, Department of Biostatistics, University of North Carolina
Susan Korrick, Department of Environmental Health, Harvard School of Public Health, and Channing Laboratory, Brigham and Women's Hospital, Department of Medicine, Harvard Medical School

This work will be motivated by discussion of a dataset for which the intended Bayesian analysis requires an informative prior, due to interactions for which the data likelihood has no direct information. We will present a method of obtaining an informative prior based on either historical data, or on information elicited from a subject matter expert. The only quantities which the expert needs to specify are the population means, variances, and pairwise correlations. Finally, we will discuss how elicited information was used to obtain a proper informative prior for this example.

e-mail: thurston@bst.rochester.edu

BAYESIAN METHODS FOR REGRESSION USING SURROGATE VARIABLES

David H. Manner*, Eli Lilly and Company
John W.Seaman, Baylor University
Dean M. Young, Baylor University

If a dependent variable in a regression analysis is exceptionally expensive or hard to obtain the overall sample size used to fit the model may be limited. To avoid this one may use a cheaper or more easily collected "surrogate" variable to supplement the expensive variable. The regression analysis will be enhanced to the degree the surrogate is associated with the costly dependent variable. We develop a Bayesian approach incorporating surrogate variables in regression based on a two-stage experiment. Illustrative examples are given, along with comparisons to an existing frequentist method.

e-mail: manner_david@lilly.com

IMPROVING MODELS AND METHODS FOR THE STATISTICAL ANALYSIS OF Tg.AC MOUSE BIOASSAYS

Wendell D. Jones*, Analytical Sciences, Inc.
Patrick Crockett, Analytical Sciences, Inc.

Dunson et al [TOXICOLOGICAL SCIENCES 55 (2): 293-302, 2000] proposed a framework for the statistical analysis of skin tumor data (papilloma burden) from Tg.AC mouse bioassays. In the chronic mouse bioassay, a chemical agent is usually applied topically for approximately 26 weeks and the resultant number of skin papillomas are monitored weekly. To account for complications in the statistical analysis of papilloma burden per animal, Dunson et al proposed a Bayes model and method that separate the effects of papilloma latency and multiplicity and that accommodate other aspects of the assay including variability of the expression of the transgene. There are several issues with the preferred implementation method (including parameter estimation using BUGS, a popular MCMC tool) given the nature of the data collection and the resultant data that may yield questionable results. This paper describes some potential improvements to the model and the method so that the bioassay study results are more rigorous and tenable. It also provides a description of an informative prior that greatly assists with more stable model parameter estimates.

e-mail: wjones@asciences.com

BAYESIAN ANALYSIS OF MULTINOMIAL DATA

Chong He*, University of Missouri

A frequently occurring problem in biology, epidemiology, education, and social sciences is the estimation of ultinomial probabilities. In this paper, we first propose using the independent beta prior distributions for the multinomial data. Unlike the previous Bayesian models, the prior distributions are placed on the conditional multinomial probabilities insead of the multinomial probabilities directly. The mathematical structure is greatly simplified by using the reparameterization from the multinomial probabilies to the conditional multinomial probabilities. Therefore the closed froms of any posterior moments of the multinomial probabilities are available. Furthermore, the independent beta prior distributions are a general prior family which includes the Dirichlet prior, the reference prior, and the matching prior as special cases. Hence the closed form of posterior mean and variance of the multinomial probabilities are obtained for the reference prior and the matching prior as well. We then compare the frequesist properties of several different noninformative priors by simulation studies. A novel two-stage stratigy is proposed for estimating the multinomial probabilities.

e-mail: chong@stat.missouri.edu

ON COMPARING POISSON RATES WHEN DATA IS UNDERREPORTED WITH APPLICATIONS TO CERVICAL CANCER

James D. Stamey*, Stephen F. Austin State University
Tom L.Bratcher, Baylor University

A procedure to compare Poisson rates when data is underreported is developed. A fully Bayesian approach is utilized in this procedure. Prior information in the form of double samples and expert opinion are considered. This procedure is applied to cervical cancer data taken in four European countries.

e-mail: jstamey@sfasu.edu

15. Topics in Mathematical Statistics

A NOTE ON FINDING GEODESIC EQUATION OF TWO PARAMETER GAMMA DISTRIBUTION

William W. Chen*, Internal Revenue Service

In this paper, we focus on finding a geodesic equation of the two parameters Gamma Distribution. To find this geodesic equation, we applied both the well-known Darboux Theorem and a pair of differential equations taken from Struik D.J.(1961) (p132, 2-2). Engineers commonly use the gamma distribution to describe the life of a manufactured item. The result of this note could be used as a solution to reference [10], Lauritzen S.L. (1987) pp 163-216.

e-mail: william.w.chen@irs.gov

UNIMODAL DENSITY ESTIMATION USING KERNEL METHODS

Peter Hall, Centre for Mathematics and its Applications, Australian National University
Li-Shan Huang*, Department of Biostatistics, University of Rochester

We suggest a method for rendering a standard kernel density estimator unimodal: tilting the empirical distribution. It is proposed that the amount of tilting bechosen in order to minimise, subject to unimodality, the integrated squared distance between a conventional density estimator and itstilted version. This approach has an interesting data-compression aspect, in that the algorithm often implicitly summarises thedataset in a relatively small subsample as part of the process of enforcing unimodality. Another feature is that, no matter what the chosen bandwidth, the algorithm produces (with probability~1, for each sample size) a proper density estimate. Thus, it may be employed as an adjunct to any of the many popular bandwidth selection rules for density estimation. We show theoretically that in classes of densities that are of practical interest, the method enhances performance without suffering any deleterious first-order impact on asymptotic accuracy, for example as reflected in integrated squared error. In such cases, and except in places where the true density is virtually flat, the constrained density estimate is first-order equivalent to its unconstrained counterpart. The case where the number of modes is constrained to equal a number greater than one is also considered.

e-mail: Lhuang@bst.rochester.edu

ESTIMATION AFTER MODEL SELECTION IN A GAUSSIAN MODEL

Vanja M. Dukic, Department of Health Sciences, University of Chicago
Edsel A.Pena*, Department of Statistics, University of South Carolina

The problem of estimating the variance of a Gaussian model which is intermediate between a model where the mean parameter is fully known and a model where the mean parameter is completely unknown is considered. This problem is motivated by the desire to understand the implications of the process of selecting a model among several competing sub-models, and then estimating a parameter of interest after the model selection, but with these sequential steps using the same sample data. This practice is common in many areas. Of particular issue addressed in this talk is to compare the following three approaches in proceeding with the inference: (I) Utilize estimators developed under a more general model, and in the extreme case, under a fully nonparametric model; (II) Perform a two-step process, with the first step being to select the sub-model, and the second step being to use an estimator developed under the chosen sub-model, with both steps using the same sample data; and (III) Form a weighted combination of sub-model estimators, with the weights, which are the sub-model posterior probabilities, being data-dependent.

e-mail: pena@stat.sc.edu

STATISTICAL ANALYSIS OF SINGLE MOLECULE EXPERIMENTS

Samuel Kou*, Department of Statistics, Harvard University
Jun S.Liu, Department of Statistics, Harvard University

Recent technological advances enable scientists for the first time to follow a biochemical reaction on a single molecule basis. It also raises many statistical challenges. This paper concerns the statistical analysis of one type of single molecule experiments — fluorescence lifetime experiment, in which the conformational dynamics of a single DNA hairpin molecule is studied. The conformational change is modeled as a continuous time Markov chain, which is not directly observable. Instead, it has to be inferred from photons emitted from a dye attached to the DNA hairpin, making the Markovian structure hidden. On top of the hidden Markov structure, the presence of molecular Brownian diffusion adds an extra layer of complexity. As a key step, we obtain a closed form likelihood function. We then show how to use Metropolis-Hastings algorithm and data augmentation method to compute the posterior distributions of the parameters of interest. We also discuss the problem of model selection, in which we show that the energy barrier between the open and close of the DNA hairpin significantly oscillates (modeled as an Ornstein-Uhlenbeck process), shedding light on its biophysical properties.

e-mail: kou@stat.harvard.edu

SEMIPARAMETRIC SKEW-ELLIPTICAL DISTRIBUTIONS

Yanyuan Ma*, North Carolina State University
Marc G.Genton, North Carolina State University

We present a semiparametric representation of generalized skew-elliptical distributions for which the probability density function has the form of a product of an elliptical pdf and a skewing function. By constructing an enumerable dense subset of the skewing function, we are able to define a flexible family of distributions which can capture skewness, heavy tails, and multi-modality systematically. It is straightforward to simulate from this family. We present illustrative examples.

e-mail: yma@unity.ncsu.edu

STATISTICAL ANALYSIS OF AN INTERACTION THRESHOLD ALONG A FIXED-RATIO RAY

Adam K. Hamm*, Virginia Commonwealth University, Department of Biostatistics
Chris Gennings, Virginia Commonwealth University, Department of Biostatistics
Walter H. Carter, Jr., Virginia Commonwealth University, Department of Biostatistics
Ken K. Liao, Colorado State University, Center for Environmental Toxicology and Technology
Raymond S. Yang, Colorado State University, Center for Environmental Toxicology and Technology

Humans are exposed to a variety of chemicals on a daily basis, thus mixture toxicology is an important and relevant issue. Risk assessors often assume additivity at low concentrations of mixture exposures. Recent studies on complex mixtures have revealed interaction thresholds in chemical mixtures through physiologically based pharmacokinetic (PBPK) modeling. We propose the use of empirical models to detect the presence of an interaction threshold along a fixed-ratio ray of the chemicals under study. Our approach is illustrated using a cytotoxicity study involving three metals (arsenic, chromium, lead) and Syrian Hamster Embryonic (SHE) cells. Initial tests for overall departure from additivity along the ray were significant. Therefore, the interaction threshold model is used to detect the presence of a concentration where significant interactions start to occur. An interaction threshold was found to exist at approximately 1.0 to 6.0 mM total concentration. However, the model parameter was not significant. Based on this initial model, Ds-optimal designs are presented that would be associated with a decrease in the variance of the interaction threshold parameter.

e-mail: akhamm@hsc.vcu.edu

---

16. Hazard Modeling and Estimation

USE OF A VALIDATION SUBSET IN DISCRETE PROPORTIONAL HAZARDS MODELS WITH MISMEASURED OUTCOMES

Amalia S. Meier*, Fred Hutchinson Cancer Research Center, HPTN

Outcome mismeasurement has been shown to bias estimation of hazards and of hazard ratios in time-to-event analyses. A general adjusted proportional hazards method (APH) was developed that correctly estimates the discrete-time baseline hazard and hazard ratios when outcomes are subject to a known rate of mismeasurement (Meier, Richardson and Hughes, unpublished work). To achieve proper inference using the APH, accurate estimates of the outcome measure's sensitivity and specificity to the true event status must be available. The current work extends the APH to research contexts in which the mismeasurement rates are unknown or uncertain. Several approaches are compared. Adapting the general methods for surrogate and auxiliary outcomes given by Pepe (1992) and Pepe, Reilly and Fleming (1994), a validation subset is employed. The mismeasurement is described using either a parametric or empirical distribution. Through simulation, results are compared to those of standard discrete-time proportional hazards methods when applied to the validation subset alone.

e-mail: amalia@scharp.org

MIXED BASELINE ADDITIVE HAZARDS MODEL FOR MULTIVARIATE FAILURE TIME DATA

Guosheng Yin*, Department of Biostatistics, University of North Carolina at Chapel Hill
Jianwen Cai, Department of Biostatistics, University of North Carolina at Chapel Hill

Multivariate failure time data often arise in biomedical studies. There might be multiple parallel events of interest, while there could be clustered individuals or organs contributing to each event type. Both the between-failure-type correlation and the within-cluster correlation need to be examined to ensure valid statistical estimation and inference. For such multivariate survival data, we study the additive hazards model and propose estimating equations for parameter estimation. The regression coefficient estimates are shown to follow multivariate normal distribution asymptotically where the sandwich-type variance-covariance matrix can be consistently estimated. Jointly across the failure types, the estimated baseline and subject-specific cumulative hazard processes are shown to converge weakly to a zero-mean Gaussian random field. The weak convergence properties for the corresponding survival processes are established as well. Through a resampling technique, we propose to construct confidence bands for the survival curve. Simulation studies are conducted to assess the finite-sample properties and the new proposal is illustrated with a real data set.

e-mail: gyin@email.unc.edu

THE COX SEMI-MARKOV ILLNESS-DEATH MODEL: A LARGE SAMPLE STUDY

Youyi Shu*, Merck Research Laboratories
John P.Klein, Medical College of Wisconsin

Multistate survival analysis is dominated by Markov chain models. This research work deals with relaxing the Markov assumption commonly made in the classical illness-death model, where patients are in one of the three states: 'healthy', 'diseased' and 'dead'. A Markov model assumes that the intensity of dying while diseased does not depend on the disease duration, whereas a semi-Markov model assumes that it only depends on the disease duration. If we model each transition rate with a Cox (1972) model, the asymptotics for the transition probabilities under the Markov model are readily available in Andersen et al. (1993), but those under the semi-Markov model are unknown. In this research, we derive the asymptotic variances for the survival function, the so-called 'probability of being in response function' (Temkin, 1978) and the 'prevalence function' (Pepe et al., 1991) under the semi-Markov model using martingale theory and a time transformation trick first proposed by Voelkel and Crowley (1984). A Monte Carlo study is conducted to investigate the robustness of Markov/Semi-Markov estimators. A real data example from the PROVA (1991) trial, a Danish multi-center randomized clinical trial concerning bleeding episodes and mortality in patients with liver cirrhosis, is used to illustrate the theory.

e-mail: youyi_shu@merckc.om

PROPORTIONAL HAZARDS ANALYSIS OF RANDOMIZED TRIALS SUBJECT TO COMPETING CAUSES OF CENSORING

Daniel O. Scharfstein, Johns Hopkins Bloomberg School of Public Health
David S.Cohen*, Johns Hopkins Bloomberg School of Public Health
Mark J. van der Laan, University of California, Berkeley
James M. Robins, Harvard School of Public Health

In this paper, we show how to draw inference about the regression parameter in a proportional hazards model when the failure time outcome is subject to competing causes of censoring. We present the result of a simulation study and illustrate our approach using data from the AIDS Clinical Trial Group 193A Study.

e-mail: dcohen@jhsph.edu

DIVERGENCE OF RELATIVE RISK, HAZARD RATIO, AND ODDS RATIO IN PROSPECTIVE EPIDEMIOLOGICAL STUDIES

Michael J. Symons, Biostatistics, School of Public Health, University of North Carolina
Dominic T.Moore*, Biostatistics, Lineberger Cancer Center, University of North Carolina

The analysis of prospective follow-up data usually includes a Cox regression modeling. The hazard ratio, obtained as the exponential of an estimated regression coefficient, consistently exaggerates relative risk and undersizes the odds ratio. Symons and Moore (J. Clinical Epidemiology, 2002) describe how the similarity, or dispartity, of these measures depends upon the product of three factors: (1)the length of follow-up, (2)the average rate of endpoint occurrence, and (3)the risk of the 'experimental' group to the referent group. Basically, these measures are similar when the product of these three factors is 'not too large'. Otherwise they diverge. Described here are accuracy zones, where any pair of the three measures agree to within a specified percentage, like 10%. The results suggest: (a)that for large risks between the groups, the endpoint needs to be even less common and (b) that for small risks between the groups, the endpoint may be quite common. These two point seem not to be emphasized in the mainstream of epidemiological texts and methods.

e-mail: symons@bios.unc.edu

COX MODEL BASED PROSPER FUNCTION ANALYSIS FOR ORGAN ALLOCATION

Jian Yu*, Dept of Biostatistics, University of Michigan
Susan Murray, Dept of Biostatistics, University of Michigan

The lung transplant community is currently considering a reprioritization for waitlisted patients based on a risk/benefit analysis of one year survival rates for patients remaining on the waitlist versus patients receiving an organ. Unless waitlist survival follows an exponential distribution, which is known to be memoryless, patients surviving for a period of time on the waitlist without a transplant will have a different survival profile than those more recently entering the waitlist. This research develops inferential tools for Cox-model based prosper functions that allow individual-specific assessments of transplantation risk/benefit according to waitlist and post-transplant characteristics as known at waitlist time, t. Furthermore, we introduce a corresponding test statistic function across waitlist times, along with relevant confidence bands, that adjusts for correlation between waitlist and post-transplant prosper functions using empirical martingale process results. A simulation study estimating coverage probabilities validates the constructed confidence bands. The method will be illustrated using the lung patient data from United Network of Organ Sharing (UNOS).

e-mail: yujian@umich.edu

ESTIMATION OF TRANSITION RATES IN A MULTISTAGE PROPORTIONAL HAZARDS MODEL FOR LONGITUDINAL DATA

Rafiqul I. Chowdhury*, Department of Health Information Administration, Kuwait University,
Ataharul M.Islam, School of Mathematical Sciences, Science University
Makhdoom A. Shah, Associate Professor & Chairman, Department of Health Information Administration, Kuwait University,

There has been growing interest in the use of multistge model to the longitudinal data sets, in recent years. An algorithm using the S-plus is proposed to estimate the parameters of a multistage model. This algorithm is applied to a longitudinal data set on pregnancy-related morbidity among rural Bangladeshi women. Complications occuring at three phases of childbearing, antenatal period, delivery, and postnatal period, were considered as different stages. This data was analysed by employing a multistage proportional hazards model for transitions and reverse transitions. Seven sets of estimates were obtained for seven different types of transitions for the complications that occured at the three stages. Model significance was measured using the global chi-square test proposed previously. Five covariates, economic status, wanted pregnancy, number of previous pregnancies, education of the respondents, and visit for antenatal care were used in the model. It was found that a women is more likely to suffer complications during delivery, if she suffered one or more major complications during antenatal period. Irrespective of the number of any complications during delivery, if a women suffered from one or more major complications during antenatal period, there was a substantially higher risk of suffering from complications during postpartum period.

e-mail: rafiq@hsc.kuniv.edu.kw

THE EFFECTS OF MODEL MISSPECIFICATION AND UNRECOGNIZABLY INCOMPLETE DATA IN FLOWGRAPH MODELS

C. Lillian Yau*, School of Public Health and Tropical Medicine, Tulane University
Aparna V.Huzurbazar, Department of Mathematics and Statistics, University of New Mexico

Flowgraphs provide an alternative methodology for analysing semi-Markov processes, especially when interest focuses on estimating the PDF, survival, or hazard functions of a waiting time. Given a stochastic process with conditionally independent states, a flowgraph model allows different waiting time distributions for the different states. It provides a method for accessing the waiting time distribution for any partial or total waiting time. Flowgraph models operate on MGFs and use saddlepoint approximations to convert the MGFs to waiting time PDFs, cumulative distribution functions, survivor functions, and hazard functions. Flowgraphs handle censoring and can be used in either a frequentist or a Bayesian framework. Yau and Huzurbazar (2002) use flowgraph for heavily censored and incomplete data to model diabetic retinopathy. We extend this work to handle model misspecification, which occurs via parametric assumptions imposed on random waiting times within the flowgraph model, and unrecognizably incomplete data, which are distinct from censored data in that these are data that are missing but not known to be missing. We illustrate our methods using large scale simulation studies.

e-mail: cyau@tulane.edu

MULTIPLE IMPUTATION IN SURVIVAL ANALYSIS OF INTERVAL- AND RIGHT-CENSORED DATA

Jaroslaw Harezlak, Department of Biostatistics, Harvard School of Public Health
Wanzhu Tu*, Division of Biostatistics, Indiana University School of Medicine
and Regenstrief Institute

In most survival analyses, time of the event is observed exactly or is right-censored. In certain situations, the event of interest is known to have happened within an interval only. We propose a non-parametric method based on multiple imputation of possible event times to improve efficiency of interval-censored methods. The methodology will be illustrated on data coming from a study of STI's in a group of young women. Available infection data comes from scheduled visits. Additionally, daily diary data on the sexual activities of the study subjects is collected. With the auxiliary information provided by the diaries, we use the times of sexual encounters to improve the estimation of the time to an infection. A random sample of possible infection times is drawn from the interval of interest. Each sample is analyzed using a method for right-censored data, and the estimates are combined in an appropriate manner. Our method is compared with other techniques used for interval-censored data. In our simulation studies, we explore one and two sample analysis in non-parametric context. Our results indicate a good performance of the proposed method in all considered settings.

e-mail: jharezla@hsph.harvard.edu

ANALYSIS OF MEDICAL COSTS WITH CENSORED DATA

Yong Zhou*, University of North Carolina at Chapel Hill
Jianwen Cai, University of North Carolina at Chapel Hill

The statistical analysis and inference of lifetime medical cost is often challenged by the fact that the survival times are censored on some study subjects and their subsequent costs are unknown. In this paper, we propose empirical likelihood procedures for making statistical inference about the mean and median of medical cost when censoring is present. Confidence intervals for those parameters are easily derived under a given confident level. We show that the proposed procedure gives better confidential intervals for the mean or median of medical cost than those in the existing literature. We also propose empirical likelihood estimators for the mean and median of medical cost. When the estimating equation based on the empirical likelihood does not have solution, we propose to use a smoothed empirical likelihood. We show that the proposed estimators are consistent and asymptoticallynormally distributed. Simulation studies are conducted to examine the finite sample properties in sample of practical size. The methods are applied to two real data sets for medical costs of patients with colorectal cancer and lung cancer.

e-mail: yzhou@bios.unc.edu

PENALIZED SPLINE OF RESPONSE PROPENSITY METHODS IN MISSING DATA ANALYSIS

Hyonggin An*, Department of Biostatistics, University of Michigan
Roderick J. A.Little, Department of Biostatistics, University of Michigan

Recently, several nonparametric or semiparametric approaches of dealing with missing data problems have been developed in order to reduce bias and improve efficiency under the assumption of missing at random. However, these methods can be suffered from the ‘curse of dimensionality’ if the dimensions of covariates are large. In this talk, we propose a semiparametric method using penalized spline and response propensity to retain the flexibility of nonparametrics, but to resolve the ‘curse of dimensionality’ problem. Simulations are conducted to evaluate its performance.

e-mail: hyongg@umich.edu

DRAWING INFERENCE FROM A REGION OF ESTIMATES: SENSITIVITY ANALYSIS FOR MISSING DATA

Stijn Vansteelandt*, Ghent University, Belgium
Els Goetghebeur, Ghent University, Belgium
Michael Kenward, London School of Hygiene and Tropical Medicine, U.K.
Geert Molenberghs, Limburgs Universitair Centrum, Belgium

Analyses of incomplete data are problematic as parameters describing the target population are typically not identified without untestable assumptions. To make sense of the incomplete data while avoiding misguided conclusions due to incorrect assumptions, sensitivity analyses have replaced the classical point estimate by a region of parameter estimates. Molenberghs, Kenward and Goetghebeur (2001) call such region an estimated ignorance region because it expresses ignorance due to the missing data.

The idea of worst case-best case regions is not new, but some major open challenges remain. First, special algorithms are needed to allow for computationally feasible estimation of ignorance regions. Second, the region itself is subject to sampling variation and this needs to be acknowledged. Third, incorporating expert opinions is not straightforward. In this talk, we address these practically important questions and illustrate them with some solutions in a number of practical applications. Our development will lead to a formalism for sensitivity analysis which naturally allows to combine model uncertainty or 'ignorance', and sampling uncertainty or 'imprecision' into overall Estimated Uncertainty RegiOns (EUROs).

e-mail: Stijn.Vansteelandt@rug.ac.be

UNBIASED ESTIMATE AND EFFICIENT TESTS IN CLINICAL TRIALS WITH DROPOUTS

Peter C. O'Brien, Mayo Clinic
David Zhang*, Genentech
Kent R. Bailey, Mayo Clinic

Last observation carried forward (LOCF) and analysis using data from subjects who complete a trial (Completers) are commonly used techniques for analyzing data in clinical trials with dropouts when the endpoint is change from baseline at last scheduled visit. We propose two alternative methods. A parametric method, referred to as the KM procedure, is conceptually similar to the life-table method and corresponding Kaplan-Meier estimation. Here we express the change from baseline at K years as the summation of annual changes. A nonparametric analogue of LOCF, referred to as LRCF, is obtained by carrying forward the rank of the change from baseline at the last observation. Both procedures retain the simplicity of LOCF and Completers analyses and free from modeling assumptions. In simulations intended to reflect chronic degenerative diseases, LOCF was observed to produce markedly biased estimates and inflated type I error rates while these problems did not arise with the Completers, KM, or LRCF methods. KM and LRCF were more powerful than Completers, and the KM procedure provided more efficient estimates with smaller variance than the Completers approach, in all simulations.

e-mail: dzhang@gene.com

AVOIDING BIAS WHEN ROUNDING IN MULTIPLE IMPUTATION

Nicholas J. Horton*, Department of Biostatistics, Boston University School of Public Health Stuart R.Lipsitz, Department of Biometry and Epidemiology, Medical University of South Carolina
Michael Parzen, Graduate School of Business, University of Chicago

With the advent of general purpose packages that support multiple imputation for analyzing datasets with missing data (e.g. Solas, SAS PROC MI and S-Plus 6.0), we expect much greater use of multiple imputation in the future. For simplicity, some imputation packages assume the joint distribution of the variables in the multiple imputation model is multivariate normal, and impute the missing data from the conditional normal distribution for the missing data given the observed data. If the possibly missing data are not multivariate normal (say, binary), imputing a normal random variable can yield implausible values. To circumvent this problem, a number of methods have been developed, including the 'predictive mean matching method', in which the imputed normal variate is rounded to the closest observed value in the dataset. We show that this rounding can cause biased estimates of parameters, whereas if the imputed value is not rounded, no bias would occur. We show that rounding should not be used indiscriminately, and thus some caution should be exercised when using predictive mean matching, particularly for dichotomous variables.

e-mail: horton@bu.edu

---

18. Statistical Methods In The Sutdy of HIV and AIDS

COMPARISON OF LINEAR, NONLINEAR AND SEMIPARAMETRIC MODELS FOR ESTIMATING HIV DYNAMIC PARAMETERS

Hulin Wu, Frontier Science & Technology Research Foundation
Caixia Zhao, Department of Statistics,Texas A&M University,
Hua Liang*, Department of Biostatistics, St. Jude Children's Research Hospital

The potency of an antiretroviral agent can be assessed on the basis of early virological response such as viral decay rates in AIDS clinical trials. Currently linear, nonlinear, semiparametric models have been proposed to estimate the viral decay rates in viral dynamic models. Whether do these models produce consistent estimates of viral decay rates? if not, which model is correct and should be used in practice. We applied these models to data from an AIDS clinical trial with potent antiviral treatments and found significant inconsistencies in the estimated rates of reduction in viral load. Simulation studies indicated that reliable estimates of viral decay rate were obtained by using the nonlinear and semiparametric models. Our analysis also indicated that the decay rates estimated by using linear models should be interpreted differently from those estimated by using nonlinear models. The semiparametric model is preferred to other models since no arbitrary data screening is needed. Based on the real data analysis and simulation studies, guidelines for estimating viral decay rates from clinical data are provided for practitioners.

e-mail: hua.liang@stjude.org

PER-ACT RATE OF HIV TRANSMISSION FROM INFECTED THAI MEN TO THEIR MONOGAMOUS FEMALE PARTNERS

Charles M. Heilig*, Centers for Disease Control and Prevention
Steve Shiboski, Univsersity of California, San Francisco

Using doubly censored current status data, we estimate the rate of HIV transmission per sexual contact from Thai men to their monogamous female partners. Our hazard regression technique allows for time-varying infectivity and time-fixed covariates and accommodates uncertainty in male seroconversion dates. The method also checks heterogeneity.

624 HIV-infected men were identified from blood bank records in Chiang Mai, Thailand. 210 of their 472 monogamous female partners were HIV-infected at enrollment. The male partners' seroconversion dates are bounded by their last known HIV-negative date (or the start of the Thai HIV epidemic, if no date is known) and their first positive HIV test. These date intervals range up to 9 years in length, admitting stable estimates up to 5 years after male seroconversion.

The highest infectivity follows male seroconversion, with 57 transmissions (95% CI 38–86) per 1000 contacts in the first month. It then drops nearly to 0, and it rises to 14 transmissions (95% CI 5–41) per 1000 contacts at 5 years after male seroconversion. Infectivity is positively associated with a history of sexually transmitted infection in either the male or the female.

e-mail: cheilig@cdc.gov

ESTIMATION OF PARAMETERS AND ASSESSMENT OF TREATMENT EFFECTS ON HIV PATHOGENESIS UNDER HAART BY STATE SPACE MODELS

Wai-Yuan Tan, The University of Memphis
Ping Zhang*, The University of Memphis
Xiao-Ping Xiong, St.Jude Children's Research Hospital
Pat Flynn, St.Jude Children's Research Hospital

To treat HIV-infected individuals by anti-viral drugs, recently important breakthrough has been achieved through the combination of the drugs-reverse transcriptase inhibitors(RT inhibitors) and protease inhibitors (PR inhibitors), referred to as the Highly Active Anti-Retroviral Therapy(HAART). To assess effects of drugs and to monitor the disease status under HARRT, one would need to estimate the effects of RT inhibitors and PR inhibitors, as well as the numbers of non-infectious HIV and infectious HIV. To date, methods for this purpose are non-existent. To fill in this gap, in this paper we have developed a method to estimate these parameters and the state variables to assess effects of HAART on HIV pathogenesis. As an illustration, we have applied the method of this paper to the data of a patient from the St. Jude Children Research Hospital treated by anti-retroviral drugs and HAART.

e-mail: pzhang1@memphis.edu

A BAYESIAN APPROACH FOR ESTIMATING ANTIVIRAL EFFICACY IN HIV DYNAMIC MODELS

Yangxin Huang*, Frontier Science & Technology Research Foundation
Hulin Wu, Frontier Science & Technology Research Foundation

The study of HIV dynamics is one of the most important developments in recent AIDS research. It has led to a new understanding of the pathogenesis of HIV infection. Although important findings in HIV dynamics have been published in prestigious scientific journals, the statistical methods (nonlinear least squares, for example) for parameter estimation and model-fitting used in those papers appear surprisingly crude and have not been studied in more details. In this paper, a viral dynamic model is developed to evaluate the effect of pharmacokinetic variation, drug resistance and adherence on antiviral response. In the context of this model describing HIV infection, we investigate a Bayesian modeling approach under a nonlinear mixed-effects (NLME) model framework. In particular, our modeling strategy allows us to estimate time-varying antiviral efficacy of a regimen during the whole course of treatment period by incorporating the information of drug exposure and drug sensitivity. Both simulation and real clinical data examples are given to illustrate the proposed approach. The Bayesian approach involves assumptions of probability distributions for model parameters prior to an analysis being performed, allowing the fitting of complex models and enabling analysis of all of the model parameters, and has great potential to be used in many aspects of viral dynamics modeling. It is suggested that Bayesian approach for estimating parameters in HIV dynamic models is more flexible and powerful than the nonlinear least squares method.

e-mail: huang@fstrf.dfci.harvard.edu

ANALYZING MULTIPLY MATCHED COHORT STUDIES WITH TWO DIFFERENT COMPARISON GROUPS: APPLICATION TO PREGNANCY RATES AMONG HIV+ WOMEN

Yan Li*, Department of Epidemiology and Public Health,Yale University
Daniel Zelterman, Department of Epidemiology and Public Health,Yale University
Brian W.C. Forsyth, Department of Pediatrics,Yale University

We develop a new statistical method to analyze multiply matched cohort studies with two different comparison groups. We employ a linear logistic model to describe the underlying log odds ratios and use a conditional likelihood approach to conduct inference. Under the assumption of homogenous log odds ratios, we provide methods to construct both asymptotic and exact confidence regions of the two log odds ratios in a simple case. We propose a score test to evaluate the assumption of homogeneous log odds ratios across strata. While our methods are general, we develop them around a specific application, namely, the study of pregnancy rates in HIV infected women. Our analyses suggest that HIV infection is associated with a decrease in pregnancy rates and that this decrease in fertility becomes significant after accounting for illicit drug use.

e-mail: yl97@omega.med.yale.edu

A NEW TWO-DIMENSIONAL DESIGN FOR PHASE I CLINICAL TRIALS

Kai Wang*, Department of Biostatistics, University of North Carolina—Chapel Hill
Anastasia Ivanova, Department of Biostatistics, University of North Carolina—Chapel Hill

Often treatment plans in Phase I oncology trials consist of two or more agents. The purpose of such trials is to find a set of maximum tolerated dose combinations—combinations that yield the pre-specified toxicity rate (usually 20%). The most common approach used in two-agent trials is to fix the dose of one agent and vary the dose of the other agent, that is, to reduce the design to one dimension. In this talk we present a two-dimensional Bayesian design for Phase I trials with two agents. The new design uses the natural assumption that the probability of toxicity is a non-decreasing function of dose level in one agent when the dose of the other agent is fixed. The new design assigns fewer patients to sub-optimal or toxic doses and provides a higher quality estimation of maximum tolerated dose combinations than do one-dimensional designs.

e-mail: kwang@bios.unc.edu

DISCUSSION OF ISSUES IN DESIGNING PROTOCOLS FOR STEM CELL TRANSPLANTATION AND PRACTICAL VARIATIONS OF THE CLASSICAL PHASE II STUDY.

Anne I. Goldman*, Biostatistics, University of Minnesota

Categorizing the designs of clinical trials into phase I, II, III, and IV is useful for trials of a new drug agent, such as in the treatment of cancer. When the intervention is complex or the disease is rare, a standard design may not be appropriate. Such is the case with many clinical trials in the field of hematopoetic stem cell transplantation (HSCT). This paper presents practical approaches to designing such trials, from the Minnesota program.

Currently, 69 protocols are open of which fewer than 30% are national multi-clinic trials. Of the 49 'local' protocols, 4 (8%) are phase I and 27 (55%) are II, while none are phase III or standard controlled/comparative trials and the remaining 37% are 'Standard of Care' protocols. Notably, a quarter are specifically for patients with rare non-malignant diseases, many of the rest are open to both cancer and non-cancer patients. That standard designs are frequently not suitable for HSCT trials is briefly discussed, but the focus is on describing variations of phase II protocols with emphasis on defining the relevant efficacy and toxicity variables for monitoring. Algorithms for finding optimal stopping boundaries are described in Goldman and Hannan, Stat. Med. 20: 1575-1589 (2001).

e-mail: anne@biostat.umn.edu

SIMULATIONS TO IMPROVE EXPERIMENTAL DESIGNS FOR U-SHAPED DOSE-RESPONSE MODELING

Daniel L. Hunt*, St Jude Children's Research Hospital

This research includes the results of simulations in which correlated binary data was generated. This data mimics binary litter data collected from developmental toxicity studies on animals. An assumption in these simulation designs is that hormesis exists, as evidenced by the dose-response pattern of the generated data. This also implies the existence of a threshold level below which the hormetic dose-response pattern occurs. Although hormesis implies several possible dose-response patterns, in this work, the hormetic pattern below threshold is assumed to be U-shaped. These simulations improve upon the design of past developmental experiments, which had standard allocations of equal litters per group and arbitrary dose spacing. The improvements include designs in which dose levels and litters are reallocated to increase the power for detecting hormesis.

e-mail: daniel.hunt@stjude.org

OPTIMAL STUDY DESIGN FOR SETTINGS WHERE THE EXPOSURE OF INTEREST IS SUBJECT TO SHORT AND LONG TERM VARIATION

Sohee Park*, Department of Biostatistcs, Harvard School of Public Health
Louise Ryan, Department of Biostatistics, Harvard School of Public Health

We discuss environmental epidemiology studies where the predictor of primary interest is long-term, average exposure, but where cost and practical considerations limit the number of times at which study subjects can be assessed for exposure. We discuss the measurement error that arises in this setting when the exposure is subject to short term (e.g. day-to-day) as well as long term (e.g. month-to-month) variation. We develop optimal designs for reliability sub-studies in this context, extending existing methods to incorporate both short and long-term variation. Our research is motivated by a study designed to assess the effects of pesticides on male reproductive health.

e-mail: shpark@hsph.harvard.edu

EVALUATING OBSERVER AGREEMENT FOR AGREEMENT STUDY DESIGN WITH REPLICATED READINGS

Jingli Song*, Department of Biostatistics, Rollins School of Public Health, Emory University
Huiman X.Barnhart, Department of Biostatistics, Rollins School of Public Health, Emory University
Michael Haber, Department of Biostatistics, Rollins School of Public Health, Emory University

In clinical studies, assessing agreement of multiple observers plays an important role in the evaluation of continuous measurement scales. The traditional agreement study is often designed either with each rater only taking one reading on each subject or with one rater taking replicated readings. The limitation of such a design is the inability to simultaneously assess both inter-rater and intra-rater agreement. Therefore, we advocate an agreement study design that collects replicated readings of each rater on each subject. In this talk, we present inter-rater, intra-rater and the overall agreement indices that can be used to fully assess the agreement of multiple observers for data collected in such a study design. We propose a new index based on the 'true' reading of each rater on each subject to evaluate the inter-rater agreement and use the ICC to assess the intra-rater agreement. We also extend the OCCC (Barnhart it et al, 2003) to evaluate the overall agreement for the case of replicated readings and investigate the relationship of the OCCC with the inter-OCCC and ICC. We propose estimation and inference approaches based on the GEE method and the method of moment.

e-mail: jsong@sph.emory.edu

A CURE MODEL APPROACH TO BAYESIAN PHASE I TRIAL DESIGNS

Thomas M. Braun*, University of Michigan, Department of Biostatistics

We use a cure model to generalize the likelihoods used in the CRM [Q'Quigley, et al., Biometrics (46) p.33-48] and TITE-CRM [Cheung & Chappell, Biometrics (56) p.1177-1182] designs of Phase I trials. Specifically, like the CRM and TITE-CRM, we model the probability of toxicity as a one-parameter monotonic function of dose. However, for those subjects who experience toxicity, we use another one-parameter monotonic function of dose to describe their times to toxicity, thereby creating a two-parameter likelihood from the mixture of subjects who experience toxicity and those subjects who do not experience toxicity. We show that each subject's contribution to thedetermination of the maximum tolerated dose (MTD) is weighted by a function of the time to toxicity parameter. In numerical examples, we compare our cure model-based approach to the CRM and TITE-CRM in terms of identifying theMTD and how doses are allocated.

e-mail: tombraun@umich.edu

ESTIMATION OF INFLUENZA ANTIVIRAL AGENT EFFICACY BASED ON HOUSEHOLD DATA

Yang Yang*, Dept. of Biostatistics, Emory University
Ira M.Longini, Dept. of Biostatistics, Emory University

The typical randomized trial design for influenza antiviral agents is as follows: Households are enrolled and followed prospectively. When the first case of influenza appears in the household, then either the drug or placebo is randomly assigned to that index case and the other exposed members. The antiviral efficacy is then measured from the resulting data. Currently, it is unclear if randomization should be on a household or individual level, and whether we need to use information on all the initially enrolled households or just those with one or more cases. Furthermore, there is a misclassification problem since influenza case data are observed, but the inference is done on the infection process. Maximum likelihood and log-linear models are applied to estimate both antiviral efficacy for susceptibility and antiviral efficacy for infectiousness, while taking the different designs and misclassification into account. For the log-linear modeling approach, we develop an EM algorithm competitive with the maximum likelihood approach but easier to implement. We find that only using households with one or more cases is almost as efficient as using all households in the study. In addition, individual level randomization is more efficient than household level randomization regardless of how households are ascertained.

e-mail: yyang3@sph.emory.edu

20. Placebo-Controlled Trials: Ethics, Science and Economics

PLACEBO OR ACTIVE CONTROLS? SCIENTIFIC AND ETHICAL CONSIDERATIONS

Susan S. Ellenberg*, Center for Biologics Evaluation and Research, FDA

Placebo-controlled trials (or more generally, trials in which active therapy is withheld from the control group) regularly generate controversy. Concerns about the ethics of such trials have sometimes been raised in settings of serious diseases where available therapies are non-existent or unsatisfactory, and a new experimental treatment raises hopes of improved outcomes. More recently, it has been argued that placebo-controlled trials are always unethical if one or more treatments have been proven effective for the disease under study and could therefore be used as an active control. It has been further argued that when known effective therapy is available, the comparison of a new therapy to placebo is of little scientific interest or value. This presentation will address the ethical concerns that have been raised, and will explore the scientific, regulatory and public health implications of each side of this debate.

e-mail: ellenberg@cber.fda.gov

THE ECONOMICS OF PLACEBO IN DRUG DEVELOPMENT

David J. DeBrota*, Lilly Research Laboratories,Eli Lilly and Company
Teresa S.Williams, Lilly Research Laboratories, Eli Lilly and Company

The most common comparator employed today in controlled clinical trials testing the safety and efficacy of novel pharmaceutical therapeutics is probably placebo. Consequently it may be supposed that more money is collectively spent in the study of placebo than on any single molecule. Pharmaceutical companies recognize clinical research to be extremely costly, and constantly seek to find ways to make drug development decisions more economically. Placebo is the comparator of choice not only for scientific reasons, but also for economic and pragmatic ones. If placebo could not be used as a comparator, clinical trials would have to be larger and might actually put more patients at risk of being treated with less than optimally effective therapies. Importantly, placebo should not be imagined to be equivalent to no therapy. Rather, placebo must be seen as potentially powerful medicine, and vital to the efficient advancement of medical research. In this presentation a number of aspects of placebo and of the economics of drug development will be examined, and in particular the role that placebo plays in clinical research decision making will be discussed.

e-mail: djd@lilly.com

INFLUENCE OF CONTROL GROUP DESIGN ON MAGNITUDE OF TREATMENT EFFECT IN DOUBLE-BLIND RANDOMIZED CLINICAL TRIALS: THE CASE OF ANTIPSYCHOTIC EFFICACY FOR SCHIZOPHRENIA

Scott W. Woods, Department of Psychiatry, Yale School of Medicine
Ralitza V.Gueorguieva, Yale University
Robert W. Makuch*, School of Epidemiology and Public Health, Yale University

The active-controlled equivalence study and the low dose-controlled study are two prospectively randomized clinical trial designs that offer alternatives to the placebo-controlled study for investigating new treatments for schizophrenia. Interpretation of results from these designs could be complicated if the same dose of medication appeared to perform differently in such studies from studies that do contain a placebo group. Method. Using a comprehensive review process, we identified placebo-, low dose-, and active-controlled randomized clinical trials of risperidone, olanzapine, or quetiapine. We used meta-analysis to evaluate the effect of design on outcome. Results. In general, patients treated with antipsychotics improved more in low dose-controlled studies than in placebo-controlled studies when dose was controlled in the analysis, and even more in active-controlled studies. At therapeutic doses, the ratio of improvement in active-controlled to placebo-controlled studies was 1.6 (95% CI 1.2, 1.9). Conclusions. One possibility that could account for the effect of design has been termed ‘active control bias.’ Such active control bias is likely to affect an attempt to evaluate assay sensitivity by inflating the comparison of a newer treatment in an active-controlled equivalence study to historical placebo. Supported by USPHS grant MH57292.

e-mail: scott.woods@yale.edu

STATISTICAL MODELS FOR ANTHRAX OUTBREAKS

Ron Brookmeyer*, Johns Hopkins University, Department of Biostatistics
Natalie Blades, Johns Hopkins University

This talk will consider the role of statistical models in understanding anthrax outbreaks. In the fall of 2001, an act of bioterrorism resulted in an anthrax outbreak in the United States. Public health officials responded by distributing antibiotics to attempt to reduce morbidity and mortality. We will consider the role of statistical models for understanding what happended in the fall of 2001. In particular, we estimate the numbers of cases of inhalational anthrax that were prevented by the antibiotics. We will also discuss the 1979 Sverdlovsk outbreak in the former Soviet to estimate the incubation period of inhalational anthrax. The statistical chalenge arises becasuse of truncation effects in the data. Simulation results will be presented to evaluate the models. Our overall findings emphasize the importance of early detection of outbreaks for minimizing mortality. The models may also be helpful in identifying the unknown source of an outbreak by estimating the date cases were esposed to anthrax spores.

e-mail: rbrook@jhsph.edu

BIOSURVEILLANCE FOR EARLY DETECTION OF TERRORISM

Ken Kleinman*, Harvard Medical School

This talk describes experience in setting up a system for collecting daily information on ambulatory patients who present symptoms compatible with bioterrorist attack (e.g. high fever, rashes, severe nausea, etc.). Those data are analyzed, on a daily basis, to discover unusual disease clusters in space as they appear in real time.

While several statistical techniques for surveillance exist, they are mainly limited to purely temporal or purely spatial data collection. The few contemporary techniques for spatio-temporal surveillance were not yet adapted to the context of daily data when the system was developed.

We thus implemented a relatively simple system that is currently in effect. We will summarize that system here. The system uses uncorrelated random effects to estimate the p-value for the observed cases in small geographic areas under the null hypothesis of the random effects model being correct.

In an effort to stimulate interest in methodological development in this area, a similar data set was disseminated. A separate session will include three other approaches to this kind of data, using this data set as an example.

e-mail: ken_kleinman@harvardpilgrim.org

RISK ANALYSIS AND GAME THEORY

David Banks*, U.S. Food and Drug Administration

Traditional game theory is a poor guide to human decision-making. This talk explores ways in which the standard method can be made more realistic through the use of statistical risk analysis, different payoff tables for different players, non-minimax rules, and conditional strategies. To illustrate the ideas, we focus upon decision-making in the context of bioterrorism, and evaluate various defense strategies that arise in the context of a smallpox threat.

e-mail: banksd@cber.fda.gov

22. Innovative Approaches and Challenges in the Analysis of Recurrent Failure Time Data

THE ACCELERATED GAP TIMES MODEL

Robert L. Strawderman*, Cornell University
Edsel A.Pena, University of South Carolina

In this talk we introduce the accelerated gap time model, a new semiparametric intensity model for recurrent event data. This model represents a useful alternative to the modulated renewal process model of Cox (1972). It shares strong similarities with the accelerated failure time model in that the role of the covariates is to shrink or expand the time to subsequent recurrence. Such models are appealing because the covariates influence the relevant time scale in a direct and interpretable manner. It will be demonstrated that estimation can be carried out in a manner that parallels the single event setting; however, the required asymptotics are different because the basic processes involved no longer form martingale processes in the relevant time dimension.

e-mail: rls54@cornell.edu

THE GAMMA-FRAILTY POISSON MODEL FOR THE NONPARAMETRIC ESTIMATION OF PANEL COUNT DATA

Ying Zhang*, University of Central Florida
Mortaza Jamshidian, California State University, Fullerton

We study nonparametric estimation of the mean function of a counting process with panel count data. We introduce the gamma frailty variable to account for the intra-correlation between the panel counts of the counting process. This estimating procedure, while preserving the simplicity of the computation, improves the estimate efficiency over the nonparametric maximum pseudo-likelihood estimate studied in Wellner and Zhang (2000). Two simulated examples are used to illustrated the method.

e-mail: zhang@mail.ucf.edu

ROBUST INFERENCE FOR NON-MARKOV EVENT DATA

David V. Glidden*, Division of Biostatistics, University of California, San Francisco

Multistate event data, in which a single subject is at risk for multiple events, is common in biomedical applications. This talk considers nonparametric estimation of the vector of probabilities of state membership at time t. Estimators, derived under the Markov assumption, have been shown to be consistent for data which is non-Markov. Inference, however, must take into account possibly non-Markov transitions when constructing confidence bands for event curves. We develop robust confidence bands for these curves, evaluate them via simulation and illustrate the method on two datasets.

e-mail: dave@biostat.ucsf.edu

EVALUATION OF TREATMENTS WITH RECURRENT EVENT DATA

Jerald F. Lawless*, University of Waterloo
Richard J.Cook, University of Waterloo

The evaluation of treatment effects in longitudinal studies where subjects may experience recurrent or multiple events has reveived considerable discussion. Marginal attributes such as expected numbers of events provide a simple basis for comparisons but may miss interesting features of the event history process. In addition, marginal and conditional analysis differ in their robustness properties and in the need to adjust for selection effects or process-related censoring. This talk will examine these issues and suggest guidelines for treatment comparisons.

e-mail: jlawless@uwaterloo.ca

23. Bayesian Hierarchical Spatial Temporal Methods

NONSTATIONARY SPATIAL MODELLING FOR MULTIPLE POINT SOURCES

Sujit K. Ghosh*, NC State University
Jacqueline M.Hughes-Oliver, NC State University

The objective of this talk is to develop, assess, and provide convenient tools for implementing parametric modeling of nonstationary processes that are driven by point sources. The Clean Air Act (CAA) of 1970 and its 1977 and 1990 amendments defined different categories of pollution sources. “A source is defined as any place or object from which pollutants are released. ..” This talk proposes a hierarchical Bayesian approach to an extension of a process decomposition model that was recently introduced in the context of modeling the effect of point sources. Many benefits afforded by the process decomposition model and the Bayesian paradigm will be discussed. No restrictions are placed on the index set of the process, so that temporal, spatial, and spatial-temporal processes are all considered. The resulting methodology will be applied to address a public health and welfare concern of the EPA, namely to determine which NOx and SO2 emission sites aremost responsible for site-specific ambient concentrations far from the emission sites.

e-mail: sghosh@stat.ncsu.edu

MODELLING REPLICATED SPATIAL DATA USING SPATIALLY-VARYING GROWTH CURVES

Sudipto Banerjee*, Division of Biostatistics, University of Minnesota.
Gregg Johnson, Department of Agronomy,University of Minnesota
Nick Schneider, Thorp Seed Co.

We look at a spatially replicated experiment, where each location is, by itself, a lattice of plots with each plot monitoring the growth of a weed through time. Along with estimating fixed effects that affect the growth of the weed, of particular interest is the spatial variation in the growth patterns. This setup allows the modelling of spatial variation at two resolutions - for the plots nested within locations and between the locations.We develop a Bayesian hierarchical framework that helps capture variations across replicates and locations, using growth curve coefficients that vary spatially. We use spatial processes to model the spatially-varying coefficients and discuss the flexibilty of our approach. We look at several competing models that arise in our context, discuss fitting algorithms and perform model comparisons. We interpret these models with respect to our experiment and indicate extensions and future work.

e-mail: sudiptob@biostat.umn.edu

CONDITIONALLY SPECIFIED SPACE-TIME MODELS FOR MULTIPLE POLLUTANTS

Michael J. Daniels*, Department of Statistics, University of Florida
Zhigang Zhou, Department of Statistics, Iowa State University
Hui Zou, Department of Statistics, Stanford University

We propose a class of conditionally specified models for the analysis of several pollutants over space and time. Such models are useful in situations where there is sparse spatial coverage of one pollutant and much more dense coverage on other pollutants. The dependence structure over pollutants, space, and time is completely specified through a neighborhood structure. We introduce several computational tricks which are integral for model fitting and implement a gibbs sampler to sample from the posterior distribution of the parameters. Model fit is assessed via the DIC and predictive ability, both over time and space, by mean squared prediction error. The models are used to analyze particular matter and ozone data collected in the LA area in 1995 over a 3 month period.

e-mail: mdaniels@stat.ufl.edu

HIGH THROUGHPUT DRUG TARGET IDENTIFICATION

S. Stanley Young*, NISS
Douglas M.Hawkins, U Minnesota
Li Liu, Aventis

With the advent of large, micro array data sets, it is now possible to identify drug targets through association analysis, "guilt by association" or "you are known by the company you keep." There is a need to execute association analysis in the presence of outliers, missing data and heavy-tailed distributions. We will examine a relatively large micro array data set where the associations are known with two statistical methods that should be relatively impervious to these data problems. As we know the answers, we should be able to judge the utility of the statistical methods.

e-mail: genetree@bellsouth.net

CHEMICAL INTERPRETATION OF CLASSIFIERS OF HIGH THROUGHPUT SCREENING DATA

Yan Yuan*, Department of Statistics and Acturial Science, University of Waterloo
William J.Welch, Department of Statistics and Acturial Science, University of Waterloo
Hugh A. Chipman, Department of Statistics and Acturial Science, University of Waterloo

Today millions of compounds are available as potential drug candidates. High throughput screening (HTS) is widely used to identify compound activity in drug discovery. A classification tree may be built using HTS data to predict the activity of unscreened compounds, and hence select compounds to be screened. The response variable of the tree model is biological activity, and the explanatory variables are descriptor sets based on compound structure. In addition to building a model for prediction, it is also useful to understand the structure and activity relationship (SAR). Similarly, it is of interest to identify groups of chemicals amongst the active compounds. Therefore, it is necessary to know which descriptor variables are important for activity. Our study shows that a descriptor variable may not necessarily be important for activity although it is used to build part of a classification tree. We describe methods for identifying the most important descriptor variables that relate to part of a tree, leading to simplification of the tree and understanding of the structure and activity relationship.

e-mail: y4yuan@math.uwaterloo.ca

DESIGN AND ANALYSIS OF POOLING EXPERIMENTS BASED ON STRUCTURE ACTIVITY RELATIONSHIP

Lei Zhu*, GlaxoSmithKline
Hughes-Oliver M.Jacqueline, North Carolina State University
S. Stanley Young, National Institute of Statistical Sciences

Pooling experiments are used in pharmaceutical companies as a cost-effective approach for screening compounds during drug discovery. Compounds are tested in pools of size 10 or so, thus making pools more information dense than individual compounds. When a biologically potent pool is found, the goal is to decode the pool, that is, to determine which of the individual compounds are potent. The specification of the compounds occurring in each of the pools is referred to as a pooling design. We propose augmenting the data on pooled testing with information on the chemical structure of compounds in order to complete the decoding and designing processes. This proposal is based on the well-known relationship between biological potency of a compound and its chemical structure. We use real data from a drug discovery process at GlaxoSmithKline to demonstrate the improvement in hit rate; namely, the number of potent compounds identified divided by the number of tests required.

e-mail: lz34757@gsk.com

25. Analysis of Microarrays and Gene Expression I

METHODS FOR ANALYZING ARRAY DNA COPY NUMBER DATA

Adam B. Olshen*, Department of Epdemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center
E. S.Venkatraman, Department of Epdemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center

Cancer progression often involves alterations in DNA sequence copy number. Array versions of comparative genomic hybridization (CGH) and representational difference analysis (RDA) allow for high-resolution assessment of copy number gains and losses across entire genomes. Similar to gene expression studies, the copy number for a mapped sequence on a genome is related to the ratio of test to reference fluorescent intensities. Our goal is to make accurate genome-wide maps of copy number of subjects undergoing array copy number studies from the noisy intensity ratios. We have developed a novel modification of binary segmentation and a hidden Markov model for this purpose. We discuss the similarities and differences between these approaches and their application to real array copy number data sets.

e-mail: olshena@mskcc.org

COMBINING THE MODEL SELECTION AND FALSE DETECTION RATES (FDRs) TO DETECT DIFFERENTIALLY EXPRESSED GENES IN MICROARRAY EXPERIMENTS

Lang Li, Indiana University
Jingjin Li*, Indiana University
Xiaohua Zhou, University of Washington

In microarray data analysis, simulateneous inference for detecting differentially expressed genes is a complex statistical problem. A methodology based on estimating false detection rates (FDRs) (Efron et al., 2001) has been introduced to allow for the realistic assessment of true expression changes adjusting for the multiplicity of genes being screened. On the other hand, model-based approaches have been proposed to control the variations in the microarray experiments and to achieve less bias and more accurate estimates.

In this paper, we combine the model selection and FDRs to better detect the differentially expressed genes under two experimental conditions and two RNA preparation time points. Analyses using various bootstrap methods with and without model selection are compared based on the data of 5032 genes. And simulations are performed to evaluate the approach we present.

e-mail: jingjinl@iupui.edu

GENE SELECTION FOR MICROARRAY CLASSIFICATION USING PRINCIPAL COMPONENT

Antai Wang*, Biostatistics Unit, Lombardi Cancer Center, Georgetown University
Aiyi Liu, Biometry and Mathematical Statistics Branch, NIH/NICHD
Edmund A. Gehan, Biostatistics Unit, Lombardi Cancer Center, Georgetown University

Principal component analysis (PCA) has been widely used in data structure detection, dimensionality reduction and classification. It has become a useful tool in microarray data analysis. For a typical microarray data set, since the number of genes tends to be much larger than the number of observations, it is often difficult to compare the overall gene expression difference between observations from different groups or do the classification based on so many genes. In this talk, we compare several variable selection methods using PCA to select a subset of genes for microarray classification. The effectiveness of these methods in microarray classification is evaluated using gene expression data with known membership of the observations.

e-mail: aw94@georgetown.edu

NYLON MEMBRANE MICROARRAY SPOT DECONVOLUTION AND OPTIMIZATION OF ARRAY GRID DESIGN

Lisa H. Ying*, Merck & Co., Inc
Vladimir Svetnik, Merck & Co., Inc

In the nylon array experiments, samples are radiolabeled using 33P isotopes and hybridized to the matching cDNA printed on the array. A radio-imaging cassette is exposed to the array, and the radiation emitted by the bound sample is captured by the detection media of the cassette. Subsequently an image processing system produces an image that displays the spots. The nylon array spots differ in size to such an extent that some big spots "bleed over" to their neighbors, thus significantly distorting the later's spot intensities. The "bleed over" is due to the facts that radiolabel emits radiation isotropically and the detection media collects signals from the array non-selectively. In this presentation we address the followings: 1)How to identify the spots, which have been affected by "bleed over"? 2)How to estimate the intensity for the affected spots? 3)How to design nylon arrays such that "bleed over" spots and their affected neighbors can be easily identified? 4)If we compensate for the "bleed over", how much accuracy & precision will be gained? The results were obtained using the software developed by T. Therneau, who also customized them for our microarrays.

e-mail: lisa_ying@merck.com

STATISTICAL DESIGN OF REVERSE DYE MICROARRAYS

Kevin K. Dobbin*, Biometric Research Branch, National Cancer Institute
Joanna Shih, Biometric Research Branch, National Cancer Institute
Richard Simon, Biometric Research Branch, National Cancer Institute

In cDNA microarray experiments all samples are labelled with either Cy3 dye or Cy5 dye. Certain genes exhibit dye bias — a tendency to bind more efficiently to one of the dyes. The common reference design avoids the problem of dye bias by running all arrays "forward," so that the samples being compared are always labelled with the same dye. But comparison of samples labelled with different dyes is sometimes of interest. In these situations, it is necessary to run some arrays "reverse" — with the dye labelling reversed — in order to correct for the dye bias. We consider three types of experiments for which some reverse labelling is needed: paired samples, samples from two predefined groups, and reference design data when comparison with the reference is of interest. We present simple probability models for the data, derive optimal estimators for relative gene expression, and compare the efficiency of the estimators for a range of designs. In each case, we present the optimal design and sample size formulas. We show that reverse labelling of individual arrays is generally not required.

e-mail: dobbinke@mail.nih.gov

DIFFERENTIAL EXPRESSION ANALYSIS VIA POOLING ERRORS FOR GENES WITH SIMILAR EXPRESSION INTENSITIES

Michael A. O'Connell*, Insightful COrporation

Many gene expression studies are implemented with a limited number of replicated arrays, because of expense and limited availability of RNA materials (Chen et al., 1997; Lee, 2002). This makes analysis of differential expression through within-gene models problematic. For example, permutation based tests dont work well with limited replication.

To alleviate this problem, we provide methods and S-PLUS functions for expression error estimation based on pooling errors within genes and between replicate arrays for genes in which expression values are similar. This approach leverages the fact that variability of gene expression intensity between replicates often decreases as a nonlinear function of the average gene expression intensity.

e-mail: moconnell@insightful.com

META TESTS IN MICROARRAY DATA ANALYSIS

Grier P. Page*, Department of Biostatistics, University of Alabama at Birmingham
Jode W.Edwards, Department of Biostatistics, University of Alabama at Birmingham
David B. Allison, Department of Biostatistics, University of Alabama at Birmingham

When describing a microarray experiment researchers often look at the data and say that there appears to be a given pathway up regulated or down regulated or genes of a certain class of genes appears to be up or down regulated. These observations are often quite qualitative based upon experience or 'gut' feeling. In order to make such analyses more quantitative we have developed meta analytical methods to quantify whether a class of genes are differentially regulated in a single experiment. These methods can also be extended to comparing the same gene's significance between multiple studies. We have considered several meta-tests: 1) Combining of individual p-values into a single p-value for all genes in a single class 2) Vote counting methods where the number 'significant' genes in a class are compared to how many were expected at random 3) Combining of estimates of effect sizes across tests by parametric and non-parametric methods.

e-mail: gpage@ms.soph.uab.edu

A TEST FOR HETEROGENEITY OF EFFECT IN CONDITIONAL LOGISTIC REGRESSION

Randall H. Rieger*, West Chester University
Clarice R.Weinberg, NIEHS

The conditional logistic regression (CLR) model assumes a constant exposure coefficient across clusters. Application of CLR when there is instead heterogeneity among clusters leads to biased variance estimation and biased hypothesis testing (and suggested a robust alternative method of analysis). Here, we propose the Test of Heterogeneous Susceptibility (THS), a simple non-parametric test to assess the assumption of homogeneity in response to exposure in the CLR model. Simulations demonstrate that the test has reasonable power to reveal violations of homogeneity. We apply the THS to a study of periodontal disease.

e-mail: rrieger@wcupa.edu

A NOVEL VARIABLE SELECTION AND SIGNIFICANCE TESTING TECHNIQUE FOR QUANTILE REGRESSION

David T. Redden*, University of Alabama at Birmingham
Jose R.Fernandez, University of Alabama at Birmingham
David B. Allison, University of Alabama at Birmingham

Quantile regression seeks to model the quantiles of a random variable as a function that is conditional upon observed variables. Research into significance testing and variable selection techniques for quantile regression is still warranted. We propose a novel and simple to implement iterative algorithm utilizing logistic regression to test significance and select variables for inclusion into a quantile regression function. Simulations have been performed to estimate the Type I error rate and statistical power of the procedure. Results from this novel procedure are compared to 1) methods proposed by Koenker and Machado (1999) and 2) bootstrap techniques commonly used for quantile regression inference. Results show that this new less computationally intense method has appropriate Type I error rates and has power comparable to both the methods of Koenker and Machado (1999) and bootstrap techniques. We illustrate the procedure using 779 cross-sectionally selected subjects from the Third National Health and Nutrition Examination Survey (NHANES III) to determine which variables best predict the quantiles of the waist circumference distribution among male adolescents ages 12 to 18.

e-mail: dredden@ms.soph.uab.edu

VALIDITY OF USING LOGISTIC REGRESSON ANALYSIS IN THE ASSESSMENT OF CORRELATION BETWEEN HOSPITAL VOLUME AND SURGICAL MORTALITY

Yen-Hong Kuo*, Jersey Shore Medical Center, Meridian Health System
Yen-Liang Kuo, Tri-Service General Hospital, Taipei, Taiwan, Republic of China

Mortality rate is one of the major indicators to evaluate the surgical outcomes of a hospital. In order to assess the correlation between hospital volume and surgical mortality, logistic regression analysis is widely utilized in the medical literature. Within the regression model, surgical outcome of each patient is used as the dependent variable and the hospital volume (either the value or the level of volume after categorizing) is used as the independent variable. However, the association among patients within the same hospital was not considered in most of the published studies. Therefore, the validity of this type of approach is of our concern. A simulation study was conducted to evaluate the effectiveness of using logistic regression analysis. The probability of Type I error from a simple logistic regression model was used as the outcome measurement. Combinations of mortality rates and distributions of hospital volume were assessed. This presentation will discuss the validity issues based on the simulation outcomes. Examples from the medical literature will be presented to illustrate the impacts.

e-mail: yhkuo@jhu.edu

ON A FAMILY OF PARAMETRIC MODELS FOR ANALYZING CLUSTERED BINARY DATA

E. Olusegun George, Department of Mathematical Sciences, University of Memphis,
Deo Kumar Srivastava*, Department of Biostatistics, St. Jude Children's Research Hospital
Roopa Seshadri, Department of Pediatrics, Northwestern University

The joint distribution of exchangeable binary random variables can be constructed using the sieve method. This procedure gives a representation of the joint distribution in terms of a finite number of completely monotone moment sequences. An important property of these moments is that they are Laplace transforms of a distribution function. We use this property to construct parametric models for exchangeable clustered data. We show that such models can be constructed so that they can be interpreted as generalized linear models for clustered data. We evaluate our models by analyzing several developmental toxicity data sets and compare our results with those obtained through the use of other procedures.

e-mail: kumar.srivastava@stjude.org

A MIXTURE MODEL FOR THE ANALYSIS OF CORRELATED BINOMIAL DATA

N. Rao Chaganty*, Old Dominion University

Statistical analysis of correlated binomial data is complicated by the fact either there are many parameters in the model involving higher order correlations or the ranges of the correlation parameters, being functions of the marginal means, are subject to unmanageable constraints. In this talk I will discuss a mixture model, which circumvents those complications, allows a wide range of dependence and requires only specification of the first two moments for the mixing distribution. I will also discuss estimation of the parameters and present an illustrative example.

e-mail: rchagant@odu.edu

ANALYSIS OF CLUSTERED BINARY DATA IN THE PRESENCE OF VARYING CLUSTER SIZES

Abigail G. Matthews*, Harvard School of Public Health
Rebecca A.Betensky, Harvard School of Public Health
Dianne M. Finkelstein, Harvard School of Public Health

A common approach to analyzing clustered binary data is to use the quadratic exponential model (Zhao and Prentice, \textit{Biometrika} 77:642-648 1990), which assumes that all clusters are of the same size. This would be violated, for example, if some data were missing. Researchers have previously developed an approach to handling this problem that calculates the distribution of the observed data by summing the complete data likelihood over all possible values of the missing data. In some applications, for example studies of familial aggregation, where families form the clusters, this approach is not ideal. We propose an alternative approach which considers all possible subsets of a fixed size of a cluster, and uses generalized estimating equations to adjust for a cluster contributing more than once to the likelihood. Our approach assumes that all subsets of clusters of a particular size behave the same as clusters of exactly that size. These methods will be compared to other standard procedures through simulation, and applied to a study of familial aggregation of cancers in families with and without BRCA1/2 mutations.

e-mail: amatthew@hsph.harvard.edu

DISTRIBUTIONAL PROPERTIES OF RATIO OF TWO RANDOM VARIABLES WITH APPLICATION TO ACTIVE CONTROL NON-INFERIORITY TRIALS

Yong-Cheng Wang*, Food and Drug Administration
Gang Chen, Food and Drug Administration
George Chi, Food and Drug Administration

What should be an appropriate level of "fraction retention" in the design of an active control non-inferiority trial is a critical and most important issue. It is often based on clinical judgment alone. The clinical judgment for such selection is often based on the disease, the control effect size, and the objective of the trial. If the objective is to demonstrate efficacy of the new treatment, then 50% retention may be reasonable if the new treatment offers other clinical benefits that are not available from the control. If the objective is to demonstrate that the new treatment is equivalent or non-inferior to the control, then the fraction retention needs to be higher. In this paper, we will show that an appropriate level of fraction retention is not only based on the clinical judgment but also the statistical properties of fraction retention. We develop the fundamental statistical theory for the distributional properties of fraction retention considered as a ratio of two independent normal random variables and provide statistical methods in the determination of a level of fraction retention. With selected level of retention, appropriate design of the trial, the statistical inference, type I error assessment and adjustment, will be presented.

e-mail: WangY@CDER.FDA.gov

TWO APPROACHES TO THE BEHRENS-FISHER PROBLEM WITH EXTENSIONS TO REPLICATE DATA

Terry Hyslop*, Biostatistics Section, Division of Clinical Pharmacology, Thomas Jefferson University

We consider the Behrens-Fisher problem of testing means from two independent populations where the variances are not assumed to be equal. Our motivations for considering this problem are two-fold. First, an application of interest is to consider bioequivalence problems in parallel design. In recent approaches to population bioequivalence, the assumption of equality of variances is removed, which required us to consider Behrens-Fisher methods. Secondly, we considered a generalization for replication as repeated measures data is frequently seen in many situations, including our motivating application. We propose an approach based on linear combinations of random variables as in Howe (1974). We also consider a test introduced by Lee and Fineberg (1991) that we generalize to replicate data. Simulation studies are used to demonstrate the level and power of these tests. A numerical example is also provided to illustrate these methods.

e-mail: Terry.Hyslop@mail.tju.edu

EXACT INFERENCE FOR BIOEQUIVALENCE BETWEEN TWO DRUGS IN INCOMPLETE 2X2 CROSSOVER EXPERIMENT

Sang-Gue Park*, Chung-Ang University &
Harvard School of Public Health
Nam-Kyu Lim, Chung-Ang University

The exact inference procedures for bioequivalence between two drugs are considered when the 2x2 crossover experiment is incomplete. The proposed inference procedure is based on conditional argument. Patel (1985)'s approximate method is discussed and compared through a simulation study.

e-mail: spark@cau.ac.kr; sgpark@hsph.harvard.edu

A NEW ASYMPTOTIC TEST FOR TESTING NON-INFERIORITY FOR PAIRED BINARY DATA

David Li*, Merck Research Labs.
Cong Chen, Merck Research Labs.

Several asymptotic statistics have been proposed for testing non-inferiority for paired binary data. It has been well accepted that score statistic performs better than others in terms of the type-I error rate and power. In this talk, we are going to present a new statistic where a new variance estimate is used. Through simulation, we show that the test using this new statistic has type-I error rate closer to the nominal level and is more powerful, compared to the score statistic in reasonable settings.

e-mail: jianjun_li@merck.com

SAMPLE SIZE RE-CALCULATION IN NON-INFERIORITY TRIALS THAT INCLUDE A PLACEBO ARM

Todd A. Schwartz*, School of Public Health, University of North Carolina at Chapel Hill
Jonathan S.Denne, Department of Global Statistics, Eli Lilly and Company

Clinical trials with an objective of non-inferiority differ from superiority trials in that they are designed to demonstrate that the difference in efficacy between a test treatment and an active control is less than a pre-determined and acceptable quantity. This predetermined quantity is usually calculated as a proportion of the anticipated superiority of the active control over placebo. Here, we operate in the context of a three-arm non-inferiority trial that includes assignment to test treatment, active control, and additionally placebo in order to estimate the magnitude of the superiority of active control over placebo in the trial. For such a design, the determination of an adequate sample size for a desired level of power can depend on several unknown parameters, including the treatment difference between active control and placebo. When there is excessive uncertainty surrounding the magnitude of these parameters, an approach to estimating these parameters and updating the sample size during the course of the trial may be desirable. Such an approach is explored here, including comparisons with fixed-sample, two-arm non-inferiority, and group-sequential designs.

e-mail: tschwart@bios.unc.edu

ASSESSING EQUIVALENCE OR NONINFERIORITY IN SCREENING TESTS WITH PAIRED BINOMIAL ENDPOINTS

Boris G. Zaslavsky*, FDA/CBER/DBE

The repeated measures logistic regression (PROC GENMOD, SAS 8.x) is an efficient tool for assessing consistency of different clinical protocols tested on the same sample with no gold standard among them. The confidence interval (CI) on the odds ratio is a useful measure of consistency. It can distinguish equivalent clinical protocols. This method is applicable to samples that are composed of both paired individuals and independent individuals. Moreover, it can compare many available clinical protocols simultaneously. The null hypothesis can be rejected either when the P-value <0.05 or when "1" doesn't lie within the 95% confidence interval, both of which are equivalent. We implemented a numerical analysis of the width of the CI for the odds ratio of two paired protocols with respect to the sample size. The first size of the sample was 398 individuals (796 outcomes) of which 8 individuals had (16) discordant outcomes. The second sample size was only 8 individuals, all of them discordant (16 outcomes). Even so, the P-value changed minimally, but the width of the CI changed 3.8 times.

e-mail: zaslavsky@cber.fda.com

NONPARAMETRIC REGRESSION FOR WEAKEST LINK MODELS

Roger S. Day*, Department of Biostatistics, University of Pittsburgh
Thomas J.Richards, Department of Biostatistics, University of Pittsburgh

Weakest link models connect predictors to responses via the assumption that a predictor can change the expected response only if all other relevant predictors are sufficiently high (or sufficiently low). Thus almost all points in covariate space have a single 'weakest link' predictor which locally determines the expected response. Such models may be more relevant to real phenomena than linear models, especially in biology and medicine.

Parametric models achieve this by positing an optimal use function for each predictor, then taking the maximum (or minimum) at each point in covariate space. These models are computationally difficult to fit and unsuitable for screening large numbers of parameter pairs, triples, etc. We propose a nonparametric approach in which the optimal use functions are determined by a quantile-matching method. This approach provides extremely rapid parameter screening. The method is demonstrated on immunologic data from a large clinical trial of interferon for melanoma.

e-mail: day@upci.pitt.edu

NONPARAMETRIC RESAMPLING-BASED TESTS FOR ASSOCIATION: WITH APPLICATION TO HIGH-DIMENSIONAL UNORDERED CATEGORICAL COVARIATE SETS

Greg DiRienzo*, Department of Biostatistics; Harvard School of Public Health
Victor DeGruttola, Department of Biostatistics; Harvard School of Public Health
Marcello Pagano, Department of Biostatistics; Harvard School of Public Health

We propose nonparametric methods to assess the significance of the association between a set of predictor variables and a response variable. Predictor variables considered are unordered categorical variables, e.g. amino acid sequence. A resampling-based method is first used to shrink the number of covariates to the set most predictive of response. Conditional on these positions, all possible patterns of amino acid sequences are formed; each one referred to as a cell. We condition on observed cell sizes and disregard empty cells; denote the number of non-empty cells by L. In general, any cell size greater than 0 is equally accommodated by our methods. Some summary statistic of the response variable is calculated for each cell, e.g. the mean or median. Conditional on the responses, resampling techniques are used to estimate the marginal distribution of the summary statistic under the null hypothesis of no association between genotype sequence and response. A confidence band for the Q-Q plot with respect to this null distribution is estimated via resampling; any of the L statistics lying outside are significant
departures from the null at the desired global significance level.

e-mail: dirienzo@hsph.harvard.edu

NEW DISCRIMINANT PROCEDURE APPLIED TO STRUCTURE ACTIVITY/PROPERTY

Kimberly S. Crimin*, Pharmacia Corporation

In drug discovery, there has been increasing interest in structure-based models for predicting activities of molecules. When a non-linear relationship exists between the descriptors and the property or when the property results from many biological mechanisms, classification methods are used to group the compounds into classes. In discriminant analysis, functions used for separation can be used for classification. Traditional discriminant procedures offer views of the data based on the space spanned by differences in group means. This is essentially the visualization procedure SIR. Similar to traditional SAVE, the spanning space is expanded to include differences in variances and covariances. The vectors in the spanning set are then ordered using a procedure based on the QR decomposition of the right unitary matrix from the singular value decomposition of the space. We call our new procedure SMVCIR. Traditional estimates of location, variance and covariance are replaced with robust estimates and results from a small simulation study are presented. Several examples are presented and the results compared to traditional discriminant procedures.

e-mail: kimberly.crimin@pharmacia.com

VARIANCE INFLATION IN STRONGLY NONNORMAL BIVARIATE REGRESSION MODELS. SIMULATION STUDY.

Simon Rosenfeld*, National Institutes of Health / National Cancer Institute
Victor Kipnis, National Institutes of Health / National Cancer Institute

Estimation of variance of regression slope is often based on the assumption of multivariate normality. However, in reality this assumption is often violated and produces misleading conclusion about accuracy of estimates. Simulation shows that inflation of confidence intervals due to deviations from normality can reach factor 5. Using Monte Carlo simulation, we compare several strategies for estimating variance of regression slope, i.e., OLS, Box-Cox transformation, sandwich estimator, bootstrap and weighted regression. It is shown that in samples of moderate size, all these strategies produce biased results. Simulation has been performed for several families of bivariate distributions, such as gamma, lognormal and 'box-coxable' distributions. A paradoxical finding is that for 2D 'box-coxable' distribution (ie, exactly transformable to normality by Box-Cox transformation) variance inflation due to model selection is pronounced much stronger than for gamma distribution. Another important result is that sandwich estimator is found completely equivalent to bootstrap for all the families of tested bivariate distributions both in terms of point estimates and coverage probabilities.

e-mail: sr212a@nih.gov

COMPARISON OF UNCONDITIONAL EXACT TESTS FOR TESTING THE DIFFERENCE OF INDEPENDENT BINOMIAL PROPORTIONS.

Jimmy A. Doi*, NC State University, Dept. of Statistics
Roger L.Berger, National Science Foundation, Math Sciences Division

Exact tests based upon the difference of two independent binomial proportions are popularly used and are especially suited for studies with small to moderate sample sizes. In the context of testing for superiority, we apply the confidence region p-value method (Berger and Boos, 1994) in proposing an exact test which has been found to perform better than the standard exact test in many situations. In both cases, the tests use unconditional distributions and the variances of the test statistics are determined via a restricted maximum likelihood method (Farrington and Manning, 1990). Inverting the tests, we derive corresponding confidence intervals and we provide coverage probability and expected length comparisons.

e-mail: jadoi@unity.ncsu.edu

ENVELOPE EMPIRICAL LIKELIHOOD RATIO OF TWO SAMPLE HYBRID MODEL WITH CENSORING

Kyoungmi Kim*, Department of Statistics, University of Kentucky

Clinical trials often compare several groups of subjects for the presence or absence of treatments, where each subject may contribute the values to the analysis, the values in different groups being highly related by constraints. While accounting for the relation of two groups provided by some constraints, we are interesting in testing whether the efficacy of treatment is the same among different groups. In this paper, we propose an alternative approach, based on a simple adjustment of empirical likelihood ratio for testing the efficacy of treatment. The suggested approach utilizes information on subjects and the provided information plays a main role to the analysis. We show that the empirical likelihood ratio with many constraints provided by the relation of two treatment groups has the asymptotic Chi-square distribution with two degrees of freedom. In addition to the issues of practicality, general explicit expressions are presented for the large sample information of the envelope empirical likelihood estimators with many constraints.

e-mail: kkim@ms.uky.edu

BAYESIAN ISOTONIC REGRESSION AND TREND ANALYSIS FOR CONTINUOUS REGRESSION FUNCTIONS

Brian Neelon*, University of North Carolina
David B.Dunson, National Institute of Environmental Health Sciences

In many applications, the mean of a response variable is known to be a monotone function of a predictor, and interest focuses on estimating the response function, while also assessing evidence of an association. We discuss a new framework for Bayesian isotonic regression based on a constrained piecewise linear model, with a fixed number of knots with unknown locations. The nondecreasing constraint is incorporated through a prior distribution, which is parameterized according to the prior probability of no association and the prior expectation for the slopes. The prior density follows a restricted autoregressive Gaussian process structure, which results in conditional conjugacy properties and smoothing of the response function. Extensions to multiple covariates are accommodated through a generalized additive structure. The methods are illustrated through application to data from an epidemiology study.

e-mail: bneelon@bios.unc.edu

29. Statistical Inference for Correlated Data

DISTRIBUTION DIAGNOSTICS FOR LINEAR MODELS WITH CORRELATED OUTCOMES: APPLICATIONS TO ENVIRONMENTAL TIME SERIES

E. Andres Houseman*, Harvard School of Public Health
Louise M.Ryan, Harvard School of Public Health
Brent A. Coull, Harvard School of Public Health

Correlated data arise often in environmental monitoring settings, which typically involve measuring processes repeatedly over time and/or space. While regression models for correlated outcomes (e.g. linear mixed models and time series models) are widely used in such settings, methodology for addressing goodness-of-fit remains relatively underdeveloped. In this paper we present an easy-to-implement approach that lends itself to graphical displays. We propose a residual vector that is asymptotically N(0,I). Projections of this vector are used to construct an empirical cumulative distribution function (ECDF), whose stochastic limit is characterized. We describe a resampling technique for generating representatives from the stochastic limit of the ECDF, and use these samples to construct a global test for the hypothesis of normal errors. We also demonstrate that the ECDF of the predicted random effects, as described by Lange and Ryan (1989), can be formulated as a special case of our approach. We demonstrate our methods on two environmental data sets pollen count data described in Stark et al. (1997) and unpublished monitoring data from the Massachusetts Water Resources Authority (MWRA). Our examples emphasize the context of the data analysis, where diagnostics may be used either to justify a simpler parametric model or to motivate a more complex semiparametric approach.

e-mail: ahousema@hsph.harvard.edu

A LATENT PROCESS MODEL FOR JOINT ANALYSIS OF TIME-TO-EVENT AND LONGITUDINAL DATA WITH APPLICATION TO THE PROSCAR LONG-TERM EFFICACY AND SAFETY STUDY

Weili He*, Merck & Co., Inc.
Weichung Joe Shih, UMDNJ School of Public Health
Yong Lin, UMDNJ School of Public Health
Hui Quan, Merck & Co., Inc.
George R. Rhoads, UMDNJ School of Public Health

Many clinical and epidemiologic studies often collect longitudinal repeated measurements of a time-dependent disease marker and time-to-event of a disease. We proposed a joint modeling approach to analyze these two types of data simultaneously using Markov chain Monte Carlo techniques. Additionally, we proposed an extension of the joint modeling approach when the longitudinal repeated measurement might be a potential surrogate endpoint for the time-to-event clinical endpoint. Extensive simulation studies and a data application to the ProscarTM Long-Term Efficacy and Safety Study demonstrated that joint modeling longitudinal and time-to-event data provided a superior alternative to modeling these two types of data separately.

e-mail: weili_he@merck.com

ANALYSIS OF TEN-YEAR PSA 'TRAJECTORIES'' AFTER PROSTATE CANCER SURGERY

Elizabeth L. Turner*, Department of Mathematics and Statistics, McGill University
James A.Hanley, Department of Epidemiology and Biostatistics, McGill University
Nandini Dendukuri, Department of Epidemiology and Biostatistics, McGill University
Peter C. Albertsen, Department of Surgery, University of Connecticut Health Center

Prostate Specific Antigen (PSA) is a biochemical marker used to monitor prostate cancer following treatment. We have analyzed serial PSA data for a cohort of men who underwent radical surgery for prostate cancer in the early 1990s. We describe the development of a statistical model that reflects the characteristics of these complex data. It accommodates, via a mixture model with constrained parameters, two sub-populations of men who are or are not cured, a further hierarchical model for the biologic pattern in the latter, along with left-censoring for undetectable PSA levels. We describe how we fit this model using Gibbs Sampling and discuss issues that require further study.

e-mail: eturner@math.mcgill.ca

INDEXES OF TRACKING IN LONGITUDINAL DATA WITH INDIVIDUALLY VARYING TIMES OF OBSERVATION.

Steven E. Hanna*, Department of Clinical Epidemiology and Biostatistics, McMaster University

In epidemiology and medicine, the longitudinal stability of a measured characteristic is often referred to as 'tracking' (Twisk et al, 1994). This is typically operationalized either as the maintenance of relative rank among subjects over time, or the prediction of future measurements from earlier ones. In either case, the usual goal of analysis is to summarize the degree of expected variation from time-to-time in some characteristic of clinical interest. Various indexes of tracking have been proposed, but their calculation and interpretation typically depend on subjects having been measured at common times of observation. We use mixed effects models to estimate tracking coefficients from data with individually varying times of observation, with particular focus on an adaptation of the index proposed by Foulkes and Davis (1981). We demonstrate utility of these mixed effects indexes in a cohort study of motor development among children with physical disabilities (Rosenbaum et al., 2002).

e-mail: hannas@mcmaster.ca

OPTIMAL CRITERION FOR DEFINING A POSITIVE TEST BASED ON MARKERS

Xuelin Huang*, U. of Texas, M. D. Anderson Cancer Center

The sensitivity and specificity of tests based on multiple continuous longitudinal variables are estimated, based on rough criteria currently used in practice. The problem of optimal criterion to define a positive test is considered.

e-mail: xlhuang@mdanderson.org

A MULTI-STATE STOCHASTIC MODEL FOR LONGITUDINAL DATA WITH DISCRETE OUTCOMES

Sujuan Gao*, Indiana University School of Medicine

We consider longitudinal studies where multiple discrete outcomes are collected and transitions between these outcomes are of interest. Examples of such data are longitudinal dementia studies where study participants are usually evaluated for disease status at fixed follow up intervals and can be classified into multiple outcomes, such as demented, cognitive impairment without dementia, normal or deceased. We propose a multi-state stochastic model for the analysis of such datasets. The model includes multiple states for clinical outcomes with some reversible states and an absorbing state for death. We assume a time homogeneous Markov chain for the multi-state model and covariates-dependent hazard functions. We introduce subject specific random effects into the model and assume that the transitional probabilities depend on the stochastic states the subjects occupy at one prior step conditional on random effects. Maximum likelihood approach will be used for parameter estimation from this model framework. Data from a community-based longitudinal dementia study will be used to illustrate our approach.

e-mail: sgao@iupui.edu

TESTING FOR SPATIAL CORRELATION IN BINARY DATA WITH APPLICATION TO ABERRANT CRYPT FOCI IN COLON CARCINOGENESIS

Tatiyana V. Apanasovich*, Graduate Student

In an experiment to understand colon carcinogenesis, all animals were exposed to a carcinogen while half the animals were also exposed to radiation. Spatially, we measured the existence of what are referred to as aberrant crypt foci (ACF), morphologically changed colonic crypts that are known to be precursors of colon cancer development. The biological question of interest is whetherthe locations of these ACF's are spatially correlated: if so, this indicates that damage to the colon due to carcinogens and radiation is localized. Statistically, the data take the form of binary outcomes (corresponding to the existence of an ACF) on a regular grid. We develop score—type test methods based upon the Matern and conditionally autoregression (CAR) correlation models to test for the spatial correlation while allowing for nonstationarity. Because of a technical peculiarity of the score—type test, we also develop robust versions of the method. The methods are compared to a generalization of Moran's test for continuous outcomes,and are shown via simulation to have the potential for increased power. When applied to our data, the methods indicate the existence of spatial correlation.

e-mail: tanya@stat.tamu.edu

IDENTIFICATION OF PLACEBO RESPONDERS AMONG DRUG TREATED SUBJECTS

Thaddeus Tarpey*, Wright State University
Eva Petkova, Columbia University
Todd Ogden, Columbia University
Ying Chen, Columbia University

Positive responses among subjects treated with an active drug may be due to the chemical component of the drug, a placebo effect, spontaneous remission or a combination of these factors. Outcome profiles over time for depressed subjects treated with antidepressants and placebos are estimated and classified. Clustering techniques are used to determine representative profiles for various outcome profiles (placebo responder, true-drug responder, etc). The profiles for placebo treated subjects are used to help validate the classification process. In addition, the methodology will be evaluated using data from discontinuation trials where all subjects received the active drug in the first phase and in the second phase, subjects either continue on the drug or they are switched to a placebo.

e-mail: thaddeus.tarpey@wright.edu

GROWTH CURVE MODELING OF PLACEBO AND DRUG RESPONSES

Michael R. Elliott*, University of Pennsylvania School of Medicine
Thomas R.Ten Have, University of Pennsylvania School of Medicine

Tarpey et al. (2002) posit outcome profiles that distinguish true drug responders from placebo responders, and use semi-parametric clustered functional profile modeling (Flury 1990, 1993) to estimate the percentage of non-responders, placebo responders, and drug responders in a clinical trial comparing Phenelzine with placebo in treatment of depression. An alternative is to use latent growth curve models (Muthen and Shedden 1999) to determine whether these posited response curves are actually represented by the data. In contrast to the semi-parametric function profile approach, latent growth curves allow subject-level posterior probabilities of placebo and drug response to be estimated. However, frequentist latent class growth curve approaches are limited in their a priori specification of functional forms. We consider a Bayesian approach that may allow for the a priori specification of the response profiles of non-responders, placebo responders, and drug responders.

e-mail: melliott@cceb.upenn.edu

INDIVIDUAL PATIENT DATA META-REGRESSION OF PLACEBO RESPONSE IN RECENT ANTIPSYCHOTIC EFFICACY TRIALS

Jeffrey A. Welge*, Department of Psychiatry
University of Cincinnati College of Medicine
Paul E.Keck, Jr., Department of Psychiatry
University of Cincinnati College of Medicine

Although the average response to placebo in clinical trials involving persons with schizophrenia is small compared to trials in other psychiatric conditions (e.g., unipolar major depression), substantial inter-patient variability remains unexplained. To determine whether patient-level covariates routinely conducted as part of such trials are associated with the degree of response to placebo, the Schizophrenia Clinical Trial Consortium for Modeling Placebo Response was formed to undertake a meta-regression of individual patient data. To date, data from twelve trials conducted to assess the short-term (< 8 week) efficacy of three atypical antipsychotics have been received. These trials contain longitudinal measurements on 764 patients who were randomized to receive placebo.

We will describe the contents of the database, strategies for model selection and validation, and treatment of missing data. Preliminary findings will be presented. Opportunities and obstacles specific to individual patient data meta-analysis will be considered.

e-mail: welgeja@email.uc.edu

---

31. Recent Developments in Bayesian Non/Semi-parametrics

SPATIAL NEUTRAL TO THE RIGHT PROCESSES

Lancelot F. James*, Hong Kong University of Science and Technology

Neutral to the right processes(NTR) were proposed by Doksum in 1974. Since then these processes have been a topic of considerable interest with particular applications to the field of Bayesian nonparametric Survival Analysis. Within this context NTR's serve as natural priors for the unknown survival distribution defined over the positive real line. Except for the case of the well known Dirichlet process and its variants an extension of the concept of NTR's to more abstract spaces is currently unavailable. In this talk we will describe how one may extend such models naturally which may be potentially useful in spatial survival models or survival models involving markers. Another important aspect of this talk will be a description of a new and considerably simpler approach to the dervivation of posterior quantities of Neutral to the Right processes and their spatial extensions.

e-mail: lancelot@ust.hk

POSTERIOR CONSISTENCY FOR REGRESSION MODEL

R. V. Ramamoorthi*, Michigan State University

In this talk I wil look at consistency of the posterior distribution in the regression model, when the distribution of the error is endowed with a nonparametric pror. The main tool is a variant of the Schwartz theorem.I will also discuss Doob's theorem in this context.

e-mail:

STATISTICAL VALIDATION OF COMPUTER MODELS

James O. Berger*, SAMSI AND DUKE UNIVERSITY

One of the major activities in science today is the development of math-based computer models of scientific and engineering processes. The most basic question in the evaluation of such computer models is "Does the computer model adequately represent reality?'' A computer model validation methodology with be discussed that is based on Bayesian nonparametric and spatial modeling. The approach is particularly suited to treating the major issues associated with the validation process: quantifying multiple sources of error and uncertainty in computer models; combining multiple sources of information; and updating validation assessments as new information is acquired. The framework has been implemented for two test bed computer models (a vehicle crash model and a resistance spot welding model) that will be used to illustrate the proposed validation process.

e-mail: berger@stat.duke.edu

A POISSON CUMULATIVE SUM APPROACH TO SPATIAL SURVEILLANCE

Peter A. Rogerson*, Department of Geography and National Center for Geographic Information and Analysis, University at Buffalo

We present an approach to spatial surveillance for incident clusters and bioterrorism using a Poisson cumulative sum approach. Spatial surveillance in this context uses geographical and temporal information about the location of each case to determine if an unusual clustering of cases exists; if so, appropriate authorities should be alerted. This kind of data has particular application to biological terrorism. As with the other presentations in the session, the approach is illustrated with a data set containing syndromic surveillance of a defined population of subjects from eastern Massachusetts.

Expected counts were modeled for the period 1996-1998 using logistic regression and using month, weekday/weekend, and a time trend as covariates. In general, weekends and summer months result in lower odds of office visits, in comparison with weekdays and winter months. The observed counts were then compared with the expected counts using the cumulative sum method. The small daily counts suggest that a Poisson cusum be employed, where expectations vary over time.

e-mail: rogerson@buffalo.edu

A SPACE-TIME PERMUTATION SCAN STATISTIC FOR SPATIAL DISEASE SURVEILLANCE

M Kulldorf*, University of Connecticut Health Center
F. Mostashari,
R. Heffernan,
J. Hartman,

We present and illustrate a space-time permutation based scan statistic for prospective disease surveillance for the early detection of localized disease outbreaks. The method is designed for situation when no denominator population at risk data is available, using only the temporal and spatial information of disease cases. Adjusting for the multiple testing inherent in the many potential times, locations and geographical size of an outbreak, it is able to both detect and evaluate the statistical significance of clusters found. Software implementing the method is available for public use.

e-mail: martink@neuron.uchc.edu

THE DISTRIBUTION OF INTERPOINT DISTANCES AND CLUSTER DETECTION

Marco Bonetti, Laura Forsberg, Al Ozonoff, Macello Pagano*, Harvard School of Public Health

The frequency distribution of the distances between individuals can be used to study the spatial distribution of individuals. It can thus be used to detect clustering in space and time. Since the presence of clustering might be indicative of a disturbance in what one considers normal behavior, these methods can be used as indicators of deviations from normalcy. Statistically, the number of distances increases with the square of the sample size. This is evidence of a complex dependency in the distribution of these distances. We investigate how to deal with this dependency and devise methods of testing for disease outbreaks.

e-mail: pagano@hsph.harvard.edu

---

33. Emerging Exact Methods in Biomedical Research

SMOOTH AND ACCURATE MULTIVARIATE CONFIDENCE REGIONS

Bo Yang, Shering—Plough Research Institute
John E.Kolassa*, Rutgers University

This paper describes multivariate approximate conditional confidence regions for canonical exponential families. These confidence regions have coverage probabilities that are better approximated by nominal levels than do traditional normal—theory regions, and have boundaries that are smoother than are obtained by inverting traditional exact tests. Our method is based on constructing one—dimensional conditional tests, combining p values, and inverting. More specifically, consider a statistical model with three parameters, of which two are of interest and one is not of interest. We generate a confidence region for the two parameters of interest by first generating a confidence interval for one of the parameters, conditional on sufficient statistics associated with the other interest parameter and the nuisance parameter. For values of this bounded parameter inside the confidence interval, determine a confidence interval for the remaining interest parameter conditional on the sufficient statistic associated with the nuisance parameter. This procedure determines the boundaries of a confidence region. This method is illustrated through applications to logistic and Poisson regression examples, in which parameters of interest are alternative representations of a single underlying physical quantity.

e-mail: kolassa@stat.rutgers.edu

ADJUSTING FOR BASELINE COVARIATES BY INDUCING A PARTIAL ORDERING

Vance W. Berger*, NIH
Judie Y.Zhou, Merck

In the context of randomized clinical trials, multiplicity arises in many forms. One prominent example is when a key endpoint is measured and analyzed both at baseline and after treatment. While it is common to analyze the baseline and subsequent measure separately, a more efficient approach is to adjust the post-treatment comparisons for the baseline values. Adjustment techniques generally treat the covariate (baseline value, in this case) as either nominal or continuous. Either is problematic when applied to an ordinal covariate, the former because it fails to exploit the natural ordering and the latter because it relies on an artifical notion of linear prediction and differences between values. We propose a new exact method for adjusting for ordinal covariates without having to treat them as either nominal or continuous. Specifically, we construct the information preserving composite endpoint, which consists of the pair of values for each patient, one at baseline and one after treatment, and determine which of these patterns indicate the most improvement. Some patterns are not comparable to others, resulting in only a partial ordering. We develop analyses to account for this.

e-mail: vb78c@nih.gov

PROVING NON-INFERIORITY/EQUIVALENCE WITH DICHOTOMOUS ENDPOINTS USING EXACT METHODS

Ivan S. F. Chan*, Merck Research Laboratories

Non-inferiority/equivalence studies are frequently conducted to evaluate the efficacy of new drugs and vaccines that have been shown to offer better safety profiles, easier administration, or lower cost compared to the standard treatment. Recently, exact methods for analyzing non-inferiority/equivalence trials with dichotomous (success/failure) endpoints have been developed to address the concern that existing asymptotic methods may fail because of small sample size or because of skewed or sparse data structure. In this talk we give an overview of the methodology and the critical issues associated with the exact methods for analyzing non-inferiority/equivalence studies. We emphasize the importance of choosing statistical methods that provide consistent inference between hypothesis testing (p-value) and confidence interval estimation. Then, we illustrate these exact methods with real clinical-trial examples and show that these exact methods are very useful tools for analyzing non-inferiority/equivalence trials, especially in situations where asymptotic methods are likely to fail. Finally, we mention the software available for these exact methods.

e-mail: Ivan_Chan@Merck.Com

COMPARISON OF EXACT TESTS AND CONFIDENCE INTERVALS FOR THE RELATIVE RISK (THE RATIO OF TWO BINOMIAL PARAMETERS)

Pralay Mukhopadhyay*, Dept. of Statistics, North Carolina State University
Roger L.Berger, Dept. of Statistics, North Carolina State University

Exact tests for the Relative Risk parameter (p1/p2) based on two independent binomial proportions are considered. These tests are desirable for studies with small to moderate sample sizes where asymptotic inference is inaccurate. We have proposed a test which is a modification of the standard exact unconditional test and is seen to perform much better than the standard test in many situations. Our proposed test is compared with the standard test with respect to size and power. We will invert these tests to get confidence intervals. Comparison of these exact intervals will be considered on the basis of lengths of the interval and coverage probability. An example will be considered where we will compare the efficacy of two vaccines.

e-mail: pmukhop2@stat.ncsu.edu

34. High-dimensional Hypothesis Testing and Model Selection for Functional Data Analysis

P-SPLINES AS A TOOL FOR SEMIPARAMETRIC LONGITUDINAL MODELLING

Naomi S. Altman*, Dept. of Statistics, Pennsylvania State University

Nonlinear Mixed Models are useful tools for modeling longitudinal data. When the nonlinear function is unknown, but the shape of the function is expected to be similar for all subjects, nonparametric regression can be used to model the shape, while retaining the mixed models formulation for the other effects. In this talk, the use of p-splines for fitting this type of 'self-modeling regression' with mixed effects is discussed, along with simulation and theoretical results for testing.

e-mail: nsa1@psu.edu

HIGH-DIMENSIONAL TESTING IN FUNCTIONAL DATA ANALYSIS WITH APPLICATIONS TO MEANS AND COVARIANCES

Dan J. Spitzner*, Virginia Tech, Department of Statistics

The analysis presented is of a data set generated for studying human tactile perception from digitized curves that represent subjects' perceptions of a moving brush stroke on the face. The interest is whether various facial preparations have measurable effects on perception. For purposes of testing, trend extraction and Fourier decomposition were applied together in an effort to represent the data in terms of statistically independent, one-dimensional components. Such preprocessing sets up a relevant framework in which numerous high-dimensional tests are available, including the test developed by Fan and Lin (1998) that was used to analyze these data. Nevertheless, basic descriptive analysis strongly suggests that straightforward trend-Fourier decomposition may not lead to independent components, which is, unfortunately, critical for the test to behave properly. Thus, the present analysis aims to understand the covariance structure within the data. In particular, principal components analysis is now adapted to explore covariance structures among model residuals, and similarly, the high-dimensional test of means is now adapted as a high-dimensional test of covariances.

e-mail: dan.spitzner@vt.edu

A BASIS FUNCTION APPROACH FOR THE ESTIMATION OF VARYING COEFFICIENT MODELS WITH LONGITUDINAL DATA

Jianhua Huang, Department of Statistics, University of Pennsylvania
Colin O.Wu*, Office of Biostatistics Research, DECA, National Heart Lung and Blood Institute
Lan Zhou, Center of Clinical Epidemiology and Biostatistics, University of Pennsylvania

We investigate a basis function approach for the stimation of time-varying coefficient curves in varying-coefficient models when the data are repeatedly measured over time. These coefficient curves describe the effects of the covariates, which may be either time-invariant or time-dependent, on the outcome of interest. Focusing specifically on the methods based on B-splines, we (a) propose a class of B-spline estimators based on least squares, (b) develop the consistency, rates of convergence and asymptotic distributions of these estimators, and (c) investigate through simulation the finite sample properties of confidence and testing procedures based on the asymptotic approximation and the 'resampling-subject-bootstrap'. Comparing with the well-known local smoothing procedures, the proposed global smoothing procedure has the advantages of (i) providing appropriate smoothing simultanously for all the coefficient curves, (ii) handeling nonparametric, semiparametric and parametric varying coefficient models in a flexible and unified way, and (iii) leading a simple framework for testing semiparametric and parametric submodels.

e-mail: wuc@nhlbi.nih.gov

VARIABLE SELECTION AND NONPARAMETRIC GOODNESS-OF-FIT TEST IN FUNCTIONAL DATA ANALYSIS

Jianqing Fan, Chinese University of Hong Kong
Runze Li*, Penn State University

Semiparametric regression models are very useful for longitudinal data analysis. The complexity of semiparametric models and the structure of longitudinal data pose new challenges to parametric inferences, variable selections and nonparametric goodness-of fit that frequently arise from longitudinal data analysis. In this paper, two new approaches are proposed for estimating the regression coefficients in a semiparametric model. The asymptotic normality of the resulting estimators is established. An innovative class of variable selection procedures is proposed to select significant variables in the semiparametric models. The proposed procedures are distinguished from others in that they simultaneously select significant variables and estimate unknown parameters. Rates of convergence of the resulting estimators are established. With a proper choice of regularization parameters and penalty functions, the proposed variable selection procedures are shown to perform as well as an oracle estimator. Furthermore, the generalized likelihood ratio test is introduced to test whether or not the baseline function can be fitted by a family of parametric models.

e-mail: rli@stat.psu.edu

35. Analysis of Linkage and Relationship

MULTIVARIATE LINKAGE ANALYSIS IN QUANTITATIVE TRAITS

Mariza de Andrade*, Mayo Clinic
Curtis Olswold, Mayo Clinic
Stephen T. Turner, Mayo Clinic

It has been shown that for correlated traits multivariate approaches for genetic linkage analyses can increase the power and precision to identify genetic effects. Therefore, using methods that can analyze several traits jointly is likely to enhance the ability to identify genes influencing the correlated traits. Allison et al. [1998] presented results from a large simulation study to assess the effectiveness of a bivariate H-E test for linkage versus the univariate H-E test. Their results showed that bivariate analyses can improve the power to detect linkage, with a greater gain in power when the genetic covariance due to a major locus linked to the marker studied is negative and the residual covariance among the traits is positive. Amos et al. [2001] also showed that bivariate approaches are more powerful than univariate analyses except for traits with very high positive polygenic correlation. Evans [2002] also reached similar conclusion. Our aim is to show that using combinations of correlated traits gives reliable results for linkage. We will present results from the Rochester Family Heart Study.

e-mail: mandrade@mayo.edu

MULTIPOINT LINKAGE DISEQUILIBRIUM MAPPING FOR COMPLEX DISEASES

Kung-Yee Liang, Departments of Biostatistics, Johns Hopkins Bloomberg School of Public Health
Fang-Chi Hsu*, Department of Public Health Sciences, Wake Forest University School of Medicine
Terri H. Beaty, Departments of Epidemiology, Johns Hopkins Bloomberg School of Public Health

Association study using trios has become one of the major approaches to locate susceptibility disease genes. With the availability of multiple dense markers, how to utilize information conveyed by multiple markers simultaneously remains challenging to investigators. Recently, Liang et al. (2001a) proposed a multipoint linkage disequilibrium (LD) method to estimate the location of the susceptibility gene from the framed region along with its sampling uncertainty. Two important features of this method are (i) the utilization of all trios whether parents' genotype are heterozygous or not and (ii) the availability of a single test statistic for no linkage or no LD to the framed region avoiding the multiple testing problem encountered when performing the transmission disequilibrium test (TDT) one marker at a time. In this talk, we discuss expanding this method to address important issues pertaining to complex disease in a unified fashion.Such issues include, among others, gene-gene and gene-environment interactions, genetic heterogeneity and phenotypic refinement.

e-mail: fhsu@wfubmc.edu

WEIGHTING OF PEDIGREES IN GENETIC LINKAGE ANALYSIS

Haydar Sengul*, Department of Human Genetics, University of Pittsburgh
Weeks E.Daniel, Department of Human Genetics, University of Pittsburgh
Eleanor Feingold, Department of Human Genetics, University of Pittsburgh

Nonparametric linkage methods for genetic mapping have been used frequently in recent years because they make no assumptions about the mode of inheritance of the trait. These methods generally involve calculating some kind of allele-sharing statistic for each pedigree in a dataset, then normalizing and summing over pedigrees. In theory, pedigrees of different sizes can be weighted differently in the sum, though it is most common to weight all the normalized pedigree statistics equally. Other common weighting schemes are based on the number of affected individuals in the pedigree. It may be possible to improve the power of allele-sharing statistics by developing more sophisticated weighting schemes. Under any particular trait model, optimal weights can be derived, but this is not useful because in practice wewill rarely know the true underlying trait model. It is desirable, instead, to find weighting schemes whose power is robust over large classes of models. We have used simulation and analytic methods to evaluate the power of several standard weighting schemes. We have also looked at several new weighting schemes, including some that incorporate information on how closely the affected individuals are related

e-mail: hsengul@hgen.pitt.edu

ESTIMATING AND APPROXIMATING EXPECTED LOG-LIKELIHOODS IN RELATIONSHIP INFERENCE

Solveig K. Sieberts*, Department of Statistics, University of Washington
Elizabeth A.Thompson, Department of Statistics, University of Washington

Expected log-likelihood differences (ELODs) are an effective way to assess the ability to infer the correct genetic model. In relationship inference, we wish to infer the correct relationship among a group of genotyped individuals, even under various model misspecifications. Because ELOD computations require taking expectations over multilocus genotypes for each individual, exact computations become infeasible for more than a few loci. Thus, it is impossible to assess the effect of a full chromosome of markers. Approximate ELODs can be calculated assuming a first order Markov dependence structure for data along the chromosome. This approximation reduces the problem to that of computing two locus ELODs, but ignores some of the data dependence. This can be especially detrimental when comparing relationships with the same marginal sharing expectations, where information comes from linkage. Another solution is to estimate ELODs using Monte Carlo simulations. These estimates account for the full dependencies among the data, but are subject to variation.

e-mail: solly@stat.washington.edu

POWER CALCULATIONS FOR FAMILIAL AGGREGATION STUDIES

Nusrat Rabbee*, Department of Biostatistics, Harvard School of Public Health
Rebecca Betensky, Department of Biostatistics, Harvard School of Public Health

Family studies are frequently undertaken as the first step in the search for genetic determinants of disease. In particular, significant familial aggregation of diseaseis suggestive of a genetic etiology for the disease and may lead to more focused genetic analysis. Many methods have been proposed in the literature for the analysis of family studies (e.g., Liang and Beaty (2000)). One model that is appealing for its simplicity of computation and the conditional interpretation of its parameters is the quadratic exponential model (e.g., Zhao and Prentice(1990), Betensky and Whittemore (1996), Hudson, et al (2001)). However, a limiting factor in its application, as well as that of the other proposed methods, is that power and sample size calculations have not been derived. These calculations are essential for investigators who are designing family studies. Here we derive analytic approximations for power for testing for familial aggregation, for both randomly sampled and non-randomly sampled families. We also present simulation studies of power for both the single and two disease cases, also under random and non-random sampling.

e-mail: nrabbee@hsph.harvard.edu

BAYESIAN VARIABLE-SELECTION WITH RELATED PREDICTORS AND ITS APPLICATION IN QUANTITATIVE TRAIT MAPPING

Cheongeun Oh*, SUNY at Stony Brook

We applied SSVS (stochastic search variable selection), a Bayesian model selection method, to two genetic problems. One is a QTL study on backcrossed mice model, in which we combined SSVS with pseudomarker imputation method by Sen and Churchill (2000), A statistical Framework for Quantitative Trait Mapping. The other is the simulated data of GAW 13. We used SSVS with Haseman-Elston method (2000), Haseman and Elston revisited, to find the markers linked to the change of cholesterol. In applying SSVS, we adopted prior structure first proposed by Chipman (1996), Bayesian Variable Selection with Related Predictors, to incorporate the relation among the predictors. This allows us to study gene-gene interaction (epistasis) since SSVS conducts a smart search in the model space. Traditional model selection method requires enormous computation in such situation. We also modified SSVS to deal with the highly correlated predictors which have been known as a major challenge for SSVS.

e-mail: cceoh@yahoo.com

36. Statistical Methods for Longitudinal Data

TEST FOR LONGITUDINAL DATA WITH SKEWED RESPONSES

Taesung Park*, University of Pittsburgh and Seoul National University
Jong-Hyeon Jeong, Department of Biostatistics, University of Pittsburgh
Sung-Hyun Kang, Department of Statistics, Seoul National University

In many clinical studies, the response information is often collected using questionnaires on an integer scale such as 1 through 6. A conventional approach is to treat the response variable as continuous and then to apply the maximum likelihood approach based on the normality assumption. When the distribution of the response variable is extremely skewed, however, the maximum likelihood approach may yield inefficient and invalid results. In this paper, we consider an analysis of longitudinal data with heavily skewed distributions of the responses. We propose a permutation test that does not depend on any distributionalassumption. Through simulation studies, we compare the powers and sizes of the proposed method with the maximum likelihood method. The proposed test is illustrated by a data set of quality of life from the National Surgical Adjuvant Breast and Bowel Project (NSABP).

e-mail: tspark@stats.snu.ac.kr

ON THE ANALYSIS OF LONGITUDINAL DATA MEASURED USING QUESTIONNAIRE INSTRUMENT

Shesh N. Rai*, St. Jude Children's Research Hospital
Shelly Lensing, St. Jude Children's Research Hospital
James M. Boyett, St. Jude Children's Research Hospital
Sean Phipps, St. Jude Children's Research Hospital

Longitudinal data encountered in practice are often 'messy', i.e. the data obtained in practice do not conform to the planned design. Therefore, inferences drawn from such studies need special attention. We critically evaluate some of the issues related specific to the longitudinal studies in terms of a St. Jude study in which the objective was to relate baseline psychometric status to longitudinally obtained quality-of-life indicators in pediatric cancer patients. Measuring quality of life or other psychosocial parameters often involves the use of a questionnaire instrument. Generally, responses to individual questions are assigned a score and then these scores are summed for a total scale or subscale score. Analyses of these data often involve parametric statistical models, which implicitly assume the underlying distribution of the data is normal. We propose a transformation of such data based on the logit so that the statistical assumptions involved in the analysis are better justified and illustrate its use in two St. Jude studies.

e-mail: Shesh.Rai@stjude.org

DETECTING DISEASE ONSET USING A LATENT CLASS MODEL FOR JOINT ANALYSIS OF LONGITUDINAL AND SURVIVAL OUTCOMES

Elizabeth H. Slate*, Department of Biometry and Epidemiology, Medical University of South Carolina
Liqiu Jiang, Department of Biometry and Epidemiology, Medical University of South Carolina

Lin et al. (2002, J. Amer. Statist. Assoc. 97:55-65) used a latent class model to uncover subpopulation structure with respect to the joint responses of longitudinal prostate-specific antigen (PSA) and time to prostate cancer (PCa) diagnosis. This model not only offers a way to handle heterogeneity in the evolution of a longitudinal biomarker (PSA) and accompanying disease (PCa) risk profile among subgroups (latent classes), but also provides for the potential for more efficient use of the longitudinal biomarker for detection of disease onset. The model naturally leads to a diagnostic rule for disease onset that uses the information from the associated longitudinal biomarker and other covariates up to the current time. We perform simulations to evaluate the performance of this rule. We then specialize to the PSA-PCa context and use ROC methodology to compare the new diagnostic rule with rules used in practice for detecting PCa, including a single elevated PSA reading or a large annual PSA velocity.

e-mail: SlateEH@musc.edu

THE USE OF SCORE TESTS FOR INFERENCE ON VARIANCE COMPONENTS

Geert Verbeke*, Biostatistical Centre, K.U.Leuven, Belgium
Geert Molenberghs, Center for Statistics, Limburgs Universitair Centrum, Belgium

Whenever inference for variance components is required, the choice between one-sided and two-sided tests is crucial. This choice is usually driven by whether or not negative variance components are permitted. For two-sided tests, classical inferential procedures can be followed, based on likelihood ratios, score statistics, or Wald statistics. For one-sided tests, however, one-sided test statistics need to be developed, and their null distribution derived. While this has received considerable attention in the context of the likelihood ratio test, there appears to be much confusion about the related problem for the score test. The aim of this paper is to illustrate that classical (two-sided) score test statistics, frequently advocated in practice, cannot be used in this context, but that well-chosen one-sided counterparts could be used instead. The relation with likelihood ratio tests will be established, and all results are illustrated in an analysis of continuous longitudinal data using linear mixed models.

e-mail: geert.verbeke@med.kuleuven.ac.be

LATENT TRAJECTORY MODELS FOR LONGITUDINAL DATA: A BAYESIAN APPROACH

Sujata M. Patil*, University of Michigan, Department of Biostatistics
Trivellore E.Raghunathan, University of Michigan. Department of Biostatistics
Jean T. Shope, University of Michigan Transportation Research Institute

Longitudinal data are often used to study individual developmental growth curves or 'trajectories'. These trajectories are not directly observed and require estimation. The analyses considered here explore the predictive relationship between latent trajectories and a future event. Analysis methods for these types of substantive questions proceed in two stages. In stage 1, individual latent trajectories are estimated and summarized by two or more latent trajectory variables. For example, in polynomial models the coefficients for linear or quadratic terms are the latent trajectory variables. In stage 2, the latent trajectory variables are used as predictors of a future event. In the traditional used approach, the stage 2 model conditions on the estimated trajectories and ignores the measurement error. This can affect inference. In a competing approach, parameter estimates and variances are corrected for measurement error. We develop a Bayesian approach that accounts for all uncertainties and allows for numerous extensions to the model not possible with the two competing approaches. We apply these three methods to study trajectories of adolescent alcohol use as predictors of vehicle crashes incurred during young adulthood. Results from a simulation study examining the three methods are also reported.
Supported by NIAAA grant and NIDA training grant.

e-mail: spatil@umich.edu

A JOINT CURE RATE AND LONGITUDINAL MODELS WITH APPLICATIONS TO CANCER VACCINE TRIALS

Elizabeth R. Brown*, Department of Biostatistics, University of Washington
Joseph G.Ibrahim, Department of Biostatistics, University of North Carolina, Chapel Hill

Complex issues arise when investigating the association between longitudinal immunologic measures and time to an event, such as time to relapse, in cancer vaccine trials. Unlike many clinical trials, we may encounter patients who are cured and no longer susceptible to the time-to-event endpoint. If there are cured patients in the population, there is a plateau in the survival function, S(t), after sufficient follow-up. If we want to determine the association between the longitudinal measure and the time-to-event in the presence of cure, existing methods for jointly modeling longitudinal and survival data would be inappropriate since they do not account for the plateau in the survival function. The nature of the longitudinal data in cancer vaccine trials is also unique as many patients may not exhibit an immune response to vaccination at varying time points throughout the trial. We present a new joint model for longitudinal and survival data that accounts both for the possibility that a subject is cured and for the unique nature of the longitudinal data. An example is presented from a cancer vaccine clinical trial.

e-mail: elizab@u.washington.edu

37. Survey Sampling: Theory and Applications

A SIMPLE METHOD FOR DETERMINING CONFIDENCE INTERVALS FOR POPULATION ATTRIBUTABLE RISK FROM COMPLEX SURVEYS

Sundar Natarajan*, Ralph H. Johnson VAMC & Medical University of South Carolina
Stuart R.Lipsitz, Medical University of South Carolina

In contrast to the other measures of association, population attributable risk (PAR) combines information on prevalence and a measure of association, usually relative risk, to provide a quantitative estimate of the proportion of disease in the population that is attributable to a particular exposure. While point estimates of PAR are valuable and provide an indication of central tendency, they do not provide insights regarding the precision of the estimates by quantifying uncertainty. Methods for computing confidence intervals associated with PAR are not readily available, particularly for data from complex surveys. In this presentation we will provide a simple, straightforward way to estimate confidence intervals for PAR based on the Bonferroni inequality and illustrate the method using simple random and complex samples.

e-mail: nataraja@musc.edu

WHEN IS STRATIFICATION DETRIMENTAL TO DESIGN? PART I

Gretchen Marcucci*, Rho, Inc.

Stratified random sampling is sometimes recommended if one wants to insure that important covariates are distributed evenly among the treatment groups. When the treatment allocations are balanced within each stratum, there is a gain in the precision of the estimates. Furthermore, if a goal of the study is to analyze the subgroups that define the strata, then stratified randomization has some clear advantages. However, stratification is not always practical, or even prudent. Under what circumstances is stratification a desirable design feature? This paper provides a literature review of the advantages and disadvantages of stratification. A particular application motivates this review: 120 subjects are to be randomized at 30 sites to one of 4 treatments should they be stratified into 2 strata, or not?

e-mail: gmarcucc@rhoworld.com

WHEN IS STRATIFICATION DETRIMENTAL TO A DESIGN? PART II

Katherine L. Monti*, Rho, Inc.

Consider the following situation: 120 subjects are to be recruited at up to 30 sites and randomized to one of 4 study drugs. A continuous covariate is considered to be predictive of the outcome. The distribution of the covariate in the study population is unknown. A client proposes to use stratified randomization, with the strata defined by dichotomized levels of the covariate. This stratification adds significantly to the cost of running the study. Is the stratification a good idea, assuming that central randomization is not an option? If the 120 subjects enroll so that there are 2 subjects per strata per site, then randomizing within site without stratification yields a perfectly balanced design, while randomizing within site with stratification could only unbalance the design. In more realistic enrollment patterns, could the stratification be beneficial? Motivated by this example, this paper uses simulations to explore the implications of using stratified randomization, assuming in a variety of possible reasonable enrollment patterns. Measures of between-treatment imbalance (over strata and within strata) and the loss of power for tests of hypotheses are explored.

e-mail: kmonti@rhoworld.com

VARIANCE-COVARIANCE FUNCTIONS FOR DOMAIN MEANS OF ORDINAL SURVEY ITEMS

A. James O'Malley*, Department of Health Care Policy, Harvard Medical School
Alan M.Zaslavsky, Department of Health Care Policy, Harvard Medical School
Department of Statistics, Harvard University

Estimates of a sampling variance-covariance matrix are required in many statistical analysis problems, particularly those with multi-level data. In univariate problems, a function relating the variance to the mean has been used to obtain variance estimates by pooling information across units. In this talk I will present variance and correlation functions for multivariate means of ordinal survey items. The cases of complete data and data with structured non-response are considered. Methods are also developed for assessing model fit, and for using sampling variances to optimally weight the direct and modeled estimates in a combined estimator.

e-mail: omalley@hcp.med.harvard.edu

COMPATIBLE ESTIMATORS OF THE COMPONENTS OF GROWTH FOR AN ANNUAL FOREST INVENTORY DESIGN

Francis A. Roesch*, USDA Forest Service, Southern Research Station

A compatible estimation system for the components of growth and instantaneous values for use with the USDA Forest Service's annual inventory design is given and evaluated. The sample design, a result of the 1998 Farm Bill, consists of a systematic triangular base grid of permanent plots containing a cluster of four circular subplots and annual measurement of a fixed proportion of these plots. We first address estimation for the traditional components of growth as presented by Meyer (1953) and show estimators that have been used with analogous designs to estimate those components for this design. Next we address the time invariant redefinition of the components of growth first given by Eriksson (1995), and how those components could be estimated from the annual inventory design.

e-mail: froesch@fs.fed.us

SAMPLING EXPERIMENTS FOR TEACHING STATISTICS

Michael J. Symons*, Biostatistics, University of North Carolina – Chapel Hill
Dana Quade, Biostatisics, University of North Carolina - Chapel Hill

In-class, real-time experiments are very effective and well received by students in our statistics classes. Examples of these simulation demonstrations, and variations thereof, are presented. Teaching points achieved with these exercises include: (a) bias in judgement sampling, (b) unbiasedness in simple random sampling, and (c) plausibility of the Central Limit Theorem, especially the decrease in variance of a sample mean with increasing sample size. Other topics equally well punctuated by simulation are the coverage of confidence intervals and the reality of Type I and Type II errors. Generally these demonstrations take five to ten times as much classtime as descriptions of these phenomena by lecturing with overheads. However, the student retention of the points being made seems much greater, and the student interest in such classes is so much greater, by a live demonstration over a traditional lecture. Topics to be covered during the semester, of course, limits the number of classes, or quater to half classes, that we are able to devote to these teaching techniques.

e-mail: symons@bios.unc.edu

38. Semi-and Nonparametric Methods in Survival Analysis

NONPARAMETRIC PAIRED TWO-SAMPLE TESTS FOR CENSORED SURVIVAL DATA WITH LONGITUDINAL COVARIATES

Shari Messinger*, University of Miami School of Medicine
Susan Murray, University of Michigan Department of Biostatistics

In this work, we present nonparametric two-sample tests for paired censored survival data incorporating longitudinal covariate information. These tests take advantage of information collected post-baseline to provide additional efficiency gains at each look when censoring is uninformative while accomodating paired designs. Additionally, these methods extend the ability to adjust for potential bias from informative censoring that is captured by the longitudinal covariates, as well as bias due to baseline covariate imbalances. Finite sample properties are investigated with simulation, and we illustrate methodology with example from the Early Treatment Diabetic Retinopathy Study.

e-mail: smessinger@med.miami.edu

MODEL ASSESSMENT AND OUTLIER DETECTION FOR A PSEUDOSPLINE-BASED SURVIVAL MODEL

Zekarias Berhane*, University of Pittsburgh
Joyce Chang, University of Pittsburgh
Lisa Weissfeld, University of Pittsburgh

Pseudosmoothers have been applied to time-to-event data, providing for a spline-based extension of the Cox proportional hazards model (Gray, 1994; Hastie, 1996). This approach allows for greater flexibility in modeling survival data and makes formal development of inferential procedures possible. Using this approach, we propose to extend the traditional approaches for model assessment and outlier detection used in linear regression to the pseudospline-based extension of the Cox proportional hazards model. We propose an estimate of the hat matrix and residuals and examine their usefulness for this model.

e-mail: ztbst1@imap.pitt.edu

NONPARAMETRIC ESTIMATION OF THE BIVARIATE SURVIVOR FUNCTION

Zoe Moodie*, Fred Hutchison Cancer Research Center
Ross L.Prentice, Fred Hutchison Cancer Research Center

While many nonparametric bivariate survivor function estimators are consistent and asymptotically Gaussian, van der Laan's NPMLE is the only one to claim efficiency. Despite its attractive asymptotic properties, it involves an iterative procedure that can be difficult to implement.

Based on the likelihood equations for the reduced data, we present an alternative expression to calculate van der Laan's NPMLE. This gives a simpler procedure for computing the estimator that is non-iterative in the special cases of censoring on only one failure time variate and censoring on both failure times at a common time. The new formulation also allows explicit calculation of a variance estimator for the bivariate survivor function estimator.

e-mail: zoe@scharp.org

SEMIPARAMETRIC BAYESIAN ANALYSIS OF THE PROPORTIONAL ODDS MODEL

Nibedita Bandyopadhyay*, Oakland University,
Ananda Sen, Oakland University,

Recurrent event data is a specific type of multivariate survival data in which subjects may experience repeated occurrences of the same type of event during follow-up period. Many extensions of survival models based on the Cox proportional hazards approach have been proposed to analyze recurrent event data. However, in many real-life situations the proportionality of the hazard ratios does not seem to be an appropriate assumption. The proportional odds model with its property of convergent hazard functions has been widely considered recently as an alternative to the proportional hazards formulation of survival data. We propose a semiparametric Bayesian approach of proportional odds model to analyze recurrent event data. The dependence between failure times of the same subject is induced by a frailty effect. In addition to inference for model parameters, we also discuss model selection issues. The proposed methodology is applied to survival data from a randomized clinical trial.

e-mail: nbandyop@oakland.edu

CONFIDENCE BOUNDS FOR THE RELATIVE RISK OF TWO BINOMIAL DISTRIBUTIONS WITH APPLICATION IN SURVIVAL ANALYSIS

Changyong Feng*, Department of Preventive Medicine, University of Kansas Medical Center

Relative risk is widely used in categorical data analysis to meansure the relative proportion of two binomial distributions. In this talk, we study the confidence bounds of the relative risk based on likelihood ratio test. We compare the coverage of this confidence interval with those of Wald, Agresti-Coull and Jeffrey confidence intervals. Simulation studies show that the coverage of the likelihood ratio based confidence interval is much better that those regulr confidence intervals. We use the likelihood based confidence interval to compare the survival distributions from a true clinical trial data.

e-mail: cfeng@kumc.edu

SURVIVAL ANALYSIS USING AUXILIARY VARIABLES VIA NONPARAMETRIC MULTIPLE IMPUTATION

Chiu-Hsieh Hsu*, University of Michigan
Jeremy M. G.Taylor, University of Michigan

We develop an approach, based on multiple imputation, to using auxiliary variables to recover information from censored observations to estimate the marginal survival distribution in survival analysis. To conduct imputation, we use two working proportional hazards models to define the imputing risk set. Based on the imputing risk set, two nonparametric multiple imputation methods are considered. In risk set imputation each censored time is replaced by a random draw of the observed times amongst those in the imputing risk set. In Kaplan-Meier imputation the imputed time is a draw from the estimated distribution of event times amongst those in the imputing risk set. In a situation with a binary covariate, we show that with a large number of imputes the estimates from both methods reproduce the weighted Kaplan-Meier estimator. We apply the approach to data from an AIDS clinical trial comparing ZDV and placebo, in which CD4 count is the time-dependent auxiliary variable. In a simulation study we show that the inclusion of a bootstrap stage in the multiple imputation algorithm improves the coverage rates of confidence intervals. In the situation with auxiliary variables, we show that the use of the multiple imputation methods can improve the efficiency of estimators and reduce bias due to dependent censoring. These findings of improved efficiency and reduced bias can be seen with both time-independent and time-dependent auxiliary variables.

e-mail: pchhsu@umich.edu

NONPARAMETRIC ESTIMATION FOR BASELINE HAZARD FUNCTION AND COVARIATE EFFECTS WITH TIME-DEPENDENT COVARIATES UNDER PROPORTIONAL HAZARD STRUCTURE

Feng Gao*, Emory University
Amita K.Manatunga, Emory University
Shande Chen, University of North Texas Health Science Center at Fort Worth

In many biomedical studies, it is important to know the hazard function. The Breslow's estimator is commonly used for the integrated hazard, but the estimator requires that the functional form of covariate effects be correctly specified. Especially with presence of time-dependent covariates, it is difficult to model a correct function of covariate effects from data. In this paper, a tree-type model is proposed which enables simultaneously estimating both baseline hazard function and the effects of time-dependent covariates. The proposed method approximates the baseline hazard and covariate effects with step-functions. The jump points in time and covariate space are searched via an algorithm based on the improvement of full log-likelihood function. Since the determination of these jump points is totally data driven, the proposed method in principle takes a non-parametric approach. In contrast to most other estimating methods, the proposed method estimates the hazard function rather than integrated hazard. The performance of the proposed method is evaluated by several simulation studies, and is also exemplified by applying to an anti-depression.

e-mail: fgao@sph.emory.edu

POPULATION INFERENCE IN CASE-CONTROL STUDIES STUDIES NESTED WITHIN A POPULATION-BASED COHORT

Charles B. Hall*, Albert Einstein College of Medicine

Nested case control studies involve sampling of controls from an established cohort. When that cohort is itself sampled from a defined population using representative weights, it is possible to infer back to the original population by taking account of the different sampling weights of the cases and controls. Examples are given using the Einstein Aging Study.

e-mail: chall@aecom.yu.edu

APPLYING A SURVEY SAMPLING PERSPECTIVE TO THE DESIGN AND ANALYSIS OF POPULATION-BASED CASE-CONTROL STUDIES

Ralph DiGaetano*, Westat
Lou Rizzo, Westat

A survey sampling perspective can provide useful guidance for the design and analysis of population-based case-control studies. A sample design should be developed to satisfy a study's analytic goals, taking into account the resources available for carrying out the study. The survey sampling literature provides tools for evaluating design trade-offs between such factors as precision and cost (for example, when making sample size decisions). Sampling methodologies appropriate for population-based controls include area sampling, RDD, or sampling from a list frame, utilizing stratification and multi-stage sampling. The control sample can be selected unmatched or matched to cases on auxiliary variables. Survey sampling analytic methods appropriate for a case-control study include the use of sample weights to reflect variable sampling rates as well as compensate for nonresponse and noncoverage. Weights can use either the general population or the case distribution as the reference population. Survey sampling software is available to estimate standard errors of study estimates, appropriately accounting for the complex sample design employed.

e-mail: ralphdigaetano@westat.com

USING SAMPLE SURVEYS TO OBTAIN A CONTROL GROUP: TWO RESEARCH EXAMPLES

Donna J. Brogan*, Biostatistics Dept., Rollins School of Public Health, Emory University

Simple or complex sample survey procedures may be used to obtain a control group for various research designs. The first example is a breast cancer case-control study of young women conducted in 1990-1992 in Atlanta, GA. All eligible diagnosed cases during the time period were included (i.e. no sampling). Two control groups were obtained, one by random digit dialing (RDD) and one by area probability sampling. Each control group was frequency matched to the cases on age. Analytic strategies are discussed for estimating odds ratios, and results are compared across the two different control groups.

The second example is a cross-sectional sample survey of women physicians in the U.S. in 1993. One research objective was to describe the health status of lesbian physicians. Two items on a self-administered questionnaire allowed classification of the probability sample of physicians into heterosexuals (the comparison group) and lesbians. Analytical strategies for using the comparison group, containing many more subjects than the lesbian group, are discussed. A second group of lesbian physicians was obtained from the membership of a gay/lesbian medical organization and compared to the lesbian group obtained from the probability sample of women physicians.

e-mail: dbrogan@sph.emory.edu

40. Multi-stage Decisions and Dynamic Treatment Regimes

ESTIMATION OF OPTIMAL TREATMENT STRATEGIES

Jamie Robins*, Harvard University

Consider observational data on the HIV infected patients receiving their health care at a large HMO. At each visit on the basis of a patient's treatment, clinical, and laboratory history, the patient's physician must decide whether to treat with anti-virals and if so the particular drugs and dosages to prescribe. In this talk, I describe a method for estimation of the optimal treatment protocol that maximizes quality adjusted survival time. I compare my method with that proposed by Susan Murphy in her seminal paper on this topic.

e-mail:

COMPARISON OF DESIGNS FOR ADAPTIVE TREATMENT STRATEGIES: BASELINE VS. ADAPTIVE RANDOMIZATION

Philip W. Lavori*, VA CSP/Stanford University
Ree Dawson, Frontier Science

Clinicians apply adaptive treatment strategies for chronic disease: at each patient visit the clinician decides whether to switch treatment, given outcomes to date. Previously we defined an adaptive randomization scheme for evaluating such strategies. In this scheme, a patient's treatment may be switched at one of the follow-up times with probability that depends on the patients observed outcomes under the initial treatment. These randomized switches generate treatment sequences thatprovide complete or partial data on adaptive treatment strategies of interest. We compare this approach to aseline randomization, where each patient is assigned to one of the adaptive strategies, and prove that there exists an adaptive design with the same efficiency as baseline randomization. We exploit randomized dropout to obtain causal comparisons beyond the original class of strategies, illustrating the flexibility of the adaptive design for evaluating adaptive treatments. We use the theory of multiple imputation to characterize the potential loss of precision due to drop-out induced by the more flexible design.

e-mail: Philip.Lavori@med.va.gov

SIMULATION-BASED SEQUENTIAL BAYESIAN DESIGN

Peter Mueller*, U. TX M.D. Anderson Cancer Center

We consider simulation-based methods for exploration and maximization of expected utility in sequential decision problems. We consider problems which require backward induction with analytically intractable expected utility integrals at each stage. We propose to use forward simulation to approximate the integral expressions, and a reduction of the allowable action space to avoid problems related to an increasing number of possible trajectories in the backward induction. The artificially reduced action space allows strategies to depend on the full history of earlier observations and decisions only indirectly through a
low dimensional summary statistic. We illustrate the proposed approach with an application to an optimal stopping problem in a clinical trial.

e-mail: pm@odin.mdacc.tmc.edu

LOCATION-SCALE MODELS FOR HIERARCHICAL CATEGORICAL DATA

Donald Hedeker*, University of Illinois at Chicago

Mixed-effects logistic regression models are described for analysis of hierarchical ordinal outcomes. In terms of the random effects, these can be the same for all subjects or allowed to vary by subject groups. Additionally, two extensions to the usual proportional odds assumption are described and compared. The first permits separate covariate effects to be estimated for each of the C-1 cumulative logits (where C = number of categories). The second extension allows covariates to influence the scale of the ordinal response, in addition to their usual influence on the location. This latter extension can be more parsimonious since it adds only one parameter for each covariate, and can be used to examine the effect of subject grouping variables on both the within- and between- sujbects variance. A maximum marginal likelihood (MML) solution is described utilizing numerical quadrature to integrate over the random-effects distrbution. An analysis is presented of a dataset from an adolescent smoking study, highlighting and comparing both extensions of the proportional odds mixed model.

e-mail: hedeker@uic.edu

A MARGINALIZED MULTILEVEL MODELING APPROACH FOR CLUSTERED LONGITUDINAL BINARY DATA WITH TIME DEPENDENT COVARIATES

Diana L. Miglioretti*, Center for Health Studies, Group Health Cooperative
Patrick J.Heagerty, Department of Biostatistics, University of Washington

Modeling multilevel binary data continues to pose significant challenges in many situations. For short time series typical in longitudinal epidemiological studies, hierarchical or random effects models using the standard assumption of normally distributed cluster-specific effects are difficult to fit and may not be reasonable. In addition, interest is often in the marginal or population-average effects, which are not directly modeled by hierarchical models. We propose a marginalized multilevel model that combines a logistic regression model to estimate the influence of covariates on the marginal mean and a separate conditional logistic regression model that captures both the serial dependence within short time series and the correlation within larger clusters. Extensions for endogenous time-dependent covariates will be discussed. The model is estimated using a Bayesian approach. The methods will be illustrated using data from the Breast Cancer Surveillance Consortium to estimate mammography accuracy in a repeatedly screened population.

e-mail: miglioretti.d@ghc.org

ON RECENT DEVELOPMENTS IN GRAPHICAL MODELS

Nanny E. Wermuth*, University of Mainz, Germany

Graphical models generalize path analysis models for only linear relation to arbirtrary distributions, permitting in particular interactive effects and nonlinear relations, as well as response variable, intermediate variable and purley explanatory variables whic may be quantiative (numerical) or qualitative (categorical).

In all graphical models variables are represented by nodes in a graph and strong direct asociations by an edge connecting the node pair. Different type of edge and graph provide flexibility in formulating conditional independencies that may be directly read off the graph.

For some graphical model classes it is possible to derive consequences for arbitrarily selected conditional distributions. This is a most desirable feature not only for causal inference, as stressed by R. Fisher but alsofor an improved understanding of the statistical models. An overview on recent results is given.

e-mail: nanny.wermuth@uni-mainz.de

42. Health Effects and Environmental Risk Assessment of Air Pollution

COVARIATE-ADJUSTED SPATIAL CDFS FOR AIR POLLUTANT DATA

Margaret Short, University of Minnesota
Bradley P.Carlin*, University of MInnesota
Alan E. Gelfand, Duke University

A spatial cumulative distribution function (SCDF) gives the proportion of a spatial domain D having the value of some response variable less than a particular level w. In this paper we provide a fully hierarchical approach to SCDF modeling, using a Bayesian framework implemented via Markov chain Monte Carlo (MCMC) methods. The approach generalizes the SCDF to accommodate block-level variables, possibly utilizing a spatial change of support model within an MCMC algorithm. We then extend our approach to allow covariate weighting of the SCDF estimate. We further generalize the framework to the bivariate random process setting, which allows simultaneous modeling of both the responses and the weights. Once again MCMC methods (combined with a convenient Kronecker structure) enable straightforward estimates of weighted, bivariate, and conditional SCDFs. We illustrate our methods with two air pollution data sets, one concerning ozone exposure and race in Atlanta, GA, and the other recording both NO and NO$_{2}$ ambient levels at 67 monitoring sites in central and southern California.

e-mail: brad@biostat.umn.edu

SHORT-TERM RESPIRATORY HEALTH EFFECTS OF AIR POLLUTION IN METROPOLITAN CHICAGO

Vanja Dukic*, University of Chicago, Department of Health Studies and Center for Integrating Statistics and Environmental Science (CISES)
Paul Rathouz, Department of Health Studies and CISES, University of Chicago
Dana Draghicescu, CISES, University of Chicago
Gidon Eshel, CISES, University of Chicago
John Frederick, CISES, University of Chicago
Ted Naureckas, CISES, University of Chicago
Alexis Zubrow, CISES, University of Chicago

Most recent studies of environmental effects on respiratory disorders have focused on aggregate analyses, with outcomes and exposure variables summarized by daily averages of available monitors over entire region of interest. In this project we use three time series of asthma-related outcomes in the Chicago metropolitan area (daily hospitalization counts, emergency department visits, and Albuterol purchases), aggregated at the ZIP code level, to identify linkages with measures of air quality at the spatial scale of individual neighborhoods, while allowing for neighborhood-specific effects. In addition, in the dataset at our disposal, each patient is given a unique identifying number, making it possible to examine subject-specific models of repeated asthma-related occurrences over the period of four summers.

This research is supported by EPA grant R-82940201-0. The abstract does not necessarily reflect EPA views.

e-mail: vdukic@health.bsd.uchicago.edu

CASE-CROSSOVER ANALYSIS OF THE HEALTH EFFECTS OF AIR POLLUTION

Holly E. Janes, University of Washington
Lianne Sheppard*, University of Washington
Thomas Lumley, University of Washington

The case-crossover study design allows researchers to assess the association between a short-term exposure and the risk of an acute health effect, while controlling for confounders by design. Using data from cases only, exposure at the event time is compared to exposure at comparable “referent” times. The referent selection strategy is critical. First, if the referents are completely determined by the event time, the conditional logistic regression estimating equations are biased. Second, if exposure at the referent times is not representative of exposure at the event time, there will be time-dependent confounding. With proper referent selection, however, there is no bias in the estimating equations, and both time-dependent and time-independent confounders are controlled for by design. In analyses of air pollution data, it is common to use a shared exposure series for all cases. We evaluate the bias, conditional on a fixed, shared exposure, associated with different referent selection strategies and plausible exposure series.

e-mail: sheppard@u.washington.edu

CONTROL OF THE FALSE DISCOVERY RATE FOR RANDOM SEARCH PROCEDURES

Kenneth M. Boucher*, Huntsman Cancer Institute and Dept. of Oncological Sciences, University of Utah
Aniko Szabo, Huntsman Cancer Institute and Department of Oncological Sciences, University of Utah

Microarray technology allows for simultaneous measurement of the expression of thousands of genes from a single sample. A typical study design is to obtain a relatively small number of microarrays from two or more cell types, with the goal of finding genes that are differentially expressed in the different tissues. Since the genes work together in genetic pathways, it may be desirable to find sets of genes that are differentially expressed. In this setting, it may be impractical or even impossible to test every set of genes. One strategy is to choose an appropriate multivariate distance function, and then to repeatedly search for a set of genes that maximizes the distance between the cell types. When employing random search, many thousands of comparisons may be made in the process of finding a single set of differentially expressed genes. Methods to control the family-wise error rate may result in an undesirable loss of power. We present a method to control the false discovery rate for such a random search procedure.

e-mail: ken.boucher@hci.utah.edu

GENOME ASSOCIATION STUDIES OF COMPLEX DISEASES BY CASE-CONTROL DESIGNS

Ruzong Fan*, Department of Statistics, Texas A&M University
Michael Knapp, Institute of Medical Biometry, Informatics and Epidemiology, University of Bonn

One way to perform LD mapping of genetic traits is to use individual single marker. Since dense marker maps such as SNPs are available, it is natural and practical to generalize one single marker LD mapping to high resolution multipoint LD mapping. This report investigates high resolution LD mapping methods of complex disease based on haplotype maps. Based on two coding methods, Genotype Coding and Haplotype Coding, Hotelling's T^2 statistics T_G and T_H are proposed to test the association between a disease locus and two haplotype blocks. The validity of the two T^2 statistics is proved by theoretical calculations. An extensional statistic T_C of traditional \chi^2 method of comparing haplotype frequencies is introduced by simply adding two \chi^2 test statistics of the two haplotype blocks together. The merit of the three methods is explored by power and type I error calculation and comparison. For each of the three statistics, the power of using two haplotype blocks is higher than that of using only one haplotype block. By power comparison, we notice that T_C has higher power than that of T_H, and T_H has higher power than that of T_G. In the absence of LD between the two blocks, the power of T_C is similar to that of T_H, and higher than that of T_G. In the presence of LD between the two blocks, the type I error of T_C is higher than those of T_H and T_G. Hence, we advocate to use T_H in the data
analysis.

e-mail: rfan@stat.tamu.edu

USE OF TREE MODELS TO INVESTIGATE GENE-GENE AND GENE-ENVIRONMENT INTERACTIONS IN LUNG CANCER RISK

Carol J. Etzel*, Department of Epidemiology, UT MD Anderson Cancer Center
Hui Zhao, UT Health Science Center
Yumei Cao, UT Health Science Center
Xifeng Wu, Department of Epidemiology, UT MD Anderson Cancer Center
Qingyi Wei, Department of Epidemiology, UT MD Anderson Cancer Center
Margaret R. Spitz, Department of Epidemiology, UT MD Anderson Cancer Center

In this investigation we developed tree models for case-control data to identify subgroups of individuals with high-risk for disease and identify possible gene-gene and gene-environment interactions. We grew classification trees and considered different splitting and pruning rules to grow optimal trees. We applied these methods on data from a case-control lung study including 1293 newly diagnosed lung cancer patients recruited from UT M. D. Anderson Cancer Center, Houston and 1091 healthy controls. Demographic, smoking, environmental exposure, and family cancer history and genetic measures were obtained on all participants. The resulting tree model revealed that individuals who were and were not sensitive to Bleomycin-induced chromosome breaks (BICS) had different observed risk for lung cancer. Exposure to dusts increased the risk of lung cancer among the BICS sensitive group. DNA repair capacity and smoking history modulated the risk of cancer among the non-sensitive BICS group. When we focused on former smokers, we observed that those with exposure to dusts and sensitive to BICS were at higher risk. However, cessation of smoking for more than 32 years proved to be protective. This investigation shows the importance of using tree models in the identification of high-risk subgroups and uncovering possible interactions among risk factors.

e-mail: cetzel@mdanderson.org

MEASURING THE EFFECT OF APO-E GENOTYPES ON CARDIOVASCULAR DISEASE EVENTS IN THE FRAMINGHAM HEART STUDY

Mark E. Glickman*, Department of Mathematics and Statistics, Boston University
Mei Fang Kao, Department of Mathematics and Statistics, Boston University

Many late-onset diseases are caused by what appears to be a combination of a genetic predisposition to disease, and health and environmental factors. One approach to making inferences about genetic factors has been developed by Glickman (2002). The framework describes the role of genetic and risk factors on the onset ages of multiple diseases, and accounts for the possibility that an individual was censored for reasons related to the diseases of interest. The framework also allows for missing genetic information, so that subjects censored prior to genetic sampling, and therefore missing such information, may still be included in the analysis. We examine extensions to this framework and the fit of various models to the original cohort of the Framingham Heart Study for measuring the effects of different Apolipoprotein-E (Apo-E) genotypes on the occurrence of various cardiovascular disease events. In particular, we will discuss the tradeoffs between univariate versus multivariate onset age components to the model, whether to incorporate health covariates measured at baseline or at a point later in the study, and whether to assume a heritability model for Apo-E genotype frequencies.

e-mail: mg@bu.edu

MULTIVARIATE CONDITIONAL ANALYSIS FOR COMPLEX TRAITS

Jixiang Wu*, Department of Plant & Soil Sciences, Mississippi State University
Dongfeng Wu, Department of Mathematics and Statistics, Mississippi State University
Johnie N. Jenkins, USDA-ARS, Mississippi State
Jack C. McCarty Jr., USDA-ARS, Mississippi State

A complex trait like crop yield is determined by several component traits. Multivariable conditional analysis in a general mixed linear model is helpful in dissecting the gene expression for the complex trait from different effects such as treatment, genotype, and genotype environment interaction. In this paper, an approach is presented for constructing a new random vector which can be equivalently used to analyze multivariable conditional variance components and conditional effects. End of season plant mapping data including lint yield and three yield components for nine cultivars of upland cotton (Gossypium hirsutum L.) were used to detect conditional variance components and conditional effects using this new approach.

e-mail: jw7@ra.msstate.edu

BAYESIAN INFERENCE OF POPULATION STRUCTURE FROM DOMINANT MARKERS USING MIXTURE OF BETAS

Rongwei Fu*, Dept. of Statistics, University of Connecticut
Dipak K.Dey, Dept. of Statistics, University of Connecticut
Kent E. Holsinger, Dept. of Ecology and Evolutionary Biology, University of Connecticut

Molecular markers derived from polymerase chain reaction (PCR) amplification of genomic DNA provide important information for accessing genetic diversity. Random amplified polymorphic DNA markers (RAPDs) allow analysis of species for which previous DNA sequence information is lacking, but dominance makes it impossible to apply standard techniques to calculate F-statistics. Bayesian hierarchical models make it possible to estimate FST directly from dominant markers. However, previous Bayesian analysis had assumed that allelle frequency distributions at one locus are independent betas across populations, which is not often true. In this study, we use mixture of betas to model the allele frequency to incorporate the correlation among populations. Inference of FST and FIS are performed and compared with independent beta models. We also investigate methods identifying and estimating different FST across loci. We illustrate the method with RAPD data from 14 populations of a North American orchid, Platanthera leucophaea.

e-mail: rongwei@stat.uconn.edu

ESTIMATING THE NUMBER OF FALSE NULL HYPOTHESES IN A MULTIPLE-TEST SITUATION

Dan Nettleton*, Department of Statistics, Iowa State University
J. T. Gene Hwang, Department of Mathematics, Cornell University

It is often scientifically interesting to obtain an estimate of the number of false null hypotheses when many tests are conducted. For example, researchers scanning for associations between markers and quantitative traits may wish to estimate the number of markers that are linked to quantitative trait loci. Researchers conducting microarray experiments may want to estimate the number of genes that change transcription rate in response to a treatment. While such estimates are interesting in their own right, they can be very important for effectively managing false discovery rates. Several proposed methods for estimating the number of false null hypotheses will be discussed. We will show that one intuitively appealing iterative method for estimating the number of false null hypotheses can be computed directly without iteration. We will compare through simulation the performance of this method with others that have been recently proposed in the literature.

e-mail: dnett@iastate.edu

44. Clinical Trials: Adaptive, Group Sequential Designs, and Interim Analysis

MID-COURSE CORRECTION TO GROUP SEQUENTIAL CLINICAL TRIALS

Cyrus R. Mehta*, President, Cytel Software Corporation
Anastasios A.Tsiatis, Professor, North Carolina State University

Sample size calculations based on initial guesses of key design parameters like the variance may lead to over or underpowered studies. We argue that the precision with which a treatment effect is estimated is, directly related to the statistical information in the data. It is thus possible to be flexible on sample size but rather continue collecting data until we have achieved the desired information. Such a strategy is well suited to being adopted in conjunction with a group sequential clinical trial where the data are monitored routinely anyway. We contrast our approach with adaptive designs, which allow the sample size to be modified based on sequentially computed observed treatment differences.

e-mail: mehta@cytel.com

SCPRT METHODS FOR CLINICAL TRIALS

Xiaoping Xiong*, St. Jude Children's Research Hospital
Ming Tan, University of Maryland Greenebaum Cancer Center
James Boyett, St. Jude Children's Research Hospital

Sequential conditional probability ratio tests (SCPRTs) is a class of sequential and group sequential test procedures derived from the ratio of maximum conditional likelihoods which are conditioned on the possible ending values of test statistic at the last stage of planned test. This class of sequential test procedures differs from other sequential test procedures by having the property that a conclusion made at early stopping is unlikely to be reversed if the trial had continued to the planned end. This property provides strongest rational for early stopping of clinical trials when data at early stages of trial provides enough information to foretell the conclusion of test at the planned end beyond doubt. We will introduce SCPRT methods for clinical trials and demonstrate design and analysis of clinical trials with SCPRTs using a window based computer program.

e-mail: xiaoping.xiong@stjude.org

AN INTERIM MONITORING STRATEGY FOR A SMALL SAMPLE INCIDENCE DENSITY PROBLEM

Shane L. Rosanbalm*, Rho, Inc.

In a recently proposed study of HIV positive patients with a history of cryptococcal meningitis, researchers were interested in determining if patients with reconstituted immune systems could safely discontinue their prophylactic antifungal medication. Because the disease of interest is potentially fatal and because the study was expected to enroll solwly, multiple interim looks were contemplated. Published approaches to interim monitoring assume that one has a sequence of multivariate normally distributed test statistics. However, for this study the Poisson mean (under all hypotheses of interest) is too small to presume a normal approximation to the Poisson. This research proposes an adaptation of Whitehead's method for conducting interim monitoring in this study. Simulation studies were conducted under a variety of settings. The resulting Type I error rate was conservative, target power levels were typically attained or surpassed, and mean study duration was reduced substantially if the risk of recurrence is extremely low.

e-mail: srosanba@rhoworld.com

ESTIMATING END OF TRIAL VARIANCE IN INTERIM ANALYSES OF CLINICAL TRIALS OF FIXED DURATION

Grant Izmirlian*, National Cancer Institute, Division of Cancer Prevention
Richard Fagerstrom, National Cancer Institute
Philip C. Prorok, National Cancer Institute

Since many trials are of fixed duration, many designs use group sequential methods for early stopping in favor of an effect, and stochastic curtailment for early stopping for no effect. In the cancer screening literature, there is a strong consensus for using cancer specific mortality as the primary endpoint. Such trials exhibit non-proportional hazards, since deaths occurring earlier in the trial are likely to be few in number and roughly balanced in the screened and control arms due to the exclusion of pre-existing cancers by design. Therefore, a weighted log-rank analysis with weighting function giving preference to deaths occurring later in the trial is preferable. Since maximum duration trials require an estimated end of trial variance for the weighted log-rank statistic, a problem arises as to which of markedly different estimates to use. We will use published screening trial data to show how these problems may be avoided.

e-mail: Izmirlian@nih.gov

A BAYESIAN SEQUENTIAL PROCEDURE FOR MULTIPLE TREATMENT OUTCOMES

Maria K. Mor*, Dept. of Biostatistics, University of Pittsburgh Graduate School of Public Health,
Stewart J.Anderson, Dept. of Biostatistics, University of Pittsburgh Graduate School of Public Health,

Evaluating treatment effectiveness often requires using information from multiple endpoints. Making decisions can be problematic in the presence of conflicting effects (e.g., increased benefit coupled with increased side effects). Typically, ad hoc procedures are used to assess overall effectiveness when the endpoints are not in agreement. However, these procedures are usually not incorporated into the study design. Furthermore, during the course of a study, one may be required to make interim decisions to determine whether the study should continue or if there is sufficient evidence to discontinue.

A multivariate Bayesian sequential procedure is proposed to facilitate a decision making process at multiple times. Utility functions are used to formalize the relationship between two variables and provide a framework for making decisions even when the variables demonstrate conflicting effects. We show how this procedure works with a hypothetical example in a cancer clinical trial that tests two therapies where one therapy has superior relapse-free survival but is inferior with respect to life threatening toxicities. Up to two interim looks were considered in our example.

e-mail: mormk@msx.upmc.edu

ESTIMATING MEAN RESPONSE AS A FUNCTION OF TREATMENT DURATION IN AN OBSERVATIONAL STUDY, WHERE DURATION MAY BE INFORMATIVELY CENSORED

Brent A. Johnson*, Department of Statistics, North Carolina State University
Anastasios A.Tsiatis, Department of Statistics, North Carolina State University

After a treatment is found to be effective in a clinical study, attention often focuses on the effect of treatment duration on outcome. In many studies, the treatment duration, within certain limits, is left to the discretion of the investigators. It is often the case that treatment must be terminated prematurely due to an adverse event, in which case a recommended treatment duration is part of a policy that treats patients for a specified length of time or until a treatment-censoring event occurs, whichever comes first. Evaluating mean response for a particular treatment duration policy from observational data is difficult due to censoring and the fact that it may not be reasonable to assume patients are prognostically similar across all treatment strategies. We propose an estimator for mean response as a function of treatment duration policy under these conditions. The method uses potential outcomes and embodies assumptions that allow consistent estimation of the mean response. The estimator is evaluated through simulation studies and demonstrated by application to the ESPRIT infusion trial coordinated at Duke University Medical Center.

e-mail: bajohns4@stat.ncsu.edu

SENSITIVITY ANALYSIS FOR LONGITUDINAL CLINICAL TRIALS

Herbert Thijs*, L.U.C. Center for Statistics
Geert Molenberghs, L.U.C. Center for Statistics
Carolien Beunckens, L.U.C. Center for Statistics
Ivy Janssen, L.U.C. Center for Statistics
Craig H. Mallinckrodt, Eli Lilly and Company
Raymond J. Carroll, Texas A&M University

Currently in clinical trials, standard methodology used to analyze longitudinal data subject to non-response is mostly based on the MCAR assumption, including Last Observation Carried Forward (LOCF) and Complete Case analysis (CC), mostly without assessment of possible influence on the final results. On the other hand, for a MAR dropout process, a so-called ignorable analysis is valid. However since reasons for dropout are varied it is difficult to justify on a priori grounds the assumption of random dropout. Furthermore, simple methods of analysis do not necessarily imply simple assumptions, and without understanding properly the assumptions being made in an analysis, we are not in a position to judge its validity or value. For the above-mentioned reasons it is considered vital to useful the discussion towards treating missing data in longitudinal clinical experiments. Given the importance of the possible inaccuracies when using overly simple methods and given the availability of more elaborate ways to deal with this problem we introduce various informal and formal ways to perform sensitivity analysis to provide useful insight in model uncertainty and nonrandom selected samples.

e-mail: herbert.thijs@luc.ac.be

COMPOSITE LIKELIHOOD BAYESIAN CLUSTER MODELLING OF SMALL AREA HEALTH DATA

Andrew B. Lawson*, Department of Epidemiology and Biostatistics,
Arnold School of Public Health, University of South Carolina

In this presentation the basic foci of small health data analysis are reviewed and their interrelation with environmental risk assessment is considered. In particular the area of cluster detection is one where there are strong links with environmental issues, specifically environmental epidemiology. Examples of the analysis of the spatial distribution of relative risk in the vicinity of putative health hazards such as incinerators or landfill sites, are many. Here I will focus on some newer developments in the area of cluster analysis where the prior focus (e.g. a putative source|) is not known. The approach described examines the use of Bayesian composite/pseudolikelihood models for clustering. Unlike some recent proposals for the analysis of clustering, these models do not require the use of birth-death reversible jump MCMC and so are easier to interpret. These models are very flexible and have close links to nonparametric approaches to smoothing risks. I will present some results from simulation studies and also a data example will be examined.

e-mail: alawson@gwm.sc.edu

USING THE K-NEAREST NEIGHBOR TECHNIQUE TO CLASSIFY SATELLITE IMAGERY

Ronald E. McRoberts*, North Central Research Station, USDA Forest Service

The k-Nearest Neighbors (k-NN) technique is an intuitive, easy to implement, non-parametric method for predicting values of attributes using a set of training data with known values of the attributes. Methods for calibrating the technique, illustrations of potential pitfalls, and techniques for increasing the efficiency of implementation are discussed. The particular application is construction of a forest/non-forest map using Landsat Thematic Mapper satellite imagery and data from geo-referenced forest inventory plots.

e-mail: rmcroberts@fs.fed.us

A DATA AUGMENTATION PROCEDURE FOR THE ANALYSIS OF CENSORED SPATIAL DATA

Brooke L. Fridley*, Iowa State University
Philip Dixon, Iowa State University

The analysis of spatially correlated data involving observations falling below a detection level is a problem in many environmental applications. Imputation of the censored observation with the level of detection (DL) or some function of the DL, like DL/2, is a common practice. The resulting parameter and standard error estimates found with the use of this single imputation method introduce bias. A data augmentation procedure is presented for the analysis of spatially censored data within a bayesian framework. Comparison of the data augmentation approach to handling censored data as compared to simple imputation methods will also be discussed. To demonstrate the data augmentation method, the analysis of contaminated soil site will be illustrated.

e-mail: bfridley@iastate.edu

A SPATIAL ANALYSIS OF SEA TURTLE NESTING PATTERNS IN PALM BEACH COUNTY, FLORIDA

Traci L. Leong*, Dept of Biostatistics, Rollins School of Public Health, Emory University
Andrew Barclay, Dept of Biostatistics, Rollins School of Public Health, Emory University
Lance Waller, Dept of Biostatistics, Rollins School of Public Health, Emory University

We explore variations in the observed spatial pattern of sea turtle nesting behavior at Juno Beach, Palm Beach County, Florida for the 1998-2000 nesting seasons. Of particular interest is an assessment of possible effects due to a 990-foot fishing pier constructed in 1998-1999. The data include approximately 8,000-10,000 emergence locations per nesting season over 6 miles of beach with locations identified by global positioning system (GPS) units with sub-meter accuracy. Typical statistical analyses suggest significant changes between years in counts of emergences by zone but do not identify exactly where the significant differences lie. We conduct a spatial analysis by estimating the density of emergences (number of emergences per unit length of beach) as a function of beach location, and compare densities of emergences between nesting seasons, densities of nesting and non-nesting emergences within each season.The approach reveals significant decreases in emergence density near the pier in the first post construction year (1999) in contrast to 1998 and 2000. The approach also reveals a possible distributional shift in nesting locations in the second year post construction even though total emergence counts are similar. Finally, the approach suggests an impact of the pier on nesting behavior, i.e. a reduced probability of nesting per emergence in the immediate vicinity of the pier.

e-mail: tlstat@hotmail.com

PREDICTION OF COUNTY CANCER RATES BY HIERARCHICAL SPATIAL MODELS

Linda W. Pickle*, National Cancer Institute

As part of an ongoing program of methodologic studies to estimate cancer incidence, prevalence, and survival rates at a small area level, we present the results of a hierarchical spatial model that predicts the number of incident cancer cases expected in each U.S. county. These counts are useful measures of the cancer burden and provide information with which to monitor cancer trends, plan cancer control activities and assess completeness of registry reporting. We have modeled age- and county-specific incidence data from 480 counties included in the NCI SEER cancer registry program as a function of each case's age, race, and sex and the corresponding county-specific mortality rates, sociodemographic and lifestyle factors. Resulting regression coefficients were used to predict incidence in all other counties. These models were developed and validated using random subsets of SEER county data, where observed data were available for comparison. Methodologic issues to be discussed include the development of small area lifestyle factor estimates from BRFSS data and inclusion of errors-in-covariates terms. Final predictions for five broad tumor groupings in 1999 are presented as state and smoothed county maps and are compared to state registry data reported to CDC and previously-published predictions from the American Cancer Society.

e-mail: picklel@mail.nih.gov

BAYESIAN HIERARCHICAL SPATIAL-TEMPORAL MODELS FOR WIND PREDICTION

Li Chen*, Statistics Department, NCSU
Montserrat Fuentes, Statistics Department, NCSU
Jerry M. Davis, Marine, Earth & Atmospheric Sciences, NCSU

Wind fields along a coastline are composed of many features that are spatially and temporally complex in nature, and they are well recognized as nonstationary spatial-temporal process. We represent the nonstationary spatial-temporal process as a mixure of orthogonal local separable stationary spatial-temporal processes, and a test for separability is proposed. Our main objective is to develop a statistical model which is capable of providing reliable forecasts of wind fields. The model is designed in such a way that it has the capability of combining observed wind data and output from the numerical meteorological models to improve the forecasts. We model the data in terms of an underlying but unobserved true wind process, which is nonstationary in space-time. We estiamte the model in a Bayesian way. This provides improved wind field via the posterior distribution of the true wind, and allows us to validate the numerical model via the posterior predictive distribution of the observations. It also enables us to remove the bias by estimating additive and multiplicative bias parameters in the model. We applied our methods to wind data on the Chesapeake Bay.

e-mail: lchen4@unity.ncsu.edu

FIXED-WINDOW SURVEILLANCE FOR BIOTERRORISM

Sylvan Wallenstein*, Mount Sinai School of Medicine
Joseph I.Naus, Rutgers-The State University of New Jersey

We contrast two methods for detecting clustering of events that could be associated with bioterrorism. We focus on ongoing evaluations of a specific geographic area, with a fixed length window, w, of time (e.g. 7 days), and construct tests based on the maximum of functions involving observed and expected numbers of events in each subinterval (t,t+w). The test based on the Generalized Likelihood Ratio Test (GLRT), requires simulation for computation of p-values, and is guaranteed to have appropriate type-I error. The test based on a generalization of the scan statistic uses an approximation that requires, at each t, calculation of only a single Poisson probability. Simulation and formal arguments, both based on a slightly different model, suggest that the generalization of the scan appears to have appropriate type I error, and that the two tests differ in how they allocate type I error. The scan statistic allocates the error evenly for different time intervals, while the GLRT allocates more of the type I error to time periods in which the event of interest is, under the null, infrequent. Extensions to the case where the width of the window is not fixed will be briefly noted.

e-mail: sylvan@camelot.mssm.edu

46. Applications of Longitudinal Data Analysis

ANALYSIS OF LONGITUDINAL DATA WITH MISSING VALUES OF TWO QUALITATIVELY DIFFERENT TYPES

Ofer Harel*, Dept. of Statistics and The Methodology Center, The Pennsylvania State University
Scott M.Hofer, Dept. of Human Development and Family Studies, The Pennsylvania State University
Joseph L. Schafer, Dept.of Statistics and The Methodology Center, The Pennsylvania State University

In many longitudinal studies, missing values are of two fundamentally different types. Consider the following examples: (a) In a clinical trial comparing various drug therapies for schizophrenic patients, some subjects drop out over time for reasons that are highly related to the severity of illness and degree of improvement, whereas others occasionally miss measurements for unrelated reasons. That is, the dropout is thought to arise from a 'nonignorable' missing-data mechanism, whereas the intermittent missed measurements are essentially 'ignorable.' (b) Researchers conduct a long-term study of cognitive functioning among the elderly. Over time, some subjects die but others drop out for other reasons. Death causes a subject to leave the population of interest, whereas other types of attrition do not. In this talk, we extend Rubin's (1976, 'Inference and missing data,' Biometrika, 63, 581-592) theory of missing data in likelihood-based or Bayesian procedures to situations where the missing values are of two qualitatively different types. The random variables that indicate whether the individual observations are missing or not—which took two possible values (0 or 1) in Rubin's framework—now take three possible values (0, 1, or 2). The joint distribution of these indicators can be factorized in various ways with different implications for inference. As a result, we can now say precisely what it means for one type of missing value to be 'ignorable' and another to be 'nonignorable.' We demonstrate how this new theory, when combined with a two-stage extension of Rubin's multiple imputation (1987, {\it Multiple Imputation for Nonresponse in Surveys}, New York: Wiley), allow us to analyze data from the two examples described above.

e-mail: oxh102@psu.edu

SEMIPARAMETRIC ANALYSIS OF MEAN RESPONSE MODEL ACCOUNTING FOR LATENT LAG AND SATURATION TIMES

Jingrong Yang*, University of California at Berkeley
Ying Qing Chen, University of California at Berkeley

Existing models for repeated measurements at subject specific and irregular time points assumes that the differentiating agent is fully effective upon initial exposure and remains fully effective for the duration of the exposure. In reality, these assumptions are highly unlikely. Usually, there is an effect lag time and a finite saturation period. In this talk, we will present a form of the multiplicative mean response model that relates the covariates to the mean of the repeated response measurements with latent lag and saturation periods. The stochastic structure of the responses are left unspecified. The resulting class of estimating functions are such that any solutions will be strongly consistent, unique and has well developed large sample approximations. We will also demonstrate the model using heart failure data.

e-mail: jingrong@stat.berkeley.edu

LIFETIME EVENT TREES : DESCRIPTIVE STATISTICS AND INFERENCE

Oscar Loureiro*, Dept of Biostatistics and Epidemiology, University of Massachusetts, Amherst
Monika Raju, Dept of Biostatistics and Epidemiology, University of Massachusetts, Amherst
Jean J. Schensul, The Institute of Community Research
Edward J. Stanek, Dept of Biostatistics and Epidemiology, University of Massachusetts, Amherst

This paper presents a new approach to summarizing sequences of distinct lifetime events when there is no predetermined ordering of events and the subjects may differ in the number of events they go through. Our approach leads to the definition of a lifetime event tree (LET) computed using an algorithm similar to those used in Genomics to build phylogenetic trees. We define bootstrap procedures for statistical inference on LETs and test their performance through simulation. Finally, we apply this new approach to a recent dataset on patterns of experimentation with new drugs among young drug users in Hartford, Connecticut.

e-mail: oscar@math.umass.edu

ESTIMATING SUBJECT-SPECIFIC VARIANCE COMPONENTS WITH MULTIVARIATE REPEATED DATA

Wei-Ting Hwang*, Department of Biostatistics and Epidemiology
University of Pennsylvania School of Medicine
Kathleen J.Propert, Department of Biostatistics and Epidemiology
University of Pennsylvania School of Medicine

Subject-specific variance components can be considered if the assumptions of common variance across individuals are violated or components of the variance-covariance matrix itself are of interest. For example, one may study if the variability of an outcome over time depends on observed individual characteristics or, when evaluating two (or more) outcomes, how the correlation varies across individuals due to observed or unobserved factors. The heterogeneity of the variance components can be modeled through a hierarchical model where subject-specific parameters are considered as random samples from a distribution. Several methods for estimating subject-specific variance components will be discussed and compared through simulation studies, such as maximum likelihood (Lin, Raz, and Harlow, 1997) and empirical Bayes estimators. Extension to multivariate longitudinal data will also be explored. The discussed approaches will be applied to the Interstitial Cystitis Data Base (ICDB) longitudinal cohort study where symptoms such as pain and urinary frequency were measured over time.

e-mail: whwang@cceb.upenn.edu

A DRIVER-RESPONSE MODEL OF PULSATILE HORMONE DATA WITH A TEST OF THE PULSATILE ASSOCIATION

Nichole E. Carlson*, University of Michigan, Department of Biostatistics
Timothy D.Johnson, University of Michigan, Department of Biostatistics
Morton B. Brown, University of Michigan, Department of Biostatistics

The associations between hormones regulate many biological systems. For instance, when hormones are secreted in pulses, a pulse of one hormone may be associated with a pulse of another hormone. We term this a driver-response association. To study pulsatile relationships, blood samples are collected for an extended period of time and assayed for various concentrations. The analysis goals are to estimate the temporal relationship between the pulses and to test whether the pulse association is significant. To accomplish this, we develop a model of pulsatile hormone data which incorporates a driver-response association by modeling the response pulse mass and location as functions of the driver pulse mass and locations. We show that when the times between the driver and response pulses are modeled as beta random variates scaled between the two surrounding driver pulses, one method of testing the association is to test whether the parameters defining the beta are equal to one. The model is written using a Bayesian framework and MCMC is used for estimation. We assess the performance of the test using simulated data and summarize the results using false positive and negative rates.

e-mail: ntcarlso@umich.edu

DIGITAL WOUND IMAGING ENABLES TIMELY IDENTIFICATION OF CHRONIC NON-HEALING WOUNDS MINIMALLY RESPONSIVE TO HYPERBARIC OXYGEN TREATMENT

Anuradha Roy*, The University of Texas at San Antonio, Dept. of Management Science and Statistics
John Kalns, United States Air Force School of Aerospace Medicine,Davis Hyperbaric Laboratory
CleAnn Loeffler, Texas Luthern University
James K. Wright, United States Air Force School of Aerospace Medicine,Davis Hyperbaric Laboratory

A common problem in medical science is the identification of two groups e.g. fast and slow healing patients on the basis of the diagnostic information measured repeatedly over time. The objective of this article is to determine if wound area measurements over time can be used to identify patients that are minimally responsive to hyperbaric oxygen treatment (HBOT) and search for prognostic factors that impact HBOT. Fast and slow healing groups of patients were first identified by visual examination of the clusters of the profile plots of percent normalized wound healing areas for the first five weeks of HBOT and then validated by statistical test. It is found that age, blood glucose and creatinine affect HBOT response. Digital images obtained during the first three weeks of treatment predict if a patient is minimally responsive to HBOT, with 100 % accuracy. Thus, wound area measurements can be used to identify patients minimally responsive to HBOT. Low-cost, rapid assessment using digital imaging, in conjunction with available laboratory and demographic data may be useful in making timely changes in wound treatment resulting in improved outcomes and reduced wound care cost.

e-mail: aroy@utsa.edu

47. Network Design

NETWORK DESIGN FOR SEMIVARIOGRAM ESTIMATION AND KRIGING

Dale L. Zimmerman*, Department of Statistics and Actuarial Science
University of Iowa

Inference for spatial data is affected substantially by the spatial configuration of the network of sites where measurements are taken. In this talk, an approach to network design that emphasizes the utility of the network for estimating the spatial dependence (as characterized by the semivariogram) is contrasted with an approach that emphasizes prediction (kriging) of unobserved responses assuming known spatial dependence. It is shown, via contrived examples, that these design objectives are largely antithetical and thus lead to quite different ‘optimal’ designs. Furthermore, a design approach that emphasizes prediction but takes the sampling variation of spatial dependence parameters into account is described and illustrated.

e-mail: dzimmer@stat.uiowa.edu

DESIGN OF LARGE-SCALE AIR MONITORING NETWORKS

David M. Holland*, U. S. Environmental Protection Agency, Office of Research and Development
Arin Chaudhuri, North Carolina State University, Department of Statistics
Montserrat Fuentes, North Carolina State University, Department of Statistics

The potential effects of air pollution on human health have received much attention in recent years. In the U.S. and other countries, there are extensive large-scale monitoring networks designed to collect data to inform the public of exposure risks from air pollution. A major criterion for modifying an existing network is the suitability of spatial predictions at non-monitored locations. These spatial predictions can be used to develop better pollution control strategies for protecting human health. To accomplish this, it is important to ask what monitoring coverage is required to allow optimal, in some quantitative sense, spatial predictions. We consider new approaches for network design based on entropy criteria and modeling the underlying nonstationary covariance structure of atmospheric pollutant processes. In general, entropy is defined as maximizing information expected about potential non-monitored locations. Sites with observations near air quality standards are given higher priority in a combined entropy-air standard design criterion. Eight-hour daily maximum ozone concentrations observed at 513 National Air Monitoring sites are used to demonstrate several network designs.

e-mail: holland.david@epa.gov

DESIGNS FOR PREDICTING THE EXTREMES OF SPATIAL PROCESSES

James V. Zidek*, U British Columbia
Nhu D.Le, BC Cancer Agency
Li Sun, Edmunds.com, Inc

This paper proposes methods for adding sites to environmental monitoring networks: (1) entropy maximization to side-step the specification of specific design objectives; (2) locating the monitoring stations so that the probability of violating a regulatory standard at unmonitored sites is minimized. We demonstrate the use of approach (1) by extending a monitoring network in Vancouver, that measures hourly concentrations of small airborne particulates. In particular, we find a spatial predictive distribution of the concentrations for potential sites, given the data from the existing network. Finding that distribution proves a challenging problem since amongst other things the data have a 'staircase' pattern. We will describe how recent improvements to a Bayesian hierarchical model can be used to meet the challenges. We will also see limitations to approach (1) when different metrics especially extremes such as daily 1-hour maxima, are of concern. That leads us to turn to approach (2) which is easily implemented using the predictive distribution from approach (1). However, that approach leads in turn to some difficult conceptual policy makers must face.

e-mail: jim@stat.ubc.ca

48. Statistical Issues in Serial Analysis of Gene Expression (SAGE)

SAGE - A STATISTICIAN'S VIEW OF THE UNDERLYING BIOLOGY

Keith A. Baggerly*, MD Anderson Cancer Center

SAGE, the Serial Analysis of Gene Expression, is a technique for measuring the RNA expression profile of a sample of cells. In this respect, it provides information similar to that provided by a cDNA microarray experiment. Unlike microarrays, however, SAGE data is derived from sequencing rather than hybridization, and the raw data arrive in the form of counts rather than relative expression values. In this talk, we attempt to clarify the relative strengths and weaknesses of SAGE through a basic description of the underlying biology, highlighting sources of variation as they occur.

This introductory talk will supply background and context for the other talks in this session.

e-mail: kabagg@mdanderson.org

NOISE AND SHADOW: STATISTICAL METHODS FOR SAGE DATA

Natalie J. Blades*, Department of Biostatistics, Johns Hopkins School of Hygiene and Public Health
Giovanni Parmigiani, Departments of Oncology and Biostatistics, Johns Hopkins University

Serial analysis of gene expression (SAGE) is a technique for obtaining information about gene expression. Data from a SAGE experiment consist of long lists of gene identifiers (tags) and corresponding frequencies. Many tags appear only a few times. Some low frequency tags represent low frequency mRNAs, but some result from sequencing errors. It is difficult to distinguish between these two cases. We present methods for enhancing signal from rare tags.

We present a method for calculation of the error rate in any library. We observe a linear relationship between the copy number for a given tag and the number of tags that differfrom the tag of interest by a single-base substitution, insertion, or deletion. The slope of this relationship may be transformed to estimate the error rate.

We also present a model for reassigning erroneously read tags by identifying probable errors and the corresponding tags that spawn them. An error in one base pair of a very common tag may result in the observation of a completely new, but similar, tag—a shadow of the common tag. The proposed method reassigns erroneously observed shadows to the tags that may have generated the shadow.

e-mail: nblades@jhsph.edu

DIFFERENTIAL GENE EXPRESSION STUDIES USING SAGE

Eleanor Feingold*, Department of Human Genetics, University of Pittsburgh
Lisa Weissfeld, Department of Biostatistics, University of Pittsburgh
Yan Lin, Department of Biostatistics, University of Pittsburgh

Serial Analysis of Gene Expression (SAGE) is a powerful molecular technique that produces an exact count of the number of each species of mRNA seen in a biological sample. SAGE has some advantages over microarray technologies, but it is much more labor-intensive. It is generally done on a small number of samples, often as a 'first-pass' to be followed by microarray studies. Statistical methods for SAGE data can be similar to those for microarray data, except that they must take into account the small sample size and the unique multinomial sampling structure produced by counting mRNAs in a subsample. We present a Poisson-gamma model for SAGE data that captures both the multinomial sampling structure andsample to sample variability. We show how to use our model to produce a ranked list of genes that are most differentially expressed between two treatments. We emphasize the situation where there are very few samples (perhaps only one) from each treatment.

e-mail: feingold@pitt.edu

BAYESIAN SHRINKAGE ESTIMATION OF THE RELATIVE ABUNDANCE

Jeffrey S. Morris*, University of Texas MD Anderson Cancer Center, Department of Biostatistics
Keith A.Baggerly, University of Texas MD Anderson Cancer Center, Department of Biostatistics
Kevin R. Coombes, University of Texas MD Anderson Cancer Center, Department of Biostatistics

SAGE is a technology for quantifying gene expression in biological tissue yielding multinomial count data with two characteristics: skewness in the relative frequencies and small sample size relative to the dimension. These factors cause a given SAGE sample to miss a large number of expressed mRNA species present in the tissue. Empirical estimators of relative abundance effectively ignore these missing species, and as a result tend to overestimate the abundance of the scarce observed species comprising a vast majority of the total. We have developed a new Bayesian estimation procedure that quantifies our prior information about these characteristics, yielding a nonlinear shrinkage estimator with efficiency advantages over the MLE. Our prior is a mixture of Dirichlets, whereby species are stochastically partitioned into abundant and scarce classes, each with its own multivariate prior. Simulations suggest our estimator has lower IMSE than the MLE and expression profiles uniformly closer in Euclidean distance to the truth. We apply our method to a SAGE library of normal colon tissue, and discuss its implications for assessing differential expression.

e-mail: jeffmo@mdanderson.org

49. Functional Analysis of Longitudinal Data

BORROWING STRENGTH IN THE ANALYSIS OF LONGITUDINAL AND FUNCTIONAL DATA

John Rice*, Department of Statisitcs, Univ. of California at Berkeley

Until recently, functional data analysis (FDA) and longitudinal data analysis (LDA) have been rather disjoint enterprises. In this talk, I will attempt to compare and contrast their perspectives and methods. Themes I will pursue include considering how strength is borrowed in FDA and LDA, how they can borrow strength from each other, the relationships between stochastic models and smoothing algorithms, and directions in which current methodology can advance beyond currently used linear methods. The talk is thus a review from my personal perspective with a view toward future directions.

e-mail:

MARGINAL NON- AND SEMI-PARAMETRIC KERNEL REGRESSION FOR LONGITUDINAL DATA

Naisyin Wang*, Department of Statistics, Texas A&M University,
Raymond J.Carroll, Department of Statistics, Texas A&M University,
Xihong Lin, Department of Biostatistics, University of Michigan,

There has been a substantial recent interest in Investigating the performance of non- and semiparametric marginal estimation using kernel methods. Most approaches adopt the strategy of ignoring the within-cluster correlation structure either in nonparametric curve estimation or throughout. When the cluster size m remains fixed, a result supporting the use of this "working independence" strategy indicates that under the conventional estimation procedure, a correct specification of the correlation structure actually diminishes the asymptotic efficiency. In this presentation, I will discuss an alternative kernel estimating equation that accounts for the within subject correlation. For nonparametric curve estimation, the variance of the proposed method is uniformly smaller than that of the most efficient working independence approach. Under the framework of marginal generalized partially linear models, the new estimator is semiparametric efficient in the Gaussian case, and is more efficient than the working independence estimator in non—Gaussian cases.

e-mail: nwang@stat.tamu.edu

MAMMOGRAPHIC SCREENING FOR BREAST CANCER: STATISTICS AND POLITICS

Donald A. Berry*, MD Anderson Cancer Center

The recent tumult regarding the potential benefits and risks of screening mammography pits two sets of people against each other. Both sides claim to have the best interests of women at heart, but women are caught in the middle. One side focuses on the uncertainty regarding the existence of benefit based on the evidence, the quality of the randomized trials, and the comparison of risks and benefits. This side wants to inform women of the benefits and risks and the associated uncertainties, and to help them make a personal decision regarding screening. The other minimizes the risks and claims that even a small reduction in breast cancer mortality is worth regularly screening all women over the age of 35 or 40. Virtually every medical association has recommended regular screening. They point to the approximately 15% reduction in breast cancer mortality over the last 10 years in the U.S. as evidence that screening is effective. I will say why and how this latest flare-up came about, and I will evaluate the evidence presented by both sides. I will also recount my personal involvement with the press, with politicians and with the medical establishment.

e-mail: dberry@odin.mdacc.tmc.edu

HOPE AND CONTROVERSY ABOUT LUNG CANCER SCREENING WITH HELICAL CT

Constantine Gatsonis*, Center for Statistical Sciences, Brown University

Screening for lung cancer has up to now not been shown to have a clear benefit. Helical CT, a new, fast CT capable of providing a clearer view of the lungs, was hailed by the media and by some in scientific community as the breakthrough imaging modality that would significantly reduce lung cancer mortality. The scientific data fueling these claims were rather modest. However, the sequel was all too familiar: daring projections about the effectiveness of screening with helical CT, extensive media coverage, strong statements from advocacy groups, and accusations insensitivity to the plight of lung cancer patients made against those who called for a more rigorous scientific evaluation of helical CT screening. Eventually (and predictably) media attention subsided; the scientific debate adopted a more nuanced tone; and a major randomized study was launched in the US to evaluate helical CT. This presentation will examine the role of the major parties in the controversy and will highlight some important methodological issues that were raised. These issues include the role of randomized comparative studies of screening interventions, the feasibility and limitations of such studies, and, more broadly, the need to reexamine the dominant paradigm of how screening should be evaluated.

e-mail: gatsonis@stat.brown.edu

PROSTATE CANCER SCREENING: RECONCILING GOOD INTENTIONS WITH GOOD SCIENCE

Barnett S. Kramer*, National Institutes of Health, Department of Health and Human Services

In the late 1980s, a well-publicized paper published in the New England Journal of Medicine reported that men with early stage prostate cancer often had elevated blood levels of prostate-specific antigen (PSA). This led to a rapid surge in enthusiasm for prostate cancer screening in the hopes that the disease could be detected at a more curable stage, and a 'pseudoepidemic' of the disease followed from widespread opportunistic screening. Some advocacy groups, hoping that a new means to decrease the mortality and suffering from the number two cancer killer in American men was at hand, embraced the screening philosophy, developed screening recommendations, and publicized the assumed benefits as one of their primary messages. Absent definitive evidence of prostate cancer mortality reduction from PSA screening or evidence that benefits outweighed harms, the National Cancer Institute began planning for a randomized controlled trial. However, in the face of growing national enthusiasm for PSA screening, the planning process encountered a variety of difficulties. The discussion will cover how each of these issues was handled.

* Opinions expressed in this talk are those of the speaker, and do not represent an official position or opinion of the Department of Health and Human Services or of the National Institutes of Health.

e-mail: bk76p@nih.gov

51. Analysis of Adverse Event Reports

POSTMARKETING DRUG ADVERSE EVENT SURVEILLANCE AND THE INNOCENT BYSTANDER EFFECT

William DuMouchel*, AT&T Shannon Laboratory

The Multi-item Gamma Poisson Shrinker (MGPS) is an empirical Bayesian method for identifying unusually frequent counts in a large sparse frequency table. This presentation focuses on estimating associations among drugs and adverse event codes in databases of postmarketing reports of adverse drug reactions, as practiced by FDA and other safety researchers. Extended methods can be used to signal frequent itemsets with more than two items, such as combinations of two drugs and one AE, or syndromes of multiple AEs. Another extension allows us to focus on detecting differences between itemset frequencies in different subsets of the data, or from one time period to another. Recent research attempts to adjust drug-adverse event associations for the effects of concomitant medications—sometimes called the "innocent bystander problem."

e-mail: dumouchel@research.att.com

SPONTANEOUS FIRE REPORTS

Michael Greene*, Consumer Products Safety Commission
Mark Levenson, Consumer Products Safety Commission

Most analyses of nationwide fire losses trends are based on combining data from the U. S. Fire Administration's National Fire Incident Reporting System (NFIRS) with the National Fire Protection Association's (NFPA) annual probability survey of fire departments. NFIRS is a voluntary reporting system with detailed characteristics of several hundred thousand individual fire incidents, that analysts scale to national-level estimates with the NFPA survey. In 2002, the U. S. Consumer Product Safety Commission staff (CPSC) launched two new surveys to supplement the NFIRS/NFPA surveys. One survey begins with acquisition of all fire death certificates (i.e. a census of all fire deaths) and supplements that information with detailed investigation of the incidents similar to NFIRS. The second survey begins with fire injuries sampled in CPSC's NEISS hospital emergency department reporting system. This is also followed with a detailed investigation of the incident. These new surveys, a census of fire deaths and a probability sample of fire injuries are unbiased and can provide valid standard error estimates. Experience with the new systems is described in this paper.

e-mail: mgreene@cpsc.gov

ESTIMATING THE NUMBER OF DISAPPEARED

Jana L. Asher*, AAAS Science and Human Rights Program and Carnegie Mellon University
Patrick Ball, AAAS Science and Human Rights Program

A central problem in the statistical analysis of human rights violations is how to use imperfect records to estimate the number and kinds of persons who suffer violations. This is particularly important when the analysis is used as evidence in a legal proceeding. The records may contain multiple reports of the same incident, sometimes with slightly different details; the records must be matched to prevent double-counting and to control for reporting bias. Similarly, many violations are never reported, and it is important to provide estimates of the true magnitude of violence. All of these issues are illustrated in our analysis of human rights violations in Kosovo. The analysis discussed in this paper was presented at the International Criminal Tribunal for Former Yugoslavia in The Hague as part of the trial of Slobodan Milosovic.

e-mail: asher@stat.cmu.edu

MIXED MODELS FOR ESTIMATING THE NUMBER OF DISTINCT SPECIES IN REPLICATED WILDLIFE SURVEYS

Robert M. Dorazio*, U.S. Geological Survey
Andrew Royle, U.S. Fish and Wildlife Service

Several approaches have been proposed for adjusting (upward) the number of wildlife species actually observed in a survey to account for imperfect detection. In likelihood-based approaches detection probabilities have been specified as functions of unobservable species-level effects and, when available, fixed effects for observable covariates, such as sampling location or habitat. An example is the logistic-normal mixture where logits of species-specific detection probabilities are assumed to be normally distributed. However, estimated levels of variation in detection of species can be quite high (owing to differences in behavior or abundance of species) and may produce considerable uncertainty in the estimated number of species.

One method of reducing this uncertainty is through replication. We consider surveys in which sampling locations are repeatedly visited within a short period of time, and the identities of all observed species are recorded at each time and location. We develop extensions of the logistic-normal mixture so that differences in detection among species, sampling locations, or observers may be modeled. Using analyses of data from the North American Breeding Bird Survey, we illustrate that uncertainty in estimates of species richness can be reduced considerably with only modest amounts of relication.

e-mail: bdorazio@usgs.gov

A MIXED EFFECTS MODELLING APPROACH TO CLUSTERED BINOMIAL DATA WITH RANDOM CLUSTER SIZES

Renjun Ma*, Department of Mathematics and Statistics, University of New Brunswick,

In developmental toxicity studies, data are generally clustered with random cluster/litter sizes. Appropriate inference about the dose effect should take account for both intra-cluster correlation and extra-variation arising from the random cluster sizes. We introduce a Paired Poisson random effects model for performing appropriate analyses of such data. This approach is illustrated with the analysis of developmental toxicity data.

e-mail: renjun@unb.ca

GENERALIZED LINEAR MIXED MODELS: THE IMPACT OF NON-NORMAL RANDOM EFFECTS FOR A POISSON COUNT MODEL

Kerrie P. Nelson*, University of South Carolina
Brian G.Leroux, Department of Biostatistics, University of Washington

The use of generalized linear mixed models (GLMM's) to model correlated data that has a non-normal response variable, has gained popularity over the last decade. One assumption underlying these models is that the random effects are normally distributed. Due to the intractability of the integrals involved in the likelihood function, it is difficult to examine violations of this assumption theoretically, and computationally intensive through the use of simulation studies. Consequently little has been done to date to examine the effects on parameter estimation (both the regression coffecients and variance components) when this assumption is not met. In this talk I will show the effects of violating this assumption using three commonly used methods for fitting GLMMS, including PQL, iterative bias correction and maximum likelihood on parameter estimation for a Poisson count data model where the random effects have an autoregressive structure.

e-mail: kerrie@stat.sc.edu

A GENERAL APPROACH FOR TWO-STAGE ANALYSIS OF MULTI-LEVEL CLUSTERED NON-GAUSSIAN DATA

Inna Chervoneva*, Biostatistics Section, Thomas Jefferson University
Boris Iglewicz, Department of Statistics, Temple University,

The focus is on nonnormal clustered data with multiple levels of clustering. Such data sets are common in biological and medical studies utilizing monitoring or image-processing equipment. Here we extend the global two-stage approach (GTS) to estimate the population parameters. We consider a class of hierarchical models with any $\sqrt {n}$-consistent and asymptotically normal cluster-specific estimates. The second stage model is a standard linear mixed effects model with normal random effects, but the cluster-specific distributions, conditional on random effects, can be non-Gaussian. For estimation of the population parameters, we consider the ML approach, which is an extension of the GTS method, and propose the restricted maximum likelihood (REML) approach for cases with moderate number of highest-level clusters. Both estimation procedures might be accomplished using SAS PROC MIXED. For tests on the fixed effect, we consider the Wald test and also approximate F-tests and compare their performance in a simulation study. The proposed REML method is here applied to nonnormal clustered data using the quartiles as cluster-specific parameters. An actual data set is used to show that the proposed methodology can lead to improved insight and more meaningful conclusions as compared to a straightforward linear mixed effects analysis.

e-mail: I_Chervoneva@lac.jci.tju.edu

ANALYSIS OF RATER AGREEMENT FOR BINARY RESPONSES

Svend Kreiner, Dept. of Biostatistics, University of Copenhagen, Denmark
Jorgen H.Petersen*, Dept. of Biostatistics, University of Copenhagen, Denmark
Klaus Larsen, Hvidovre University Hospital, Denmark

An item response model for the assessment of multiple rater agreement for categorical responses is discussed. Differences between raters are modelled using random effects which allows a straightforward description of the inter-rater variability. Further, a measure of the degree of rater agreement is presented. This measure has an attractive interpretation and is independent of the sampling distribution of the assessed individuals. The model's applicability is illustrated with data on the potential benefit with respect to perinatal mortality of Doppler ultra sound velocimetry. In the study, 32 international experts were asked to evaluate 139 ultrasound pictures, and give their expert opinion on whether perinatal mortality was potentially avoidable if the child was electively delivered subsequent to abnormal ultra sound results.

e-mail: J.H.Petersen@biostat.ku.dk

MULTIVARIATE LOGISTIC ANALYSES USING SCALE MIXTURES OF NORMALS

Sean M. O'Brien*, National Institute of Environmental Health Sciences
David B.Dunson, National Institute of Environmental Health Sciences

In analyzing multivariate binary or ordinal outcomes, probit models are widely used, since the underlying normal linear structure results in simplified computation and modeling of dependency. Motivated by difficulties in parameter interpretation, this article proposes to link together logistic regression models for the individual outcomes via an underlying normal structure. Each logistic regression is approximated by a scale mixture of underlying normal linear models. Following a Bayesian approach, this approximation is then used to facilitate posterior computation for the exact model via an efficient Markov Chain Monte Carlo algorithm. The methodology can be used for a variety of data structures, including multidimensional longitudinal data and mixtures of categorical and continuous outcomes. The approach is illustrated using data from a study of neurological development.

e-mail: obrien4@niehs.nih.gov

NONPARAMETRIC ASSESSMENT OF SPATIAL ISOTROPY FOR POINT PROCESSES AND MARKED-POINT PROCESSES

Yongtao Guan*, Texas A&M University
Michael Sherman, Texas A&M University
James A. Calvin, Texas A&M University

A common requirement for spatial modelling is the development of the correlation structure. While isotropy is often used for this structure, it is not always appropriate. A conventional practice checking for isotropy is to informally assess plots of direction-specific sample (semi)variograms. These graphical diagnostics are difficult to assess and open to interpretation.

We propose a formal approach testing for isotropy which is both objective and appropriate for a wider class of models. We here pay specific attention to irregularly-spaced data where the locations of observations can be modeled by a homogeneous, isotropic point process. Our testing approach is purely nonparametric in that no explicit knowledge of the marginal distribution of the process is needed. We also demonstrate the use of our approach to test if a homogeneous point process itself is isotropic.

e-mail: guanyt@stat.tamu.edu

A CONVOLUTION KERNEL APPROACH TO MODELLING DIFFUSIVE PROPORGATION IN SPATIO-TEMPORAL PROCESSES

Bill Xu*, Department of Statistics, University of Missouri - Columbia
Christopher K.Wikle, Department of Statistics, University of Missouri - Columbia

Some high-dimensional spatio-temporal processes exhibit diffusion and propagation simultaneously. An obvious example is a pollutant drifting and diffusing in the presence of dominant advecting wind. We propose a convolution kernel approach to model this behavior. The diffusive propagation is modelled through the displacement of the convolution kernel. In addition, non-separable behavior is modelled by allowing the kernel displacements to vary with space. Computational efficiency is achieved by using the FFT and spectral dimension reduction. Estimation is carried out via MCMC. The method is applied to a thunderstorm nowcasting problem in Sydney, Australia.

e-mail: xu@stat.missouri.edu

ON A TIME DEFORMATION REDUCING NONSTATIONARY STOCHASTIC PROCESSES TO LOCAL STATIONARITY

Marc G. Genton*, North Carolina State University
Olivier Perrin, University Toulouse 1

A stochastic process is locally stationary if its covariance function can be expressed as the product of a positive function multiplied by a stationary covariance. In this paper, we characterize nonstationary stochastic processes that can be reduced to local stationarity via a bijective deformation of the time index, and we give the form of this deformation under smoothness assumptions. This is an extension of the notion of stationary reducibility. We present several examples of nonstationary covariances that are locally stationary reducible. We also investigate the particular situation of exponentially convex reducibility which can always be achieved for a certain class of separable nonstationary covariances.

e-mail: genton@stat.ncsu.edu

BAYESIAN INFERENCE FOR PAIRWISE INTERACTING POINT PROCESSES

Matthew A. Bognar*, University of Iowa
Mary K.Cowles, University of Iowa

Pairwise interacting point processes are commonly used to model spatial point patterns. To perform inference, the established frequentist methods can produce good point estimates when the interaction in the data is moderate, but some methods may produce severely biased estimateswhen there is strong interaction present in the data. Furthermore, because the sampling distributions of the estimates are unclear, interval estimates are typically obtained by parametric bootstrap methods. In the current setting however, the behavior of such estimates is not well understood. In this article we propose Bayesian methods for obtaining inferences in a pairwise interacting point process. The requisite application of Markov chain Monte Carlo (MCMC) techniques is complicated by an intractable function of the parameters in the likelihood. The acceptance probability in a Metropolis-Hastings algorithm involves the ratio of two likelihoods evaluated at differing parameter values. The intractable functions do not cancel, and hence an intractable ratio must be estimated within each iteration of a Metropolis-Hastings sampler. We propose the use of importance sampling techniques within MCMC to address this problem.

e-mail: mbognar@stat.uiowa.edu

COVARIATE-ADJUSTED AND BIVARIATE SPATIAL CDF MODELING

Margaret B. Short*, University of Minnesota
Bradley P.Carlin, University of Minnesota
Alan E. Gelfand, Duke University

A spatial cumulative distribution function (SCDF) gives the proportion of a spatial domain D having the value of some response variable less than a particular level w. We provide a fully hierarchical approach to SCDF modeling, using a Bayesian framework implemented via Markov chain Monte Carlo (MCMC) methods. The approach generalizes the SCDF to accommodate block-level variables, possibly utilizing a spatial change of support model within an MCMC algorithm. We then extend our approach to allow covariate weighting of the SCDF estimate. Such an extension is natural in assessments of environmental justice, where we wish to determine if a particular sociodemographic group is being excessively exposed to harmful levels of certain pollutants. We further generalize the framework to the bivariate random process setting, which allows simultaneous modeling of both responses and weights. Again, MCMC methods enable straightforward estimates of weighted and bivariate SCDFs. We illustrate our methods with two air pollution data sets, one concerning ozone exposure and race in Atlanta, GA; the other recording NO and NO2 ambient levels at 67 California monitoring sites.

e-mail: margares@biostat.umn.edu

BAYESIAN FACTOR ANALYSIS FOR SPATIALLY CORRELATED DATA, WITH APPLICATION TO SUMMARIZING AREA-LEVEL MATERIAL DEPRIVATION

Joseph W. Hogan, Brown University
Rusty Tchernis*, Harvard University

This paper addresses the problem of quantifying area-level material deprivation using data from the US Census. A factor-analytic approach is used, framed in a Bayesian context and elaborated to allow spatial correlation. The latent factor is interpreted as an index of material deprivation.

An existing and widely-used measure of the same construct is the Townsend index, an unweighted sum of four census variables standardized as Z-scores. The Townsend and many related indices are computed as linear combinations of measured census variables, which motivates the factor-analytic structure adopted here. Our model-based approach allows several improvements over the Townsend and similarly constructed indices: (1) the index can be represented as a weighted sum of (standardized) census variables, with data-driven weights; (2) by using posterior summaries, the indices can be reported with corresponding measures of uncertainty; and (3) information from neighboring areas can be used to inform a specific area's index, which improves precision and can be useful for sparsely-populated areas.

e-mail: tchernis@hcp.med.harvard.edu

PREDICTION PERFORMANCE OF CORRELATED ERROR PROCEDURES FOR SPATIAL COUNTS

Robert Downer*, Dept. Experimental Statistics, LSU & LSU AgCenter

Spatial counts occur in a variety of disciplines including agronomy, entomology and ecology. Few guidelines exist for the application of geostatistical methods to spatial counts and the prediction to unsampled areas is an important aspect of experimental field research. The prediction performances of kriging and a correlated errors Poisson model are compared through simulation. Counts with a known spatial covariance structure are generated in an investigation involving several factors: area size, overall mean, range of correlation, spatial covariance function, and the presence of trend. The correlated errors Poisson model generally gives superior prediction performance when an exponential covariance structure is used.

e-mail: rdowner@lsu.edu

54. Latent Variable Modeling

LATENT PREDICTORS OF LATENT CLASS OUTCOMES

Melanie M. Wall*, University of Minnesota

Eating disorders are often assessed through a battery of self-report questionnaire items asking whether certain behaviors have been used within the last year. Like other questionnaires of this type for assessing health conditions, a common technique is to create a score based on the observations and choose a cut-off value to indicate individuals as high or low. An alternative technique is to use latent variable modeling which instead treats the observed variables as indicators of the true variable of interest, in this case an indicator of whether the individual has an eating disorder or not. When the latent variable is categorical, latent class modeling can be used and is well developed as a technique for classifying individuals. Besides classifying individuals, it is often also of interest to examine the relationship between possible latent predictors and the latent class outcome variable. While this model has been considered and is implemented in software for non-latent predictor variables, it has not been well developed when the predictor variables are latent as well. This presentation considers a model where a latent variable measuring body satisfaction is predicting the latent class for eating disorders. The estimation method is presented and a comparison is made with the results obtained if simple sum scores are used instead.

e-mail: melanie@biostat.umn.edu

LATENT VARIABLE MODELING FOR MISCLASSIFIED POLYTOMOUS OUTCOME VARIABLES

Jens C. Eickhoff*, Department of Biostatistics & Medical Inform, University of Wisconsin-Madison
Yasuo Amemiya, Department of Statistics, Iowa State University

Latent variable modeling is a multivariate statistical technique commonly used in behavioral and observational epidemiological studies. The models used in such analysis relate all observed variables to latent common factors. The analysis is traditionally carried out under the assumption that the observed variables are continuous with a multivariate normal distribution. In many observational studies, the response variables are in polytomous form which are often affected by misclassification errors. In this paper, we propose a new latent variable modeling approach which takes into account the response error associated with the measured polytomous outcome variables. A computationally efficient Monte Carlo EM algorithm is implemented to compute the maximum likelihood estimates. To validate the benefits of our approach, a simulation study is conducted. The approach is illustrated using a substance abuse prevention study.

e-mail: eickhoff@biostat.wisc.edu

MISSPECIFICATION ERROR IN LATENT VARIABLE MODELS

Yun Ju Sung*, School of Statistics, University of Minnesota

When a model is incorrect, the MLE is inconsistent, converging to the minimizer θ^* of Kullback-Leibler information relative to the true probability distribution of the data.We propose a Monte Carlo method for finding θ^*, thus characterizing the bias due to model misspecification, even when there are missing data or latent variables and the observed data likelihood doesn't have closed form.

We prove consistency and asymptotic normality of the Monte Carlo estimate of θ^*. It involves generating two samples, the first for latent variables (or missing data) from an importance sampling density and the second for observed variables from the true density.The entire first sample is used with each member of the second sample. Thus the first sample is used a number of times equal to the size of the second sample. We show that this results in an asymptotic variance for the estimate smaller than that obtained by using the first sample only once. The proof requires the theory of empirical processes, instead of the ordinary central limit theorem.

e-mail: yjsung@stat.umn.edu

LONGITUDINAL MEASUREMENT ERROR MODEL FOR SLEEP DISORDERED BREATHING

Liang Li*, Department of Statistics, University of Wisconsin-Madison
Mari Palta, Department of Population Health Sciences, University of Wisconsin-Madison
Jun Shao, Department of Statistics,University of Wisconsin-Madison

In the study of measurement error models, it is typically assumed that the measurement error follows an additive or multiplicative model. However, such models do not hold for the measurement error of sleep disordered breathing (SDB). The true covariate and its surrogate are both non-negative with a point mass at zero. We propose a latent variable model to characterize the error structure in this situation and use regression calibration to estimate the parameters. Unlike the usual cases in functional methods, a distributional assumption on the true covariate has to be made in this model. We use mixture of normal distribution to attain flexibility against model misspecification. The model is extended to incorporate longitudinal measurements of SDB and the response.

Key Words: measurement error, longitudinal, mixture of normal

e-mail: liangli@stat.wisc.edu

THE QUASI ML ESTIMATION METHOD FOR MODELING NONLINEAR EFFECTS IN A GENERAL LATENT VARIABLE MODELING FRAMEWORK

Andreas G. Klein*, Univ. of Illinois at Urbana-Champaign
Bengt O.Muthen, Univ. of California Los Angeles

In a statistical research context, a linear model sometimes provides only a questionable representation of reality. This situation particularly arises when the size of an effect of an exogenous variable on an endogenous variable (e.g., the effect of a treatment) itself depends on the outcome of third variables (e.g., individual compliance level). Then, in addition to linear effects, an interaction effect becomes an integral part of the model structure. The modeling of interaction or other nonlinear effects by using structural equation modeling has been an issue of ongoing research in recent years. In practice, theory-based hypothesized interactions have often failed to become replicated. With the Quasi-ML estimation method, a newly developed estimation technique is presented. This method provides an approximate ML estimator and outperforms the currently available methodology (covariance structure analysis, two-stage least squares, methods of moments) considerably with respect to statistical power and model flexibility. In connection with the Quasi-ML estimation method, a new general latent variable modeling framework is presented which incorporates the flexible modeling of multiple nonlinear effects in both cross-sectional and longitudinal models. The applicability of the new modeling framework and Quasi ML estimation method is illustrated by a an empirical data set.

e-mail: agklein@uiuc.edu

APPLICATION OF GROWTH MIXTURE MODELING TO ASSESS THE DEVELOPMENT OF CHILDHOOD OBESITY

Chaoyang Li*, Department of Preventive Medicine, University of Kansas Medical Center

Obesity has become an increasingly important health problem in children. Various developmental trajectories of obesity may be linked to different etiological factors. Growth mixture modeling (Muthen & Shedden, 1999; Muthen & Muthen, 2001) provides a unique framework for examining the distinct growth trajectories for the development of obesity and identifying related correlates. Data were drawn from the National Longitudinal Survey of Youth (NLSY) merged child-mother files. Body mass index (BMI) was modeled in both continuous and categorical scales in growth mixture models. Maternal characteristics and exposure to risk factors during pregnancy, children's weight, length, and gestational age at birth, and demographic characteristics were analyzed as covariates in predicting different growth trajectories of childhood obesity.

e-mail: cli2@kumc.edu

55. Survival Analysis and Clinical Trials

OPTIMAL ESTIMATOR FOR THE SURVIVAL DISTRIBUTION AND RELATED QUANTITIES FOR TREATMENT POLICIES IN TWO-STAGE RANDOMIZATION DESIGNS IN CLINICAL TRIALS

Abdus S. Wahed*, North Carolina State University
Anastasios A.Tsiatis, North Carolina State University

Two-stage designs, where patients are initially randomized to an induction therapy and then depending upon their response and consent, are randomized to a maintenance therapy, are common in cancer and other clinical trials. The goal is to compare different combinations of primary and maintenance therapies to find the combination that is most beneficial. In practice, the analysis is usually conducted in two separate stages which does not directly address the major objective of finding the best combination. Recently Lunceford et. al. (2002, Biometrics, 58, 48-57) introduced ad hoc estimators for the survival distribution and mean restricted survival time under different treatment policies. These estimators are consistent but not efficient. In this paper we derive estimators that are easy to compute and are more efficient than previous estimators. We also show how to improve efficiency further by taking into account additional information from auxiliary variables. Large sample properties of these estimators are derived and comparisons with other estimators are made using simulation. We apply our estimators to a leukemia clinical trial data set that motivated this study.

e-mail: aswahed@unity.ncsu.edu

PARTIAL LEAST SQUARES FOR THE ACCELERATED FAILURE TIME MODEL WITH RIGHT CENSORED DATA

Jie Huang*, Harvard School of Public Health
David Harrington, Harvard School of Public Health

In the accelerated failure time model with high dimensional explanatory variables, the parameter estimates may be unstable, or even not obtainable when the covariates outnumber the events. We propose an iterative partial least squares fitting algorithm based on the Buckley-James estimating equation to predict the covariate effect and the response for a future subject with a given set of covariates. In addition, we propose a leave-two-out cross validation method for selecting the number of variables used in the partial least squares fitting. Simulation studies are conducted to compare our proposed methods with other prediction procedures in different settings. An AIDS data from AIDS Clinical Trials Group (ACTG) protocol 333 is used to illustrate our method.%In AIDS research, it is of great interest to predict the in vivo antiviral change in HIV-1 RNA level (log$_{10}$ copies/mL) from pretreatment levels as a function of the treatment assignment and other clinical measurements. Though it is natural to use the accelerated failure time model in this setting for its ease of interpretation,

e-mail: huangj@hsph.harvard.edu

ESTIMATING SUBJECT-SPECIFIC SURVIVAL FUNCTION UNDER ACCELERATED FAILURE TIME MODEL

Yuhyun Park*, Harvard School of Public Health
L.J. Wei, Harvard School of Public Health

The semi-parametric accelerated failure time model relates the logarithm of the survival time linearly to its covariates without specifying a parametric distribution for the error term. In this article, we are interested in utilizing this model to predict the survival function and its related quantities for future subjects with a given set of covariates. Specifically, we derive the large-sample distribution for the subject-specific cumulative hazard function estimate. We then propose a simple resampling technique to construct point-wise confidence intervals and simultaneous bands for the corresponding survival function and its quantile function over a properly selected time interval. The new proposals are illustrated with two well-known examples in survival analysis.

e-mail: ypark@hsph.harvard.edu

SIMULTANEOUS ESTIMATE OF TREATMENT EFFECTS ON TWO BINARY ENDPOINTS AND THEIR ASSOCIATION IN RANDOMIZED CLINICAL TRIALS

Wenquan Wang*, Department of Biostatistics, University of Iowa College of Public Health
Robert F.Woolson, Department of Biometry and Epidemiology, Medical University of South Carolina
William R. Clarke, Department of Biostatistics,University of Iowa College of Public Health

Sometimes two or more endpoints are needed to assess impact of a new treatment in a randomized clinical trial. In assessing the treatment effects, we are not only interested in the treatment effects on the endpoints, but also in the positive correlation between the endpoints, i.e., we want to see positive effects of the new treatment happening on the same patients. A three-degree-of-freedom test in each of three analytic approaches, weighted least squares, generalized estimating equations and maximum likelihood, is developed and compared through simulation study for two binary endpoints in this study. Some adjustments are made, especially to weighted least squares approach.

e-mail: wenquan-wang@uiowa.edu

AN EVIDENTIAL ALTERNATIVE TO THE LOGRANK TEST

Wesley D. Eddings*, The Johns Hopkins University

The logrank test is the most popular tool for evaluating the association between treatment and survival time. Its sole output is the P value, intended to quantify the strength of the evidence against the null hypothesis of identical time-to-event distributions across treatment groups. The Cox proportional hazards model (the probability model implicit in the logrank test) yields a more appropriate measure of statistical evidence—ratios of the Cox partial likelihood, which compare the relative support for different values of the hazard ratio. Direct use of the partial likelihood avoids the logical difficulties of P values and emphasizes the magnitude of the observed hazard ratio, rather than statistical significance. When more than one covariate is included in the proportional hazards model, a profile likelihood based on the partial likelihood represents the available information about a particular regression coefficient. The generalized profile likelihood retains the properties of the standard parametric profile likelihood, including its superiority to the estimated likelihood obtained by replacing nuisance parameters with their global
maximum likelihood estimates.

e-mail: weddings@jhsph.edu

VALIDATION AND CALIBRATION: AN EMPIRICAL STUDY OF PROGNOSTIC SURVIVAL MODELS

Somesh Chattopadhyay*, Department of Statistics, Florida State University
Daniel L.McGee, Department of Statistics, Florida State University

Whether a prognostic model based on one population is applicable to other populations is of interest to both epidemiologists and clinicians. In this report we present an empirical example of model validation and calibration based on person-level data from 22 studies that include both sexes, black and white groups and samples from the U.S. and other countries. The database analyzed includes almost 250,000 individuals with almost 14,000 deaths from coronary heart disease occurring during follow-up. In our example we develop a model based on one study and examine whether it is applicable to the other twenty-one studies. We demonstrate several aspects of the model validation process including calibrating the model through a linear transformation of the model at the appropriate scale. Among the other possibilities, we examine whether combining models from multiple studies, accompanied by calibration performs better than depending on the model from a single study. We also examine methods of augmenting a model by incorporating new covariates not available to all of the studies. We will examine these possibilities in context of survival models for death from coronary heart disease.

e-mail: somesh@stat.fsu.edu

56. Practical Issues in Nonlinear Mixed Models

CASE STUDIES ILLUSTRATING UNANSWERED QUESTIONS IN INFERENCE WITH NONLINEAR MIXED MODELS

Walter Stroup*, University of Nebraska – Lincoln, Biometry Department

Improvements in computers and statistical software have made nonlinear mixed models widely available. As a result, researchers in many disciplines have “discovered” these models. They are often used in conjunction with designs whose sample sizes are relatively small. This raises questions about the validity of the asymptotic theory upon which inference in widely available software packages is typically based. This talk will present three example applications of nonlinear mixed models that essentially amount to cautionary tales about the current state of the art.

e-mail: wstroup@unl.edu

APPROXIMATION TECHNIQUES IN NONLINEAR MIXED MODELS: STRENGTHS, WEAKNESSES AND IMPLEMENTATION

Edward F. Vonesh*, Applied Statistics Center, Baxter Healthcare Corporation

Generalized linear and nonlinear mixed models are used extensively in such fields as population pharmacokinetics (PK), population pharmacodynamics (PD), bioassay, studies of biological or agricultural growth, and epidemiology. As these models are typically nonlinear in the random effects, maximum likelihood estimation (MLE) requires maximizing an integrated likelihood function (i.e., marginal likelihood) that generally has no closed form expression. To circumvent the computational challenges associated with maximizing an integrated likelihood, a number of estimation techniques have been proposed based on various Taylor series approximations. We will review the strengths and weaknesses of several estimation techniques including those based on the nonlinear mixed effects (NLME) algorithm of Lindstrom and Bates (1990), the penalized quasi-likelihood (PQL) algorithm of Breslow and Clayton (1993), the conditional second-order generalized estimating equations (CGEE2) algorithm of Vonesh et. al. (2002) and a Laplacian-based maximum likelihood (LMLE) algorithm similar to that described by Vonesh (1996) and Wolfinger and Lin (1997). We do so by briefly examining the theoretical basis for and limitations of these approximations, and by investigating their asymptotic properties, both theoretically and via simulation. We will also discuss some common difficulties encountered when implementing these techniques (e.g., starting values for variance-covariance parameters) and methods for potentially overcoming these difficulties. The methods will be illustrated using both simulated and real data with all analyses carried out using SAS-based software.

e-mail: voneshe@baxter.com

INFERENCE FROM COMPLEX EXPERIMENTAL DESIGNS USING NONLINEAR MODELS: A SIMULATION STUDY

Erin E. Blankenship*, University of Nebraska—Lincoln
Walter W.Stroup, University of Nebraska—Lincoln

Many disciplines conduct studies in which the primary objectives depend on inference based on a nonlinear relationship between the treatment and response. However, such data often arise from experiments with split-plots or other features that result in multiple error terms or other non-trivial error structures. The advent of PROC NLMIXED has increased the ease of fitting models resulting from such complex designs, but most of the inference depends on asymptotic procedures whose assumptions may not be satisfied by the relatively small sample sizes typically encountered in designed experiments. In this paper, we present the results of a simulation study to examine the validity of inference for nonlinear models fit to data from complex experimental designs. The results are based on type I error rates, and the simulated experiments were split-plot experiments in which the whole plot experimental units wre arranged in randomized complete blocks.

e-mail: eblankenship2@unl.edu

SMALL SAMPLES AND THE USE OF NONLINEAR MIXED EFFECTS MODELS: A COMPARATIVE SIMULATION STUDY

Mohamed Ould-Moustapha, University of Maryland
Estelle Russek-Cohen, University of Maryland
*,

Since Nonlinear Mixed Effects Models (NLMM) is driven by large sample theory there are several questions that are still open to debate:

(1) The issue of the degrees of freedom or alternatively selection of a correct null hypothesis distribution. (2) Robustness in small samples to assumptions such as normal distribution of subject specific random effects. (3) Use of different Taylor expansions to derive an approximate MLE, and its Estimators. In this paper a comparative simulations study motivated by two widely used datasets to compare four different methods to approximate the NLMM to a 'linear mixed model', discuss the degrees-of-freedom concept, and also point out an issue common to many Nonlinear models, that of convergence especially in the setting of small to medium sample data.

e-mail: er19@umail.umd.edu

57. The Measurement of Health Disparities

STUDYING NEIGHBORHOOD EFFECTS ON HEALTH: AN OVERVIEW OF METHODOLOGICAL CHALLENGES

Stephen W. Raudenbush*, Univeristy of Michigan

Over the past decade, epidemiologists have become increasingly interested in the contributions of neighborhoods to the health of their residents. Neighborhood environments may encourage or discourage physical exercise. They may provide social support or promote fear of crime. They may provide easy access to health care, but they may also expose residents to toxic air. Because neighborhoods are somewhat segregated by social class and ethnicity, neighborhood variation in health-promoting potential may account, in part, for social and ethnic disparities in health outcomes. In this paper, I consider key methodological challenges that arise in studying neighborhood effects on health: 1) causal models for neighborhood effects must relax the "Stable Unit-Treatment Value Assumption" (SUTVA) articulated by Donald B. Rubin because of spatially generated spillover effects; 2) neighborhood-level "treatments" can be hard to specify and are often measured with error; 3) selection bias and endogeneity post special problems that do not arise when treatments are at the individual level; 4) optimal design requires an optimal sample sizes of neighborhoods and of residents within neighborhoods. To contend with these challenges, I propose a hierarchical model with spatially linked causal effects and spatially dependent random effects. Causal variables may be measured with error. Within this framework, I outline propensity-score methods for reducing bias in observational studies.

e-mail:

ECOLOGICAL INFERENCE IN EPIDEMIOLOGY

John Wakefield*, Departments of Statistics and Biostatistics, University of Washington - Seattle.

Ecological regression studies in which an aggregate level response is regressed upon an aggregate level predictor, are a dangerous enterprise since relationships at the level of the group only agree with relationships at the level of the individual under very strict circumstances. This can lead to the possibility of the so-called ecological fallacy which occurs because of within-group variability in exposures and confounders. Usual biases in individual observational studies are far more complicated in an ecological setting, and there are additional biases that are unique to ecological studies. In this talk I will describe the various types of bias and characterize the direction and magnitude in a number of situations. Examples from spatial epidemiology will be presented, in particular the association between cancer incidence
and socio-economic variables.

e-mail: jon@stat.washington.edu

58. GEE and GMM: A Bridge between Biostatistics & Econometrics

A BIOSTATISTICAL PERSPECTIVE ON THE USE OF GENERALIZED ESTIMATING EQUATIONS

John S. Preisser*, Department of Biostatistics, University of North Carolina

Generalized estimating equations (GEE), in its most common usage, is an estimation procedure for parameters in population-averaged regression models where primary interest is in the marginal means. Since its introduction by Liang and Zeger in papers published in Biometrika and Biometrics in 1986, GEE has evolved to address a wide range of problems in both longitudinal and clustered data settings including extensions of the original procedure that fit regression models to within-cluster association. This talk presents a biostatistical perspective on GEE methods that have applications in epidemiology, clinical trials and health behavioral interventions.

e-mail: jpreisse@bios.unc.edu

THE GENERALIZED METHOD OF MOMENTS APPROACH TO ECONOMETRIC ANALYSIS

Alastair R. Hall*, North Carolina State University

Generalized method of moments (GMM) was first introduced into the econometrics literature by Lars Hansen in a paper published in Econometrica in 1982. Since then it has been widely applied to analyze economic and financial data. This interest has both stimulated and been facilitated by the development of numerous statistical inference techniques based on GMM. In this talk, the basic framework of GMM estimation is presented and a number of examples are preseneted to illustrate why the method is so well suited to the problem of performing statistical inference about economic phenomena. An overview of the statistical literature on GMM is provided with particular attention paid to recent work on moment selection.

e-mail: alastair_hall@ncsu.edu

THE GENERALIZED METHOD OF MOMENTS (GMM) AND GENERALIZED ESTIMATING Equations (GEE): Similarities and Differences

L A. Stefanski*, North Carolina State University

Generalized Method of Moments (GMM) and Generalized Estimating Equations (GEE) are theories of estimation based on assumptions about moments of the observed data. Starting from this common ground the similarities and differences of GMM and GEE will be discussed with the objective of identifying features of both methods that have potential for broader application in statisticalinference.

e-mail:

EXACT AND APPROXIMATE METHODS FOR INTERNAL PILOTS: BEYOND TWO GROUP COMPARISONS

Christopher S. Coffey*, University of Alabama at Birmingham
Keith E.Muller, University of North Carolina at Chapel Hill

Uncertainty surrounding the variance or covariance matrix often presents the biggest barrier to accurate power analysis. A poor choice gives either an overpowered or an underpowered study. Internal pilot designs were introduced to resolve such uncertainty about error variance. Most research on internal pilot designs focuses on comparing two groups. Not all designs and even not all clinical trials, involve only two groups. We summarize our recent developments for internal pilot designs which apply to any univariate linear model with fixed predictors and Gaussian errors. We also describe the free program GLUMIP (http://main.uab.edu/show.asp?durki=8533) which makes it easy to apply these methods to perform power calculations and avoid test size inflation. We also describe extensions to the "univariate" approach to repeated measures. We recommend approximating the misspecification of the covariance matrix in terms of the first and second moments of its eigenvalues, for both fixed sample and internal pilot designs. We also describe an unadjusted approach for internal pilots with repeated measures, which leads naturally to suggestions for future research.

e-mail: ccoffey@uab.edu

SAMPLE SIZE REDETERMINATION FOR REPEATED MEASURES STUDIES

David M. Zucker, Dept of Statistics, Hebrew University,
Jon S.Denne*, Eli Lilly and Company, Lilly Corporate Center,

It is common for clinical trials to involve taking repeated measurements of a continuous response over time on each subject, and for such data to be analyzed using a linear mixed model. To calculate precisely the sample size required for such a trial, estimates are needed for both the variances and covariances of the errors. Sample size also depends on the level and pattern of dropout and missing data, phenomena that will almost inevitably be present in any trial. Estimates of all these quantities based on previous studies are often unreliable. Incorrect estimates can lead to a study that is substantially underpowered. An appealing strategy for dealing with this is the two-stage design: (1) examine the data at some interim point, (2) update the estimates of the covariance parameters and the dropout and missing data rates, and (3) suitably modify the sample size or the visit schedule. We propose a broad framework for the two-stage design in the context of the general linear mixed model. We consider designs that allow modifying the sample size, modifying the visit schedule, or modifying both. In addition, we consider blinded estimation of variance at the end of the first stage.

e-mail: denne_jonathan@lilly.com

INTERNAL PILOT DESIGNS FOR CONFIDENCE INTERVALS WITH GOOD PROPERTIES

Michael R. Jiroutek*, Bristol-Myers Squibb Pharmaceutical Research Institute
Keith E.Muller, University of North Carolina at Chapel Hill

A variance value of uncertain accuracy makes choosing a sample size for constructing a confidence interval extremely difficult. To minimize the impact of such inaccuracy, we describe how to update the sample size by using internal pilot designs. A predetermined fraction of observations from the target study provides a new variance estimate, which is then used to conduct an interim sample size adjustment (without an interim data analysis). We extend the methods of Jiroutek, Muller, Kupper and Stewart (2002, in review), who described rules for choosing a sample size which leads to well-behaved confidence intervals for a scalar parameter in a linear model with Gaussian errors. The approach centers on the simultaneous consideration of hypothesis testing and confidence interval properties which leads to the improved alignment of sample size calculations with study objectives. An exact form for the new probability criterion is derived for an internal pilot design. Tables highlight the magnitude of the errors that can occur should the variance value used for planning be misspecified. Strategies for implementing the approach are presented.

e-mail: michael.jiroutek@bms.com

ADAPTIVE DESIGNS IN CLINICAL TRIALS – ARE THEY READY FOR PRIME TIME?

Janet Wittes*, Statistics Collaborative, Inc.

Confirmatory clinical trials employ simple robust statistical methods for design and analysis. The statistical literature offers methods that allow flexibility in the design of clinical studies without sacrificing statistical rigor. The field has already widely adopted sequential analysis, but is currently stepping only gingerly into other flexible designs. Several of these latter designs focus on protection of statistical power by allowing modification of sample size either on the basis of observed nuisance parameters, the so-called internal pilot designs, or on the basis of the observed treatment effect. Some schemes protect the Type I error rate by robbing from Peter (the rejection region for harm) to pay Paul (the region for benefit); others assign different weights to the early and late observations; still others appeal to maintenance of conditional power. Some flexible designs focus on the allocation to active and control groups during the trial. Some researchers have criticized adaptive designs because they fail criteria of statistical optimality. This talk briefly reviews available methods, describes their obvious benefits, and addresses their real and imagined liabilities.

e-mail: janet@statcollab.com

CAUSAL INFERENCE IN CHOICE-BASED INTERVENTION STUDIES

Roderick J. Little*, Department of Biostatistics, University of Michigan
Xihong Lin, Department of Biostatistics, University of Michigan
Qi Long, Department of Biostatistics, University of Michigan

Randomized trials only allow inferences on the selected set of participants who agree to be randomized to the alternative treatments, and attrition may be more likely when individuals are randomized to treatments they view as less desirable. Choice-based designs avoid these problems by allowing participants to choose their treatment, but they are clearly subject to selection bias and confounding. A compromise two-stage design was adopted by the 'Women Take Pride' study, an intervention study for ways of managing heart disease. Participants were randomized to a randomized arm, where treatment was randomly assigned, or to a choice arm, where treatment was chosen by the participant. Causal inferences for this design are studied using ideas of principal stratification' presented in Frangakis and Rubin (2000 Biometrics). Attention is paid to structural assumptions needed to identify parameters, and possible modifications of the design that may reduce these assumptions. Pros and cons of these designs are discussed.

e-mail: rlittle@umich.edu

ON THE INEFFICIENCY OF THE ADAPTIVE DESIGN FOR MONITORING CLINICAL TRIALS

Anastasios A. Tsiatis*, Department of Statistics, North Carolina State University

During the design stage of a clinical trial, detemining the clinically important treatment difference to power a study is often a difficult task. Adaptive designs, which allow the sample size to be modified based on sequentially computed observed treatment differences, have been advocated recently for monitoring clinical trials. Although such methods have a great deal of appeal on the surface, we show that such methods are inefficient and that one can improve uniformly on such adaptive designs using standard group-sequential tests based on the sequentially computed likelihood ratio test statistic.

Several examples of popular adaptive designs are used for illustration.

e-mail: tsiatis@stat.ncsu.edu

EVALUATING THERAPEUTIC STRATEGIES IN MULTI-COURSE CLINICAL TRIALS

Peter F. Thall*, Department of Biostatistics, M.D. Anderson Cancer Center
Hsi-Guang Sung, Department of Statistics, Rice University
Elihu H. Estey, Department of Leukemia, M.D. Anderson Cancer Center

Therapy of rapidly fatal diseases often requires multiple treatment courses. In each course, the treatment may achieve the clinical goal, "response,' the patient may survive without response, "failure,' or the patient may die. If a treatment fails in a given course, physicians usually switch to a different treatment for the next course. Most designs for such settings waste information by ignoring the multi-course and multivariate structure and identifying only one treatment per patient. We describe a family of clinical trial designs to evaluate outcome-adaptive, covariate-adjusted, multi-course treatment strategies. The designs utilize a Bayesian framework incorporating historical data and accommodating multiple treatment courses, trinary outcomes, and patient covariates. Based on a trade-off function between the probabilities of response and death, within prognostic subgroups, the design drops inferior strategies interimly and selects the best strategy at the end. The method is illustrated by a randomized two-course, three-treatment leukemia trial, and a simulation study is described.

e-mail: rex@mdanderson.org

---

61. Selection Bias, Verification Bias, and Propensity Scores

EXPLAINING POST-PROPENSITY SCORE COVARIATE BALANCE

Thomas E. Love*, Center for Health Care Research & Policy, Case Western Reserve University,

In observational studies of treatment effects, we regularly adjust for the possible effects of selection bias through methods such as propensity scores. In presenting the results of such studies, it is crucial to convince skeptics that, after adjustment, baseline covariates are balanced (comparable) between treated and untreated groups. Publications in the health care literature typically display this balance with massive tables of summary statistics. We develop more efficient and effective means of communication through graphical illustrations of the impact of propensity score matching, stratification or adjustment. The results should aid researchers to present the results of these sometimes complicated analyses more effectively to clinicians and other consumers.

e-mail: tel3@po.cwru.edu

REDUCING BIAS IN OBSERVATIONAL STUDIES USING PROPENSITY SCORE: A SIMULATION STUDY

Daniel C. Park*, GlaxoSmithKline
Kwan R.Lee, GlaxoSmithKline
Sam S. Joo, GlaxoSmithKline
Xiwu Lin, GlaxoSmithKline
Timothy W. Victor, AstroZeneca

The application of the propensity score method to create matched-controlled groups in observational studies has recently received great attention. However, it also raises questions as to whether propensity scores can produce estimates comparable to experimental conditions. In this paper, we perform a simulation study to demonstrate how baseline bias in observational studies can be reduced through the propensity score method and how closely the propensity score method can estimate the true value that we assign in the simulation study. Three different population groups (treatment, control, and 'common covariates' groups) with 10 variables each were created with multivariate normal for continuous variables and bivariate normal for binary variables. Variables in the study are generated to be similar to those in any clinical data. The simulation specs included sample size, percent of 'common covariates' and noise to signal ratio, which was two to one. We then compared three different matching methods to observational data; the matching methods were optimal matching and greedy matching using the propensity score method, and naive matching. In this simulation study, we used the regression model to estimate the effect of treatment. This regression model consists of one treatment and ten baseline charaterastics variables as independent variables. if all the baseline characteristics are perfectly balanced out by propensity score, the original value of parameter of treatment would be obtained .

e-mail: daniel_c_park@gsk.com

A GENERALIZED IMPUTATION METHOD BASED ON PROPENSITY SCORE

Hsiao-Lan Wei*, Department of Biostatistics, University of Pittsburgh
Howard E.Rockette, Department of Biostatistics, University of Pittsburgh

To address missing data problems, non-parametric imputation such as the propensity score imputation method proposed by Lavori, Dawson, and Shera (LDS, 1995) is an applicable approach to avoid potentially complicated modeling and laborious computation that may occur for the parametric approach, especially for data with a non-monotone missing pattern. We generalize the LDS method by conditioning on the previous observation indicators when estimating the propensity score and by incorporating a smoothing function of the propensity score prior to stratifying the data for imputation. The proposed imputation methods improve the estimates and inferences when the missing data pattern has some underlying correlations with the repeated measurements, and when the missingness relates to a moving average of the repeated measurements. The evaluations were conducted for monotone and non-monotone missing pattern, for binary and continuous measurements, and for MAR and non-ignorable missingness. The generalized procedure is applied to a study attempting to determine the risk factors of depression and onset of episodes of depression for midlife women from a mental health study.

e-mail: hswst7@pitt.edu

UNCERTAINTY ENVELOPE FOR DIAGNOSTIC TEST IGNORANCE REGION

Ying Chen*, Emory University
Kosinski Andrzej, Emory University

To estimate the sensitivity and specificity of a diagnostic test requires verification of true disease status for each study patient. In practice, not all study patients have disease status verified, and using only verified cases when estimating sensitivity and specificity often leads to biased results, commonly known as verification bias. Due to the incompleteness of data, Kosinski et al. conducted global sensitivity analysis and proposed a test ignorance region (TIR) containing all sensitivity and specificity values consistent with the observed data. The concept of the region of ignorance was introduced by Molenberghs et al., who also considered region of uncertainty, i.e. the confidence region of the true ignorance region. In this paper, we first give the uncertainty interval for the general bounds of the TIR. Second, an approach for the construction of uncertainty envelope for TIR is proposed. Finally, we provide simulation results and present two clinical examples. Key words: Uncertainty envelope; Sensitivity; Specificity; Diagnostic test; Verification bias; Missing data

e-mail: ychen7@sph.emory.edu

---

62. Estimation and Modeling in the Presence of Missing Data and Nonresponse

ESTIMATION OF PARAMETERS FOR BIVARIATE SURVIVAL DATA IN THE PRESENCE OF MISSING COVARIATE DATA

Gina M. D'Angelo*, University of Pittsburgh GSPH Biostatistics Department
Lisa Weissfeld, University of Pittsburgh GSPH Biostatistics Department

Traditional methods of analysis model survival endpoints marginally, assuming independence between the outcomes, and perform complete case analysis assuming that the data are missing completely at random. Although data are often modeled assuming independence among outcomes with the complete case analysis, this approach is not necessarily valid. In this presentation, a method is proposed to handle covariate data that are missing at random when jointly modeling the two survival endpoints.

Dependence between two survival endpoints will be measured via copulas. Marginal functions will measure relationships between each survival outcome and a set of covariates. The 2-stage method proposed by Shih and Louis (1995) will be used to estimate the dependence and marginal functions separately. Since missing data is present among the covariates, we propose the use of inverse probability weighting estimating equations (Robins, Rotnitzky, and Zhao, 1994) to estimate the parameters of the marginal survival functions. Our extension will be compared to complete-case analysis in a breast cancer data example.

e-mail: dangelo@upci.pitt.edu

STATISTICAL MODELING IN TUMOR XENOGRAFT EXPERIMENTS WITH INCOMPLETE DATA

Ming Tan, University of Maryland Greenebaum Cancer Center
Hong-Bin Fang*, University of Maryland Greenebaum Cancer Center
Guo-Liang Tian, University of Maryland Greenebaum Cancer Center
Peter J. Houghton, St. Jude Children's Research Hospital

In cancer drug development, xenograft models where mice are grafted with human cancer cells are used to elucidate the mechanism of action and to assess efficacy of a promising compound. The demonstrated activities in these models is an important step to bring the compound to human. To estimate the dose-response relationship, we propose a class of regression models to characterize the dose-response relationship in xenograft experiments. However, the incomplete repeated measurements are caused mouse death or sacrifice during the experiment, drastic tumor shrinkage ($< 0.01 cm^3$) or random truncation. Because of the small sample sizes in these models, asymptotic inferences are usually not appropriate. The intrinsic growth of tumor in the absence of treatment constrains the parameters in the regression and causes further difficulties in statistical analysis. We develop two inter-related methods, a MLE approach via the ECM algorithm and a Bayesian approach that takes advantage of the likelihood results, to estimate the dose-response relationship. A real xenograft study on two new antitumor agents is analyzed.

e-mail: hfang@umm.edu

A HYBRID MODEL FOR LONGITUDINAL BINARY RESPONSES SUBJECT TO NON-IGNORABLE DROPOUT

Kenneth J. Wilkins*, Department of Biostatistics, Harvard School of Public Health
Garrett M.Fitzmaurice, Department of Biostatistics, Harvard School of Public Health

This paper presents a likelihood-based method for handling non-ignorable dropout in longitudinal studies with binary responses. The methodology developed is appropriate when the target of inference is the marginal distribution of the response at each occasion and its dependence on covariates. A 'hybrid' model is formulated which is designed to retain the most advantageous features of the selection and pattern-mixture model approaches. This formulation accommodates a variety of assumed forms of non-ignorable dropout, while maintaining transparency of the constraints required for identifying the overall model.
Once appropriate identifying constraints have been imposed, likelihood-based estimation is conducted via the EM algorithm. The paper concludes by applying the approach to an example using data from a randomized clinical trial comparing two doses of a contraceptive.

e-mail: kwilkins@hsph.harvard.edu

ESTIMATING VACCINE EFFICACY USING AUXILIARY OUTCOME DATA AND A SMALL VALIDATION SAMPLE

Haitao Chu*, Department of Biostatistics, Emory University
M. Elizabeth Halloran, Department of Biostatistics, Emory University

In vaccine studies, a specific diagnosis of a suspected case is usually conducted by some expensive biological tests on some hard-to-collect samples. Implementing validation sets in the study costs less and is easier to carry out. Estimates of vaccine efficacy (VE) could be severely attenuated if based only on the non-specific measure. Applying missing data analysis techniques could correct the bias while maintaining statistical efficiency (Halloran and Longini, 2001). However, when the sample size in the validation sets is small and the vaccine is highly efficacious, a zero count is likely to occur in the vaccinated validation set. Two commonly used methods, the mean score method (Pepe and Fleming, 1994) and multiple imputation (Rubin, 1997), depend on the ad hoc continuity correction, and the normality or log-normality assumption of relative risk. We propose a Bayesian method to estimate VE and its highest probability density (HPD) credible set using MCMC methods. Comparing the performance of these approaches using data from a field study of influenza vaccine and simulations, we demonstrate the superiority of
Bayesian method in this specific situation.

e-mail: hchu2@sph.emory.edu

ESTIMATION BASED ON RESPONSE-BIASED DATA IN THE PRESENCE OF SUPPLEMENTARY DATA

Jason Schroeder*, University of Memphis

We consider distribution estimation in situations where the response is missing for a portion of the individuals in a sample. We suppose that the missingness is response-related, and that the missing-data mechanism is difficult or even impossible to model. It is well known that simply ignoring the missing data in this situation will lead to biased estimation. A further assumption is that supplementary data are available in some form. We present methods of estimation based on the combination of the incomplete and supplementary data. We give asymptotic properties of the estimators and results from simulation studies examining the finite-sample properties of the estimators. Two examples are presented illustrating these methods. The first example involves measurement of HIV-RNA in blood samples by assays having detection limits, beyond which reliable measurements cannot be recorded. Supplementary data related to the missing or unreliable measurements then contain valuable information about the distribution of the response. In the second example, IQ test scores are available for only some of the subjects in a sickle cell disease study; however, complete covariate data recorded for all subjects provide auxiliary information about the missing scores.

e-mail: rschrodr@memphis.edu

IMPLICATIONS OF POOLING RNA SAMPLES FOR MICROARRAY EXPERIMENTS

Xuejun Peng*, University of Kentucky
Constance L.Wood, University of Kentucky
Arnold J. Stromberg, University of Kentucky

RNA samples extracted from different subjects are often pooled and applied to a single microarray ~{!0~}chip~{!1~}. Modeling the resulting gene expression as a mixture of individual responses, we derive expressions for the resulting experimental error and provide both upper and lower bounds for its value in terms of the variability among individuals and the number of RNA samples pooled. Using data from real experiments and computer simulations, we investigated the statistical properties of RNA sample pooling. Our study reveals that pooling biological samples in a technically correct way is both statistically valid and cost effective for microarray experiments and that optimal pooling design with optimal estimate of the number of samples and arrays can be found to meet statistical power requirement with minimal total cost. One of the most important implications of these results is that RNA pooling improves efficiency and cost-effectiveness for microarray experiments.
Keywords: Microarray analysis, Pooling, Replicates, Power, Sample size determination

e-mail: peng@ms.uky.edu

STATISTICAL MODELING OF MICROARRAY DATA: VIABLE ALTERNATIVES TO THE DIFFERENCE MODEL

Janet K. Holt*, Dept. of Educational Technology, Research and Assessment, Northern Illinois University
T. Mark Beasley, Department of Biostatistics, The University of Alabama at Birmingham
David B. Allison, Department of Biostatistics, The University of Alabama at Birmingham

The general linear modeling approach to oligonucleotide microarray data subsumes various statistical models for data analysis: (a) a covariate model, in which perfect match signal is some estimated linear function of mismatch signal and other variables, (b) a perfect match-only model, in which mismatch data is not utilized, and (c) a difference model of perfect match - mismatch. (Chu, Weir, & Wolfinger, 2002). In this paper we propose a structural model that is useful in choosing among these analytic methods. By decomposing the correlations among the variables in the structural model and making certain assumptions, we theoretically derive the statistical model that most accurately reflects the actual gene expression level. Contrived data are used to illustrate the comparisons among the models under varying conditions. Results show that when modeling non-systematic variation, the covariate model provides maximum flexibility and often more accurately reflects the actual gene expression levels as compared to the difference model. Further, the PM-only model can be a useful alternative in certain circumstances.

e-mail: mbeasley@uab.edu

ASSESSING DIFFERENTIAL EXPRESSION WITH FEW OLIGONUCLEOTIDE ARRAYS

Jianhua Hu*, Dept. of Biostatistics, UNC-Chapel Hill
Fred A.Wright, Dept. of Biostatistics, UNC-Chapel Hill

Expression microarray techniques allow researchers to monitor the relative expression of thousands of genes simultaneously. One problem that arises repeatedly is the need to efficiently identify differentially expressed genes across two conditions. Due to the expense or difficulty of obtaining biological material, it may not be possible to hybridize more than a few arrays in each condition. Thissituation poses serious difficulties for standard t-test or rank-based comparisons, and fully Bayesian approaches may not be attractive. By developing a model for expression indexes in oligonucleotide arrays, we derive a modified t-statistic and show that it has a natural likelihoodinterpretation. Analytic and numerical results demonstrate that our proposed statistic results in lower false discovery rates than a variety of competing statistics.

e-mail: jhu@bios.unc.edu

ON THE PROBLEMS OF SPORADIC POINTS AND LOCAL MINIMUM IN UNSUPERVISED LEARNING: WITH APPLICATIONS TO MICROARRAY ANALYSIS

George C. Tseng*, Harvard School of Public Health, Department of Biostatistics
Wing H.Wong, Harvard University, Department of Statistics and Biostatistics

In this paper we propose methods for two critical problems encountered in cluster analysis: the problem of convergence to a local minimum, and the problem of identifying points that should not be clustered. We find that the first problem can often be avoided by the use of the early truncated hierarchical tree as a robust initial value in the optimization. For the second problem, most current clustering algorithms assign all points into clusters. In many applications, however, there exist points that should not be assigned into any cluster. For example, when clustering genes in microarray analysis, some genes essentially do not belong to any of the significant biological pathways. These 'unclusterable' points are called sporadic points. We propose a method to identify sporadic points by calculating stability strength in randomly split data. By removing these points, the remaining points will be better clustered and interpreted. The stability strength can also serve as a measure of the tightness of the clusters. Finally we validate our methods in two simulations of 3 and 14 clusters with noises. We then examine a set of yeast cell cycle expression data using this method.

e-mail: ctseng@hsph.harvard.edu

INCORPORATING ESTIMATION VARIABILITY INTO DETECTION OF DIFFERENTIAL GENE EXPRESSION IN MICROARRAY EXPERIMENTS

Lei Shen*, School of Public Health, The Ohio State University

The main interest of a microarray experiment often lies in detecting genes that are differentially expressed between two experimental conditions, and many statistical approaches have been proposed for this problem. On the one hand, the usual single-slide cDNA array method and some methods for oligonucleotide arrays are based on the relationship between the variance and the mean. On the other hand, for a decent number of replicates, a common choice is a permutation-based test, which does not make use of the afroementioned relationship. We first discuss how estimation variability is separate from sampling variability, and that the former can be assessed when gene expressions are estimated using a model-based method, with weights assigned to take into account the mean-variance relationship. Then we show how estimation variability can be incorporated into various methods to better determine differential gene expression. If time allows, situations more complex than a simple two-group comparison, such as paired samples and longitudinal measurements, will be discussed.

e-mail: lshen@stat.ohio-state.edu

IDENTIFYING DIFFERENTIALLY EXPRESSED GENES IN UNREPLICATED MULTIPLE-TREATMENT MICROARRAY TIMECOURSE EXPERIMENTS

Rhonda R. DeCook*, Iowa State University
Dan Nettleton, Iowa State University

Microarray technology has become widespread as a means to investigate gene function and metabolic pathways in an organism. A common experiment involves probing, at each of several time points, the gene expression of experimental units subjected to different treatments. Due to the high cost of microarrays, such experiments are often unreplicated. Though an experiment with replication would provide more powerful conclusions, it is still possible to identify differentially expressed genes while controlling the false discovery rate. We present such a method thatutilizes polynomial regression models to approximateunderlying expression patterns over time for each treatment and gene. Models involving treatment effects, terms polynomial in time, and interactions between treatments and polynomial terms are considered. The 'best' model is chosen for each gene using Schwarz's Bayesian InformationCriterion. Genes whose 'best' model differs significantlyfrom the simplest possible model involving only an overall mean are considered potentially biologically interesting. A two-stage permutation testing approach is used to identify such genes. The expected proportion of false positive results among all positive results is estimated using a method presented by Storey(2001).

e-mail: rdecook@iastate.edu

SMOOTH BACKGROUND ESTIMATION IN SPECTRA AND IMAGES

Paul H.C. Eilers*, Department of Medical Statistics, Leiden University Medical Centre

Many instrumental signals, like spectra and chromatograms, and many images, like those of microarrays, show a a smoothly varying background. To better estimate the real signal, it is of great value to estimate the background reliably and then subtract it from the data. I will discuss several new approaches to this problem.

The smooth curve (surface) is modelled with (tensor products of) B-splines. Smoothness is tuned with differences penalties on the B-spline coefficients. For the fit to the data asymmetric measures are used. This means that the weight of a residual depends on its sign. Positive residuals get a small weight, while negative ones are weighted heavily. With an L1 norm (sum of absolute values), a smooth percentile curve is being estimated. The computations, using linear programming, are relatively involved. An asymmetric least squares (L2) norm, leads to simple and fast iterative weighted smoothing. A more principled alternative models the data as a mixture of 1) a smooth curve plus small noise, and 2) a signal with an asymmetric distribution. The computations are simple and fast and useful diagnostic statistics are obtained in the process.

To illustrate the performance of the algorithms, I will presented several real-live applications.

e-mail: p.eilers@lumc.nl

MODELING DRIFT OF FMRI DATA WITH PENALIZED SPLINE

Wen-Lin Luo*, Department of Biostatistics, University of Michigan
Thomas E.Nichols, Department of Biostatistics, University of Michigan

One of key features of fMRI data is the drift that exhibits Diverse shapes and is usually highly nonlinear and heterogeneous over time. Drift often far exceeds both the ambient noise level and the amplitude of the task-related signal change. It is important to model drift, since it can either be confounded with the task-related signal, or it can add to the estimate of the random noise.

In this work, we propose to model the fMRI signal using semiparametric model, where the task-related signal is estimated parametrically and drift is modeled nonparametrically as a spline. Penalized spline (P-spline) is applied to model the nonconstant drift, since usually more than hundred of time points are in a fMRI study and P-spline uses much less knots (<40) than smoothing spline.

Mixed model representation of P-spline is applied to allow estimation of smoothing parameter, where the coefficients of the basis functions are treated as random. Several issues about P-spline, like the number of knots, and degree of basis function, are investigated to check their effects on the estimation of fMRI task-related signal and removal of possible correlation between observations. Further, byusing mixed model, the correlation between observation can be estimated simultaneously with smoothing parameter.

e-mail: wluo@umich.edu

SENSITIVITY OF FWE-CORRECTED INFERENCES ON T STATISTIC IMAGES

Thomas E. Nichols*, University of Michigan Biostatistics
Satoru Hayasaka, University of Michigan Biostatistics

Functional brain imaging methods, such as Positron Emission Tomography (PET) and Functional Magnetic Resonance Imaging (fMRI), create statistic images to measure evidence for an effect of interest. Choosing a threshold for such images represents a massive multiple testing problem, as there are typically 100,000 spatial elements in the image. The standard approach to this multiple testing problem is to use the theory of random t fields to find a threshold which controls the familywise error (FWE) over the whole brain. We present evidence that such an approach is quite conservative for low degrees of freedom, even when the assumptions appear to be satisfied. An alternative, nonparametric approach is to permute the data to build empirical null distributions of interest. While parametric tests are usually more powerful than corresponding nonparametric tests, we find that the permutation test consistently outperforms the random t field results for analyses with low degrees of freedom.

e-mail: nichols@umich.edu

SEMI-LOCAL WAVELET DENOISING OF PET IMAGES

Richard Charnigo*, Case Western Reserve University, Statistics
Raymond Muzic, Case Western Reserve University, Radiology, Oncology, Biomedical Engineering
Jiayang Sun, Case Western Reserve University, Statistics

Wavelet techniques are useful for denoising one-dimensional signals and two-dimensional images. These same techniques can be carried over to three-dimensional problems, such as the denoising of volumetric PET images. However, even the most judicious selection of thresholding parameters does not provide the flexibility to enforce different degrees of smoothing across different regions of an image. Ideally, one would like to enforce a high degree of smoothing in regions where noise dominates physiological activity without discarding valuable information in regions where the physiological activity dominates. To this end, we propose a "semi-local" paradigm for denoising. Semi-local denoising involves the division of an image into blocks, which are then individually thresholded in a way that takes into account the physiological activity within and just outside of the block. Experiments with phantom data show that the semi-local paradigm provides effective denoising, both in terms of MSE reduction and visual quality. Experiments with clinical data provide further insight about some practical aspects of semi-local denoising.

e-mail: rjc12@po.cwru.edu

MULTIVARIATE SELECTION OF BIOMARKERS FOR METABOLIC SYNDROME FROM 2D GEL PROTEOMICS DATA

Kwan R. Lee*, Data Exploration Sciences
GlaxoSmithKline Pharmaceuticals R&D
Xiwu Lin, Data Exploration Sciences
GlaxoSmithKline Pharmaceuticals R&D
Daniel C. Park, Data Exploration Sciences
GlaxoSmithKline Pharmaceuticals R&D
Sergio Eslava, Data Exploration Sciences
GlaxoSmithKline Pharmaceuticals R&D

Two dimensional polyacrylamide gel electrophoresis (2D-PAGE) separates proteins by isoelectric point (pI) and molecular weight (mw). Problems in the technology include image processing, missing values and identifying significantly different spots in two different gels. This talk will be about an multivariate approach to the 2-D gel data obtained from the metabolic syndrome (MTS) clinical trials where the major goal is to identify protein markers responsible for the discrimination of the affected group from the normal group.

The multivariate approach was useful in the sense that it could separate the biological variation from the random variation but we will discuss potential problems. Selection of protein markers seems to be a nontrivial exercise and there may be competing sets of markers that can differentiate the MTS as well as the chose markers by a single data model.

e-mail: kwan.lee@gsk.com

STATISTICAL TECHNIQUES IN PROTEOMICS – ANALYSIS OF 2D GEL DATA

Sreelatha Meleth*, Medical Statistics Section, Dept of Medicine, UAB
Jessy Deshane, Department of Pharmacology & Toxicology, UAB
Helen Kim, Department of Pharmacology & Toxicology, UAB

Technological advances, as well as advances in image analysis software are making the use of 2D gels in the study of complex biological systems, ever more accessible to protein biologists. In spite of the availability of several software packages however, a number of statistical issues are not addressed in the analyses of 2D gels that routinely appear in publications. Some of the outstanding issues that limit current 2D gel analysis include: 1) Reducing dimensionality of the data without foregoing valuable biological information.2)The best models to fit data that is highly non-normal. 3)The best values to use as minimum values for spot intensity when using a transformation (e.g. log transformation). 4)Differentiating statistically between a spot that is missing because due to low intensity, versus a spot with a lowered intensity due to a specific treatment? 5)The biological cost of adjusting for multiple testing? Should biological validity overtake statistical validity in exploratory techniques such as these? In this study, we address these issues using simulations as well as real data, and propose some answers.

e-mail: dhanya@uab.edu

ESTIMATION PROBLEMS FROM DATA WITH CHANGE POINTS AND IMAGES

Stephen J. Ganocy*, Case Western Reserve University
Jiayang Sun, Case Western Reserve University

This paper presents new procedures for estimation problems from data with change points and images. The motivation for this work comes from the analysis of 'footprint' data in the tire industry. A reasonable model for the data involves a regression model with 2 change points and heteroscedastic errors. This model is more general than those found in the literature such as one change point with independent or dependent error, or multiple change points with homoscedastic error.

We develop estimators for the regression coefficients and associated variances in two ways. One is based on the likelihood approach. An algorithm is established to solve for the unknown parameters which can be accelerated by choosing good starting values obtained by visual inspection of the raw data plot. Finite sample performance is studied via simulation studies. We show our estimates are consistent and have some good asymptotic properties. Secondly, a Bayesian approach is used to estimate the parameters. A new Bayesian algorithm, MMP (Maximization-Maximization-Posterior) is proposed. The algorithm has fast convergence to the true change points.

e-mail: sjganocy@aol.com

65. Inference on Risks and Rates

EFFICIENT ESTIMATION OF THE RISK OF A DISEASE BY QUARTILE-CATEGORIES OF A PREDICTOR VARIABLE USING GENERALIZED ADDITIVE MODELS (GAMS).

Craig B. Borkowf*, Centers for Disease Control and Prevention
Paul S.Albert, National Cancer Institute

Suppose that one wishes to make inference to the risk of a disease by the population quartile-categories of a particular risk factor. In the standard approach, one categorizes the continuous risk predictor variable by its empirical quartiles. One then constructs a generalized linear model (GLM), perhaps with the logistic link, to relate the probability of disease to a linear combination of known functions of predictor variables, including the empirical quartile-categories of the chosen risk factor.

Alternatively, one may construct a GAM, which nonparametrically estimates the functional form of the contribution of the risk factor to the GLM. Unlike the standard approach, this method employs information about the actual values or empirical percentiles of the risk factor, which makes the estimation process more efficient. One can then integrate under the estimated curve to estimate the risk of disease by quartile-category and, in turn, one can compute odds ratios and other desired statistics. Graphical methods enable one to make inferences about the true relationship between the disease and the risk factor. An example from nutritional epidemiology is presented.

e-mail: CBorkowf@cdc.gov

JOINT ANALYSIS OF INCIDENCE, PROGRESSION, REGRESSION AND DISAPPEARANCE RATES

Guan-Hua Huang*, Dept. of Population Health Sciences, Dept. of Biostatistics and Medical Informatics, University of Wisconsin-Madison
Ronald Klein, Dept. of Ophthalmology and Vision Sciences, University of Wisconsin-Madison
Barbara E.K. Klein, Dept. of Ophthalmology and Vision Sciences, University of Wisconsin-Madison

Age-related maculopathy (ARM) is a leading cause of vision loss in people aged 65 or older. ARM is distinctive in that it is a disease, which can transition through incidence, progression, regression, and disappearance. The purpose of our study is to develop a methodology for studying the relationship of incidence, progression, regression and disappearance rates with risk factors. We can constrain the population to a certain disease status at the baseline and then analyze the probability of the constrained population to develop a different status. While this approach is intuitive, we risk losing available information and encountering insufficient sample size. We develop a transitional model for analyzing such disease. Our approach provides the "conditional probability" of a current disease level based upon a previous level, and can thereby jointly analyze the incidence, progression, regression and disappearance rates. An analysis to determine the birth cohort effect on ARM is used for illustration.
Keywords: birth cohort effect, eye disease, longitudinal data, transitional model.

e-mail: guanhuahuang@facstaff.wisc.edu

DATA DEPTH AND INFANT MORTALITY DIFFERENTIALS

Ann F. Von Holle*, UNC, Chapel Hill.
Chirayath MSuchindran, UNC, Chapel Hill.

Infant mortality in the U.S. has declined more than 90 percent over the past century. Despite this decline, differentials still exist between racial groups. An obvious way to detect these differences is through parametric comparisons. However, I will describe an alternative for cases when the data distribution, as in this case for county-level infant mortality rates, may not lend itself well to parametric measures. Simplicial depths when used in this context provide an approach to explore data characteristics such as kurtosis and dispersion with few assumptions. More calculations based on these depths follow, including data-depth plots, shrinkage plots, convex hull volume plots and maps by depth status. Repeating the calculations for each of four groups of years enables comparisons that show infant mortality differentials mostly unchanging for blacks and whites over the past twelve years. Another highlight of this approach is its ability to robustly identify extreme values for this set of variables, important for detecting very high infant mortality rates while considering other characteristics of the county.

e-mail: avonholle@unc.edu

EFFECTS OF DATA LIMITATIONS WHEN MODELING FATAL OCCUPATIONAL INJURY RATES

James F. Bena*, National Institute for Occupational Safety and Health
A. JohnBaliler, Dept. of Mathematics and Statistics, Miami Univesity.
National Institute for Occupational Safety and Health.

The effects of sources of error and variability on models for fatal occupational injury rates are explored. Underreporting of fatalities, variability in employment estimates, and choice of employment metric (e.g. number employed vs. hours employed) can change the results obtained from the modeling of fatality data. Poisson regression models are used to explore the trends in fatality rates over time. Models representing true population values are compared with those affected by undercount errors and employment variability, via simulation. The choice of an employment measure, the denominator in rate calculations, is evaluated to investigate the effects of linear interpolation and differences in hours worked among subgroups of the population. Changes in point estimates and standard errors are used to evaluate the effects. While the underreporting of fatalities, if consistent across time, appears to have little effect on the trend, the choice of employment metric can change the trend estimate dramatically, especially among subgroups of the population such as the youngest and oldest workers. These results are related to recent findings in occupational fatal injury rate trend analyses.

e-mail: jbena@cdc.gov

A Comparison of Variable Dimension Models and Overparameterized Models for Spatially Clustered Disease Rates

Ronald E. Gangnon*, University of Wisconsin - Madison
Murray K.Clayton, University of Wisconsin - Madison

Maps of regional disease rates are potentially useful tools in examining spatial patterns of disease and for identifying clusters. We describe a partition model for estimation of disease rates. This model includes spatially structured fixed effects associated with the partitions and spatially unstructured random effects. In one variant of this model, we regard the number of partitions as a parameter to be estimated. Inference is then based on a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm. In a second variant of the model, we regard the number of partitions as being fixed. In this case, we intentially overparameterize the model by selecting an overly large number of partitions. The variable-dimension approach provides a more solid inferential base, but the overparameterized model is much simpler to implement. We compare inferences from the two models using both real and simulated datasets.

e-mail: ronald@biostat.wisc.edu

ESTIMATING AND SUMMARIZING INTERACTIONS IN MULTIPLE STRATA

Daniel L. McGee, Department of Statistics, Florida State University
Xu-Feng Niu*, Florida State University

In this report we present an empirical example of detecting and summarizing age by risk factor interactions in multiple studies, which is based on person-level data from 22 studies that include both sexes, black and white groups and samples from the U.S. and other countries. The multivariate proportional hazards model was used to establish the relationships between CHD death and 5 risk factors: age, systolic blood pressure, serum total cholesterol, current cigarette smoking, and diabetes status. Chi-square values for age by risk factor interactions in the 22 studies are calculated after adjusted for sex and race effects. Several methods of combining statistical tests including Fisher's inverse Chi-square method are used to combine the observed significance levels. Initial results show that all the four interactions are statistically significant. The estimated coefficients of the interactions are all negative, which indicates that high blood pressure, high cholesterol, smoking, and diabetes may have less impact on CHD death for old patients than for young patients. Additionally, we present fixed and random effects models to estimate the average interactions in the studies.

e-mail: niu@stat.fsu.edu

A PARAMETRIC MODEL FOR STUDYING ORGANISM FITNESS USING STEP-STRESS EXPERIMENTS

Sonja Greven, University of North Carolina, Department of Biostatistics
A. J.Bailer*, Miami University, Department of Mathematics & Statistics
Lawrence L. Kupper, University of North Carolina, Department of Biostatistics
Keith E. Muller, University of North Carolina, Department of Biostatistics

We propose a method based on parametric survival analysis to analyze step-stress data. Possible applications include experiments on swimming performance of fish using stepwise increasing water velocity, treadmill tests on humans, and step-stress reliability testing. A likelihood-ratio test is developed for comparing the failure times in two groups based on a piecewise constant hazard assumption. The test can be easily extended to other piecewise distributions and to include covariates. An example data set is used to illustrate the method and highlight experimental design issues. A small simulation study compares the new analysis procedure and a currently used method with regard to type I error rate and power.hyun

e-mail: baileraj@muohio.edu

METHODOLOGY FOR EVALUATING A PARTIALLY CONTROLLED LONGITUDINAL TREATMENT USING PRINCIPAL STRATIFICATION, WITH APPLICATION TO A NEEDLE EXCHANGE PROGRAM

Constantine E. Frangakis*, Johns Hopkins University Bloomberg School of Public Health
Ronald S.Brookmeyer, Johns Hopkins University Bloomberg School of Public Health
Ravi Varadhan, Johns Hopkins University Bloomberg School of Public Health
Mahboobeh Safaeian, Johns Hopkins University Bloomberg School of Public Health
David Vlahov, The New York Academy of Medicine
Steffanie A. Strathdee, Johns Hopkins University Bloomberg School of Public Health

We consider studies for evaluating the short term effect of a treatment of interest on a time-to-event outcome. The studies we consider are only partially controlled in the following sense: (1) subjects' exposure to the treatment of interest can vary over time, but this exposure is not directly controlled by the study; (2) subjects' follow-up time is not directly controlled by the study; and (3) the study directly controls another factor that can affect subjects' exposure to the treatment of interest as well as subjects' follow-up time. When factors (1) and (2) are both present in the study, evaluating the treatment of interest using standard methods, including instrumental variables, does not generally estimate treatment effects. We develop the methodology for estimating the effect of treatment (1) in this setting of partially controlled studies under explicit assumptions using the framework for principal stratification for causal inference. We illustrate our methods by a study to evaluate the efficacy of the Baltimore Needle Exchange Program to reduce the risk of HIVtransmission, using distance of the Program's sites from the
subjects.

e-mail: cfrangak@jhsph.edu

DYNAMIC TREATMENT EFFECTS

Karsten T. Hansen*, Kellogg School of Management, Nortwestern University
James J.Heckman, Department of Economics, The University of Chicago

In this paper we consider the formulation, identification and computation of dynamic treatment effects models. Time to treatment is modelled as a discrete time duration model which we link to a set of counterfactural outcomes. Unobserved heterogeneity is modelled using factor analytic representations. We derive conditions for the semi-parametric identification of the models. Estimation is accomplished using Markov Chain Monte Carlo

e-mail: karsten-hansen@northwestern.edu

67. Multivariate Mixed Effects Modeling of Discrete and Continuous Outcomes

MULTIPLE OUTCOMES WITH APPLICATIONS TO RISK ASSESSMENT IN TOXICOLOGY STUDIES

Paul Catalano*, Harvard School of Public Health, Department of Biostatistics

Multiple outcomes in the non-cancer toxicology setting pose significant challenges for dose response modeling in general and quantitiative risk assessment in particular. Within the context of multiple observations there can be complications with respect to whether risk should be "averaged" or "totaled", leading to complexities in the construction of the relevant likelihood and translating the joint distribution into a statement of overall multivariate risk, especially when the outcomes are of mixed discrete and continuous nature. Some studies, such as those in developmental toxicity, involve clustering in addition to the multiplicity on each animal, further complicating the analysis. Longitudinal assays present some of the same problems. In this talk, we review some of the philosophical issues in developing multiple outcome models for overall risk, discuss some of the recent methods for tackling these problems in light of the philosophy and propose some new methodology for quantitative risk assessment. Examples will be drawn from neurotoxocity and developmental studies in animals.

e-mail: pcata@jimmy.harvard.edu

LATENT VARIABLE MODELS FOR MULTIVARIATE DISCRETE AND CONTINUOUS OUTCOMES

Louise M. Ryan*, Department of Biostatistics; Harvard School of Public Health

Latent variable models provide a convenient tool for constructing models involving multivariate outcomes. Such models can easily handle complex settings, such as mixtures of discrete and continuous outcomes. The models are also easily extended to accommodate high dimensional predictors as well. Simpler versions of such models can be easily fit using readily available software such as Proc NLMIXED in SAS, while more complex models are likely to be better fit using a Bayesian framework.
However, we will see that model robustness is a serious concern. Results and findings are illustrated with several public health datasets.

e-mail: ryan@hsph.harvard.edu

BAYESIAN JOINT MODELS OF CLUSTER SIZE AND SUBUNIT-SPECIFIC OUTCOMES

David B. Dunson*, Biostatistics Branch, National Institute of Environmental Health Sciences
Zhen Chen, Biostatistics Branch, National Institute of Environmental Health Sciences

In applications that involve clustered data, such as longitudinal studies and developmental toxicity experiments, the number of subunits within a cluster is often correlated with outcomes measured on the individual subunits. Analyses that ignore this dependency can produce biased inferences. This article proposes a Bayesian framework for jointly modeling cluster size and multiple categorical and continuous outcomes measured on each subunit. We use a continuation ratio probit model for the cluster size and underlying normal regression models for each of the subunit-specific outcomes. Dependency between cluster size and the different outcomes is accommodated through a latent variable structure. The form of the model facilitates posterior computation via a simple and computationally efficient Gibbs sampler. The approach is illustrated through application to developmental toxicity data, and applications to joint modeling of longitudinal and event time data are discussed.

e-mail: dunson1@niehs.nih.gov

JOINT MIXED MODELS FOR DISCRETE AND CONTINUOUS OUTCOMES

Ralitza V. Gueorguieva*, Div. of Biostatistics, Dept. of Epidmiology and Public Health, Yale University

Biomedical researchers are often interested in estimation and testing of treatment effects on multiple outcome variables. Joint analysis of the outcomes variables results in better control of type I error, may allow for estimation of overall treatment effect and may lead to greater power. However, it involves complicated statistical modeling especially when repeated measures are collected on each subject and/or when the outcomes are of different types. In this talk we consider subject-specific mixed models for repeated measurements on multiple discrete and continuous variables. A distinction is made between situations when the outcome variables are affected differently by the treatment, and when the outcome variables measure the same underlying process. We also discuss maximum likelihood estimation, which is computationally intensive but is some special cases may be available in standard software. Comparison of joint and separate modeling of the outcome variables is also considered. Motivating examples from the area of psychiatry are used for illustration.

e-mail: ralitza.gueorguieva@yale.edu

COMPOSITE CONDITIONAL LIKELIHOOD FOR SPARSE DEPENDENT DATA

John J. Hanfelt*, Department of Biostatistics, Emory University

Sparse dependent data arise in finely stratified genetic and epidemiologic studies and pose at least two challenges to inference. First, it can be difficult to model and interpret the full joint probability of dependent discrete data, which precludes the use of full likelihood methods. Second, standard methods for dependent data, such as pairwise likelihood and the generalized estimating function approach, are unsuitable when the data are sparse due to the presence of many nuisance parameters. We present a composite conditional likelihood for use with sparse dependent data that provides valid inferences about covariate effects on both the marginal response probabilities and the intracluster pairwise association. Our primary focus is on sparse dependent binary data, in which case the proposed method utilizes doubly discordant quadruplets drawn from each stratum to conduct inference about the intracluster pairwise odds ratios. We use the approach to conduct inference about the household aggregation of high blood pressure and the familial aggregation of schizophrenia in two finely stratified studies.

e-mail: jhanfel@sph.emory.edu

ADJUSTED PROFILE ESTIMATING FUNCTION

Molin Wang*, Harvard School of Public Health
John J.Hanfelt, Emory University

We propose a simple adjustment to the profile estimating function, extended from the adjusted profile score approach of Cox & Reid (1987), that achieves first order bias correction of the profile estimating function. This adjustment method can be used in both parametric and semiparametric settings. In the semiparametric setting, typically only the first two moments of the response variable are needed to form the adjustment. Important applications of this method include the estimation of the pairwise association in stratified, clustered data and estimation of the main effects in a matched pair study. A brief simulation study shows that the proposed method considerably reduces the impact of the nuisance parameters.

e-mail: mwang@jimmy.harvard.edu

A TYPE OF RESTRICTED MAXIMUM LIKELIHOOD ESTIMATOR OF VARIANCE COMPONENTS

Jiangang Liao*, UMDNJ

The maximum likelihood estimator of the variance components in a linear model can be biased downwards. Restricted maximum likelihood corrects this problem by using the likelihood of a set of residual contrasts and is generally considered superior. However, this original restricted maximum likelihood definition does not directly extend beyond linear models. We propose a REML-type estimator for generalised linear mixed models by correcting the bias in the profile score function of the variance components. The proposed estimator has the same consistency properties as the maximum likelihood estimator if the number of parameters in the mean and variance components models remains fixed. However, the estimator of the variance components has a smaller finite sample bias. A simulation study with a logistic mixed model shows that the proposed estimator is effective in correcting the downward bias in the maximum likelihood estimator.

e-mail: jg_liao@yahoo.com

ELIMINATION OF NUISANCE PARAMETERS VIA LOCALLY-ANCILLARY QUASI-SCORES

Paul J. Rathouz*, University of Chicago

We propose a quasilikelihood extension of the projected score method of Waterman and Lindsay (1996) for the elimination of nuisance parameters. The procedure addresses cases where only the mean and the variance of the response variable are specified and where the mean function involves both parameters of interest and nuisance parameters. Due to the optimality and information-unbiasedness of the quasiscorefunction, a second-order quasiscore basis of estimating functions for the nuisance parameter is derived. Second-order locally ancillary estimating functions are then obtained by solving a linear system that corresponds to a true projection for canonical exponential family distributions. Important applications include models for matched designs and for errors-in-covariates, the latter of which is explored in more detail.

e-mail: prathouz@health.bsd.uchicago.edu

SEMIPARAMETRIC MODEL-ASSISTED ESTIMATION OF DISTRIBUTION FUNCTIONS IN SURVEYS WITH AUXILIARY INFORMATION

Alicia A. Johnson*, Colorado State University
Jay Breidt, Colorado State University
Jean D. Opsomer, Iowa State University

The availability of auxiliary information in many surveys facilitates an increase in efficiency of survey estimators. Classical survey regression estimators incorporate this information through linear, parametric models. Nonparametric methods may also be applied to minimize parametric restrictions and better represent complex relationships between auxiliary information and variables of interest. Multiple auxiliary variables, including categorical variables, are often available. In this case, the nonparametric approach can be extended to an additive semiparametric model, which incorporates both parametric and nonparametric techniques. A semiparametric approach to the estimation of population parameters, including distribution functions, is developed and applied to aquatic resources data from the Environmental Monitoring and Assessment Program. In this application, auxiliary information is obtained from a variety of remotely-sensed images and geographic information system layers. The semiparametric model-assisted distribution function estimator has good statistical properties and desirable operational features.

e-mail: johnsona@stat.colostate.edu

NONPARAMETRIC SURVEY REGRESSION ESTIMATION IN TWO-STAGE SPATIAL SAMPLING

Siobhan P. Everson-Stewart*, Colorado State University
Jay Breidt, Colorado State University
Jean D. Opsomer, Iowa State University

A nonparametric model-assisted survey estimator for status estimation based on local polynomial regression is extended to incorporate spatial auxiliary information and two-stage sampling designs. Under mild assumptions, this estimator is design-unbiased and consistent. Simulation studies show that the nonparametric regression estimator is competitive with standard parametric techniques when a parametric specification is correct, and outperforms those techniques when the parametric specification is incorrect. The methodology is applied to water chemistry data from the EMAP Northeastern Lakes Survey.

e-mail: everson@stat.colostate.edu

USE OF RANDOM PERMUTATION MODEL IN RATE ESTIMATION AND STANDARDIZATION

Wenjun Li*, Biostatistics Research Center, Tufts-New England Medical Center
Edward J.Stanek III, Dept. of Biostatistics and Epidemiology, University of Massachusetts Amherst

Through integrating techniques from several areas in survey sampling, we develop an alternative method of deriving estimators using random permutation models under (stratified) simple random sampling without replacement. The finite random permutation model links the samples to the population. The joint permutation of response and auxiliary variables is modeled using seemingly unrelated regression. We use prediction theory from the super-population sampling literature to derive the linear unbiased minimum variance predictors of population means under the design-based framework using finite estimating equation approach of Binder and Patak (1994). The predictors have similar functional form to those derived using design-based, model-assisted and calibration approaches, but depend on neither superpopulation nor regression model assumptions. We applied the results to standardization of multiple rates, and illustrate how their use accounts for the covariance of the standardized rates, unlike conventional standardization methods.

e-mail: wli1@lifespan.org

A HISTORY OF DATA MINING ACCIDENTS IN THE 20TH CENTURY

Russel M. Upsumgrub, University of Pittsburgh
Ilene Over*, San Hill Institute of Technology
Eileen Forward, San Hill Institute of Technology

'Data mining' became a popular tool for hypothesis generating in the last century. This technique was applied in a number of application areas, e.g., health sciences, economics, and psychology. Unfortunately, a number of problems have arisen due to the 1) quality of databases available; 2) the attempt by investigators to make inference on data which is not representative of the 'target population'; and 3) the identification of false positive results in many applications. Sometimes the data miners have been trapped so far underground that even the best 'inferential tools' cannot save them. We will cover some of the spectacular accidents that occurred during the latter part of the 20th century.

e-mail: sja@pitt.edu

ESTIMATION PROBLEMS FROM BIASED DATA

Bin Wang*, Statistics Department
Case Western Reserve University
Jiayang Sun, Statistics Department
Case Western Reserve University

Data from observational studies may come with selection biases. Based on one biased sample, the density of a population and the bias function are unidentifiable non-parametrically. However, if there are two overlapping samples drawn consecutively, the problem is identifiable. Further, an intelligent subject sampled before, if sampled again, may subject to a memory effect. This paper incorporates the memory effect into a general biased sampling model, with two biased samples from a finite population. Non-parametric estimators of the density, bias, memory functions and the population size are provided.
Asymptotic properties of these estimators are studied. Our procedures are compared with those ignoring the memory effect. The simulation results illustrated that the method proposed gives out stable estimators and may be applied to survey data, gene data and other data with sampling biases.

e-mail: bwang@po.cwru.edu

70. The Study of Health Effects and Health Care

MUTAGENIC POTENCY ESTIMATION IN THE AMES/SALMONELLA ASSAY

Susan J. Simmons*, Univeristy of North Carolina, Wilmington
Walter W.Piegorsch, University of South Carolina
Errol Zeiger, Errol Zeiger Consulting; Chapel Hill, NC

The ability to quantify environmental and other detrimental risks from exposure to hazardous agents is an important component in the process of assessing human and environmental health effects. One such
intellection is in the form of potency estimation, which is viewed as a quantitative measure of a chemical agent's relative toxic capability. Properly calibrated, a positive potency indicates a positive toxic potential, while a zero or negative potency is not considered detrimental.

The focus of this paper is potency estimation for the mutagenesis system known as the Ames/Salmonella
Assay. The energy devoted to this assay throughout the years has produced various forms of statistical potency estimators. In order to gain a better understanding of the stability and distributional properties of several of these estimators, a large simulation study was conducted that examined these features. Although the results of the study indicated potency estimators of mutagenic damage in Ames assay data are highly varied, a few estimators were identified as having desirable properties.

e-mail: simmonssj@uncw.edu

MODELING THE EFFECTS OF A BI-DIRECTIONAL LATENT PREDICTOR FROM MULTIVARIATE QUESTIONNAIRE DATA

Amy H. Herring*, Department of Biostatistics, The University of North Carolina at Chapel HillDavid B.Dunson, National Institute of Environmental Health Sciences
Nancy Dole, Carolina Population Center, Chapel Hill, NC

Researchers often measure stress using questionnaire data on occurrence of potentially anxiety-inducing life events and strength of reaction to these events, characterized as negative or positive and assigned an ordinal ranking. In studying the health effects of stress, one must obtain measures of an individual's negative and positive stress levels to be used as predictors. We propose a latent variable model characterized by event-specific negative and positive reaction scores. If the positive score dominates the negative score for an event, then the individual's response to that event will be positive, with an ordinal ranking determined by the value of the score. The event-level reaction scores are related to individual-level reactivity scores through a factor analytic linear model. Measures of overall negative and positive stress are obtained. By incorporating these measures as predictors in a regression model and fitting the stress and outcome models jointly using Bayesian methods, inferences can be conducted without the need to assume known weights for the different events. We use the approach to study the effects of stress on preterm delivery.

e-mail: aherring@bios.unc.edu

EFFECTS OF MEDICAL MARIJUANA LAWS ON DRUG USE AMONG HOSPITAL EMERGENCY ROOM ADMISSIONS

Li Zhu*, Department of Epidemiology and Biostatistics, Texas A&M University
Dennis Gorman, Department of Epidemiology and Biostatistics, 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. Both log-linear and Poisson regression models were fitted. Confounding factors including time, place, and demographic characteristics were controlled for in the analysis.

e-mail: lizhu@srph.tamushsc.edu

ESTIMATION CONCERNS IN THE MEASUREMENT OF TRANSITIONS IN THE CONCENTRATION OF MEDICAL EXPENDITURES

Steven B. Cohen*, Center for Cost and Financing Studies , Agency for Healthcare Research and Quality
Trena Ezzati-Rice, Center for Cost and Financing Studies, Agency for Healthcare Research and Quality

The capacity to conduct detailed analyses of the health care expenditure experience of the population each year, and to examine patterns over time, are two major features of the Medical Expenditure Panel Survey. The MEPS was designed to produce national and regional annual estimates of the health care utilization, expenditures, sources of payment and insurance coverage of the U.S. civilian non-institutionalized population.The survey design and related estimation specifications feature targeted statistical and methodological strategies to improve the accuracy of resultant health care expenditure estimates. Prior studies have demonstrated that one percent of the civilian non-institutionalized population are associated with over twenty five percent of the medical expenditures incurred in a given year. This paper provides a summary of the statistical and methodological strategies adopted in the MEPS to improve the accuracy of resultant health care expenditure estimates. In addition, the capacity of estimation strategies to adjust for survey attrition are assessed with respect to national health care expenditure estimates and transitions over time.

e-mail: scohen@ahrq.gov

THE USE OF THE KAPLAN-MEIER ESTIMATOR OF THE CENSORING DISTRIBUTION IN THE ESTIMATION OF MEDICAL COSTS

Corina M. Sirbu*, Division of Biostatistics, Department of Epidemiology, Michigan State University
Joseph C.Gardiner, Division of Biostatistics, Department of Epidemiology, Michigan State University

Health care studies are typically designed with a period of recruitment in which patients enter the study, and an additional period of follow up. At study termination however, some patients would have not reached their end-point of interest, which results in right censoring of their time-to-event response and their associated medical cost. We address the estimation of expected net present value for cost incurred over a defined period, based on a selected sub-sample from the original observations in a sample of event times and costs. Selection probabilities are estimated from the censored sample of event times using both parametric and the nonparametric (Kaplan-Meier) estimators of the censoring distribution. Using mean bias and mean-squared error as measures of performance, we investigate the sensitivity of estimators of net present value to the extent of censoring and smoothness assumptions on the censoring and event time distributions.

e-mail: sirbucor@msu.edu

INCORPORATING DEATH INTO SF-36 PHYSICAL COMPONENT SCORE ANALYSES

Laura L. Johnson*, National Cancer Institute
Paula Diehr, University of Washington

The SF-36 is collected in many clinical trials and observational studies. When collected longitudinally, and in studies involving the elderly or chronic diseases, a number of study participants may die while the study is still collecting information. These data are not missing because of compliance problems and to analyze it in such a manner often will not make sense. It may produce biased results to exclude patients who die from the analysis. This methodological issue has been discussed in the statistical literature using several different ways of incorporating death into the analysis of longitudinal data. We jointly model survival and SF-36 Physical Component Score data while trying to ensure statistical accuracy and ease of interpretation. We believe this work has the potential to provide important information to investigators on ways to analyze trial data and in turn to provide better information on changes in physical health status to clinicians and patients.

e-mail: lauralee@alumni.virginia.edu

COMPARISON OF STANDARD VERSUS MODIFIED EXPIRATORY TECHNIQUES DURING SPIROMETRY

Madhuri S. Mulekar*, Mathematics & Statistics, University of South Alabama
Tony Cowan, Mobile Diagnostic Center, Providence Hospital
William Pruitt, University of South Alabama

Current instructions for patients performing spirometry call for maximal effort throughout the expiratory effort. The patients often show signs of discomfort and fatigue when performing this test and are reluctant to return to the PFT lab for serial measurements. We compared the effects of current standard instructions and modified instructions for performing spirometry on the measurements, the patient condition during each test and their willingness to repeat each of the tests. Outpatients were tested in the PFT lab using a crossover design with both the standard and the modified instructions for spirometry. Both techniques resulted in comparable clinically significant spirometry measurements, but with the modified technique, patients showed decreases in discomfort, fatigue, lightheadedness, and dyspnea, and increases in willingness to repeat the study. Compliance with testing was also higher using the modified technique.

e-mail: mmulekar@jaguar1.usouthal.edu

SMOOTH AND ITERATIVE ESTIMATORS FOR COPULAS

Kouros Owzar*, Department of Biostatistics and Bioinformatics, Duke University Medical Center
Pranab K.Sen, Departments of Biostatistics and Statistics, University of North Carolina at Chapel Hill

A bivariate copula can be statistically interpreted as a bivariate distribution function with uniform marginals. Sklar (1959) argues that for any bivariate distribution function, say H with marginals F1 and F2, there exists a copula function, say C, such that H[x,y]=C[F1[x],F2[y]], for (x,y)^T in the support of H.

There exists a rich family of parametric copulas which yield any bivariate distribution expressible as a function of its marginals and a finite-dimensional dependence parameter. For these models, the copula stipulates the structure of the dependence, while the finite-dimensional parameter controls the degree of the dependence.

An existing methodology, for estimating the dependence parameter, consists of plugging-in the marginal empirical distribution functions for the corresponding marginals in the likelihood function. The dependence parameter is then estimated by the M-estimator of the resulting plug-in pseudo-likelihood. This approach has been discussed, among other papers, in Genest, Ghoudi and Rivest (1995) and in Shih and Louis (1995).

The aforementioned methodology is potentially beset by a number of shortcomings. What is to be presented in this work, is a brief discussion on these potential shortcomings along with the proposition of new methodology to address and remedy these very issues.

e-mail: kouros.owzar@duke.edu

A COPULA MODEL FOR MIXED DISCRETE AND CONTINUOUS OUTCOMES

Wei Lang*, Department of Public Health Sciences, Wake Forest University School of Medicine
Lisa A.Weissfeld, Graduate School of Public Health, University of Pittsburgh
Ralph B. D'Agostino, Jr., Dept. of Public Health Sciences, Wake Forest University School of Medicine

Mixed outcome data are generated in many different settings where a binary outcome denoting the presence or absence of a characteristic is collected in addition to continuous outcome data. Analysis methods proposed include latent variable model, extension of the general location model, and a correlated probit model. Lack of appropriate multivariate distribution family causes difficulty in constructing parametric model for the joint analysis of correlated discrete and continuous outcomes. We propose a copula-based parametric model (CPM) as an alternative analytical approach. The CPM does not require the factorization of the joint density into several components, therefore regression parameters can be directly interpreted in the model. Data from the family based treatment of severe pediatric obesity study are analyzed using the CPM with mixed outcomes being subject's weight measures and the development of Type II diabetes.

e-mail: wlang@wfubmc.edu

CONFIDENCE INTERVALS OF THE RATIO (OR DIFFERENCE) OF TWO INVERSE GAUSSIAN MEANS

Lili Tian*, Department of Statistics, University of Florida
Gregory Wilding, Biostatistics, University of Buffalo

Inverse Gaussian (IG) distribution is an ideal candidate for modeling positive, right-skewed data and many of its inference methods are analougous to those for Gaussian family. However, there have always been reservations about the use of IG in data analysis partially due to the fact the confidence interval involving two IG means (difference or ratio) is not available. In this paper, we presents a couple of approaches to obtain the approximate confidence intervals involving two Inverse Gaussian means. The performance of the proposed approaches for small sample size is assessed by simulation studies.

e-mail: ltian@biostat.ufl.edu

URN SAMPLING AND INTERVAL CENSORING

Lu Zheng*, Harvard School of Public Health
Marvin Zelen, Harvard School of Public Health and the Dana Farber Cancer Institute

This paper extends the urn sampling analogue of the score statistic relating survival to covariates to interval-censored data. Three large sample approximations to the urn sampling model are proposed and numerical studies show that the approximation works at significance levels 0.01 and 0.05 when the interval length is small relative to the mean survival time and the proportion of interval-censored observation is less than 40%. The conservativeness of the approximations is discussed.

e-mail: lzheng@hsph.harvard.edu

A COMPARISON OF POWER OF LOG RANK AND WILCOXON TESTS: AN APPLICATION TO HEMATOLOGIC DISORDERS STUDY

Pingfu Fu*, Department of Epidemiology and Biostatistics, Case Western Reserve University
Mary J.Laughlin, Dept. of Medicine/University Hospitals of Cleveland, Case Western Reserve University

The power of several two-sample tests is compared by simulation for small samples from piece-wise distribution of log-logistic and Weibull distributions with different percentage of censoring. The tests considered are Mantel-Haenszel or log rank test and Peto-Prentice's Wilcoxon test (Peto and Peto, 1972; Prentice, 1978). Our simulation study shows that when the hazards of underlying survival distributions are nonproportional early on, say but proportional when , the Wilcoxon test has more power than log rank test. We use the simulation results to justify which test should be used in a transplantation study of leukemia cancer patients, and show that overall survivals of patients with two graft sources are statistically significantly different partially because of the delay of neutrophil recovery of one patient population immediately after transplantation.

e-mail: pxf16@po.cwru.edu

CHANGEPOINT DIAGNOSTICS FOR COMPETING RISKS MODELS: A CONDITIONAL POSTERIOR PREDICTIVE P-VALUE APPROACH

Chen-Pin Wang*, University of South Florida
Malay Ghosh, University of Florida

Competing risks analysis (David and Moeschberg, 1978) deals with time-to-event data while accounting for the causes that jointly operate the event occurrence. We study two families of changepoint competing risks models (CCRM's) that are built upon the cause-specific model proposed by Chiang (1968), and changepoints are further added to accommodate possible distributional heterogeneity (mixtures). Fit of each CCRM is quantified by two types of model discrepancy indices, measures of cumulative model departure between each pair of adjacent changepoints, with respect to posterior predicted distribution. Indeed, each of these indices correponds to model deficiency due to missing changepoints associated with either the overall survival distribution, or cause-specific probabilities. We show that asymptotically, the conditional posterior predictive p-values associated with these discrepancy indices attain their greatest departure from 0.5 at the changepoints that are missing from the assumed model. Drawing from this asymptotic property, a diagnostic procedure is proposed to guide the search of changepoints in CCRM.

e-mail: cwang@hsc.usf.edu

RANDOM EFFECTS MODELS FOR REPEATED MEASURES OF ZERO-INFLATED COUNT DATA

Yongyi Min*, Department of Statistics, University of Florida
Alan Agresti, Department of Statistics, University of Florida

For count responses, the situation of excess zeros (relative to what standard models allow) often occurs in many biomedical and sociological applications. Modeling repeated measures of zero-inflated count data presents special challenges. This is because in addition to the problem of extra zeros, the correlation between measurements upon the same subject at different occasions needs to be taken into account. This article presents two types of random effects models for repeated measurements on this type of response variable. The first model is a hurdle model with random effects, which separately handles the zero observations and the positive counts. In maximum likelihood model fitting, we consider both a normal distribution and a nonparametric approach for the random effects. The second model is a cumulative logit model with random effects, which has the simplicity of using a single model to handle the zero-inflation problem. We motivate and illustrate the proposed methods with an example from an occupational injury prevention program.

e-mail: ymin@stat.ufl.edu

HIERARCHICAL MODELS FOR REPEATED BINARY DATA USING IBF SAMPLER

Ming TAN, University of Maryland Greenebaum Cancer Center
Guoliang TIAN*, University of Maryland Greenebaum Cancer Center

Hierarchical models have emerged as a promising tool for the analysis of repeated binary data. However, the computational complexity in these models have limited their applications in practice. Several approaches have been proposed in the literature to overcome the computational difficulties including MLE from a frequentist perspective and MCMC from a Bayesian perspective. Although MCMC methods provide the whole posterior of the parameter of interest, the mixing rate and convergence problems of the Markov chain are still unresolved. This article develops a noniterative sampling approach, the inverse Bayes formulae (IBF) sampler, for the hierarchical model for repeated binary data for which current methods encounter difficulty. The proposed method obtains iid samples approximately from the observed posterior and thus eliminates the convergence problem associated with the MCMC methods. An efficient IBF sampling is implemented by utilizing the estimated posterior modes obtained via Monte Carlo EM algorithm. Real datasets from six cities children's wheeze study and children's ear fluid study illustrate the proposed methods.

e-mail: gtian2@umm.edu

DYNAMIC LINEAR MODEL OF MORTALITY DATA

Amitabha Bhaumik*, University of Connecticut
Dipak K.Dey, University of Connecticut
Nalini Ravishanker, University of Connecticut

Compositional time series data comprises of multivariate observations that at each time point are essentially proportions of a whole quantity. This kind of data occurs frequently in many disciplines such as economics, geology and ecology. Usual multivariate statistical procedures available in the literature are not applicable for the analysis of such data because of constrained nature of these observations. This article describes new techniques for modeling compositional time series data in a hierarchical Bayesian framework. Modified dynamic linear models are fit to compositional data via Markov chain Monte Carlo techniques. The distribution of the underlying errors is assumed to be a scale mixture of multivariate normals of which the multivariate normal, multivariate-t, multivariatelogistic, etc., are special cases. In particular, multivariate normal and Student-t error structures are considered and compared through predictive distributions. The approach is illustrated on mortality data collected from Los Angeles between 1970 and 1979. The data consists of a trivariate time series of percentages of daily deaths in Los Angeles county for 1970 to 1979.

e-mail: amitabhabhaumik@hotmail.com

BAYESIAN APPROACHES TO MEASURING AND MAPPING SMALL AREA DISPARITIES IN HEART DISEASE MORTALITY IN SOUTH CAROLINA

Eric C. Tassone*, Emory University
Lance A.Waller, Emory University
Michele Casper, CDC
Ishmael Williams, CDC
Kurt Greenlund, CDC
Linda Neff, CDC

We measure county-level heart disease mortality disparity between African Americans and whites in South Carolina during 1991-1995, for men and women separately. We extend the typical hierarchical Bayesian disease mapping model to measure disparity by exploiting the race-sex group structure of the study population. We offer two disparity measures, both of which extend the typical disease mapping model, though in different ways. We obtain similar results under the two disparity measures for both men and women. Point estimates of disparity are comparable between the two approaches, though associated variability shows some differences. We discuss the relative merits of the two disparity measures, and detail possible enhancements of the measures.

e-mail: etasson@sph.emory.edu

ASCENT BASED MONTE CARLO EM

Brian S Caffo*, Dept of Biostatistics, Johns Hopkins University
Wolfgang S.Jank, RH Smith School of Business, University of Maryland
Galin L. Jones, Department of Statistics, University of Minnesota

The EM algorithm is a popular tool for iteratively maximizing likelihoods in the presence of missing data. One of the reasons for its popularity is the so-called ascent property. That is, parameters output by EM are guaranteed to have increasing likelihood values with each iteration. Unfortunately, in many practical problems EM requires the evaluation of analytically intractable and high-dimensional integrals. The Monte Carlo EM (MCEM) is the natural extension of EM that attempts to solve this problem by estimating the relevant integrals using Monte Carlo methods. Though MCEM often possesses many of the desirable numerical properties of EM, its naive application does not inherit the ascent property, or even eventual convergence of the parameter estimates to a maximum or saddle point. The latter deficiency can be solved by increasing the Monte Carlo sample size with each MCEM iteration. However, such a fix does not succeed in recovering the ascent property. In this talk we propose a data-driven automated MCEM algorithm that recovers the ascent property with high probability. We illustrate these techniques in several examples.

e-mail: bcaffo@jhsph.edu

BAYESIAN INFERENCES ON SHAPE CONSTRAINED HORMONE TRAJECTORIES IN THE MENSTRUAL CYCLE

Laura H. Gunn*, Duke University, Institute of Statistics and Decision Sciences
David B.Dunson, Biostatistics Branch, National Institute of Environmental Health Sciences

In studies of hormone patterns in the menstrual cycle, it is often reasonable to assume that the mean trajectory increases monotonically to an unknown peak and decreases thereafter. To account for the dependency in hormone measurements, one can apply a hierarchical model having women-specific random effects and autocorrelated errors. In the unconstrained case, Bayesian computation can proceed via a Gibbs sampling algorithm. Unfortunately, standard Gibbs sampling approaches for incorporating parameter constraints cannot be used when the constraints are on higher level parameters in the hierarchy. To solve this problem, this article proposes a transformation approach in which samples from the unconstrained posterior density for the cycle-specific trajectories are transformed to the restricted space using a minimal distance mapping. This approach is shown to result in substantially improved efficiency relative to unconstrained analyses and analyses that place constraints on the population parameters. The methods are illustrated through an application to progesterone data from the literature.

e-mail: laura@stat.duke.edu

ON CLUSTERING METHODS FOR EXPLORING HSC GENE EXPRESSION DATA

Jie Chen*, University of Missouri-Kansas City
Xi He, Stowers Insititute for Medical Research
Linheng Li, Stowers Insititute for Medical Research

Clustering algorithms have been used to analyze microarray gene expression data in many recent applications. In this paper, we make a comparison among popularly used clustering methods, including hierarchical clustering with average, complete, and single linkages, k-means clustering, k-means
clustering with hierarchical initialization, and self organization map (SOM), by making use of our hemotopietic stem cell (HSC) microarray data. To understand the biological pathways from HSC to proliferative multipotent progenitor (MPP), and from MPP to either common lymphoid progenitor (CLP) or common myeloid progenitor (CMP), statistical clustering is an important tool. Our results demonstrated that the HSC microarray data set casts some challenge on clustering algorithms as different clustering algorithms resulted in clusters that were not all consistent. We compared the results by using the total within-cluster sum of squares of dispersions, the cluster sizes, and the biological functions of the genes, and reached the conclusion that k-means clustering with hierarchical average or complete linkage initialization performed the best among all the methods we compared. Our review and investigation of the clustering methods with HSC microarray data provide a useful approach to medical researchers who use clustering algorithms in analyzing their microarray or related data sets.

e-mail: chenj@umkc.edu

EXACT POWER UNDER INDEPENDENCE FOR THE FALSE DISCOVERY RATE IN GENE EXPRESSION ARRAY EXPERIMENTS

Deborah H. Glueck*, Dept.of Preventive Medicine and Biometrics, Univ. of Colorado Health Sciences Ctr
Keith E.Muller, Department of Biostatistics, University of North Carolina
Larry Hunter, Department of Pharmacology, University of Colorado Health Sciences Center

The false discovery rate is widely used for multiple comparison problems, including gene expression array studies. In that case, choosing the number of chips is an important but poorly treated problem. We have been unable to find published reports of analytic expressions for power and sample size for the false discovery rate. We derive the distribution of the total number of rejections, the number of false rejections conditioning on the total number of rejections, and the joint probability distribution of the number of total and false rejections. We demonstrate the results for independent but not identically distributed arbitrary distributions, and then describe the special case of independent and identically distributed distributions. Thus we provide methods for exact small sample power and sample size for gene expression array experiments, based on the common assumption of independence. The results also apply to any other use of false discovery rate which meets the same assumptions. Simulation studies are used to confirm the analytic results. An example is given for research on the genetic basis of breast cancer.

e-mail: Deborah.Glueck@uchsc.edu

MANTEL STATISTICS TO CORRELATE GENE EXPRESSION LEVELS FROM MICROARRAYS WITH CLINICAL COVARIATES

Bill Shannon*, Washington Univ. in St. Louis School of Medicine

Mantel statistics provide an additional step to standard approaches in the analysis of gene expression and covariate data. They allow the calculation of standard statistics (e.g., correlation, partial correlation, regression coefficients) and permutation P values to relate the sample covariates to the expression levels. In this talk we describe the Mantel statistics and illustrate their use and interpretation with data from a study of seven human oligodendrogliomas (brain tumors) where expression levels of 1,013 genes and five covariates were previously analyzed using standard approaches. In the previous analysis of this data, qualitative relationships were found between gene expressions and two of the clinical covariates. We show in this talkhow the Mantel statistics are able to formally quantify and provide P values to determine statistical significance of these relationships. We also show how the Mantel statistics can be used to rank subsets of genes, found using standard clustering methods, in terms of differential expression across samples.

e-mail: shannon@ilya.wustl.edu

BAYESIAN MIXTURE MODEL FOR DIFFERENTIAL GENE EXPRESSION

Feng Tang*, Biostatistics Department, M.D. Anderson Cancer center
Peter Muller,
Kim-Ahn Do,

We propose model-based inference for differential gene expression, using a non-parametric Bayesian probability model for the distribution of gene intensities under different conditions. The probability model is essentially a mixture of normals. The resulting inference is similar to a recent empirical Bayes method proposed by Efron et al. (2001). The use of full model-based inference mitigates some of the necessary limitations of the ad-hoc empirical Bayes method. However, the increased generality of our method comes at a price. Computation is not as straightforward as in the empirical Bayes scheme; but we argue that inference is nothing more difficult than posterior simulation in the traditional non-parametric mixture of normal models. We illustrate the applicability of our proposed method through (i) a simulation study and (ii) to the analysis of a colon data set with the aim of identifying genes that differentially express for tumor versus normal tissue samples. The full model-based Bayesian method allows us to make joint inferences about a sub group of genes being differentially expressed. We illustrate how it handles multiplicity automatically.

e-mail: ftang@odin.mdacc.tmc.edu

BADGE: BAYESIAN DIFFERENTIAL ANALYSIS OF GENE EXPRESSION DATA

Paola Sebastiani*, Department of Mathematics and Statistics, University of Massachusetts
Marco Ramoni, Children's Hospital Informatics Program, Harvard Medical School

BADGE is a novel method for the differential analysis of gene expression data measured with Affymetrix microarray technology. BADGE models gene expression data by log-normal and gamma distributions and uses model averaging techniques to compute the posterior probability of differential expression of each gene. Features of BADGE are the ability to account for the within and between chip variability without need for arbitrary normalization of the data or filters. Furthermore, the model-based approach of BADGE can be used to describe a probabilistic molecular profile, which is the basis for the development of diagnostic models. Experimental evaluations on real data sets show that BADGE is able to detect differentially expressed genes in relatively small samples, and with a higher reproducibility and accuracy than non-parametric methods implemented in SAM or GeneCluster.

e-mail: sebas@math.umass.edu

LEAST SQUARES AND MIXED EFFECTS MODELS TO ADJUST FOR BATCH EFFECTS IN SINGLE-CHANNEL CDNA MICROARRAYS

Mary E. Putt*, Department of Biostatistics and Clinical Epidemiology,University of Pennsylvania,
Lan Zhou, Department of Biostatistics and Clinical Epidemiology, University of Pennsylvania,
Heping Hu, Department of Biostatistics and Clinical Epidemiology, University of Pennsylvania,
Tom TenHave, Department of Biostatistics and Clinical Epidemiology, University of Pennsylvania,

We were motivated by a study to distinguish gene expression differences between controls and patients with lymphoma. However, the microarrays were generated in batches in the laboratory, and the case and batch effects were confounded. Using least squares (LS) we considered each gene separately and estimated its adjusted case effect. Using mixed effects (ME) we fit a comparable model, that adjusted for confounding by batch at the population mean level as well as random effects due to batch. Case entered the model as a random effect. We are interested in genes with large BLUPS for the case effect. Using a permutation-based estimate, we found false discovery rates of 0.00 (95% CI 0.04) for LS and 0.04 (95% CI 0.16) for a BLUP-based estimate for the 25 most significant genes. Nine out of 25 (36%) genes were common to both methods. LS and ME approaches offer an alternative to global normalization that is easily implemented using standard statistical software and minimal assumptions. The approach may be useful in analyzing data from observational studies with other array-specific covariates and for combining data from more than one study.

e-mail: mputt@cceb.upenn.edu

EXPLORATORY SURVIVAL ANALYSIS WITH APPLICATIONS TO GENE EXPRESSION MICROARRAY DATA

Cheng Cheng*, Department of Biostatistics, St. Jude Children's Research Hospital

This paper will address the issues, challenges, and methodologies for exploring the association between a vast number of potential factors and the risk of an adverse event (e.g., relapse of primary cancer, development of a secondary maligancy) after the treatment of the primary cancer. With microarray gene expression data, the factors are gene expressions. In this setting the 'phenotype' is represented by the time to the adverse event which can be censored at the last follow up. A gene ranking and screening method, called weighted rank regression association (WRRA), and a method of permutation assessment called Monte-Carlo exact profile significance (MCEPS) will be presented.

e-mail: cheng.cheng@stjude.org

74. Genome-Scale Biology: Statistical Challenges and Solutions

PREDICTION OF SURVIVAL WITH GENE EXPRESSION

Mark J. van der Laan*, UC Berkeley, Biostatistics and Statistics

We provide methods for predicting survival based on high dimensional covariate vectors such as gene expression profiles. We propose a new cross-validation criterian using double robust locally efficient estimation of risk of a predictor based on the validation sample. We prove that the corresponding feature/covariate selection method is consistent.

e-mail: laan@stat.berkeley.edu

EXPLORING INTERACTIONS IN GENOMIC DATA

Ingo Ruczinski*, Johns Hopkins University

Exploring higher order interactions is a statistical challenge frequently occuring in the analysis of genomic data. For example, interactions may be present between single nucleotide polymorphisms or chromosomal deletions when considering associations with disease or disease stages, respectively. We discuss some of the currently available techniques to search for these interactions, how to make statistical inference, and show some examples.

e-mail: ingo@jhu.edu

RESOLVING GENE EXPRESSION PROFILES WITH TAG-BASED TECHNOLOGIES

Jennifer F. Bryan*, University of British Columbia
Biotechnology Laboratory and Dept of Statistics
Genome BC Array Facility

We will present methods for estimating gene expression profiles from a universal array, tag-based profiling technique that is under development. This technique has features in common with Serial Analysis of Gene Expression (SAGE) and with microarrays, although it is distinct from both. Issues that have implications for estimation and inference include tag degeneracy (when multiple transcripts cannot be distinguished from one another) and various substantial sources of noise and bias in sample preparation (for example, PCR bias) that affect both new and current expression profiling techniques. This work is part of a collaboration with Dr. Charles Haynes, in the Biotechnology Laboratory and Chemical & Biological Engineering Dept. at UBC.

e-mail: jenny@stat.ubc.ca

BACKGROUND, CONCEPT, AND REGULATORY CONSIDERATIONS FOR MULTI-STAGE DESIGNS IN THE IMPLEMENTATION OF THE ACCELERATED APPROVAL MECHANISM

George Y.H. Chi*, Office of Biostatistics, Center for Drug Evaluation, Food and Drug Administration
Pei L.Yang, Office of Biostatistics, Center for Drug Evaluation, Food and Drug Administration
Rajeshwari Sridhara, Office of Biostatistics, Center for Drug Evaluation, Food and Drug Administration
Gang Chen, Office of Biostatistics, Center for Drug Evaluation, Food and Drug Administration

For serious and life-threatening diseases, it is crucial to make available to the patients as soon as possible promising drugs when there are no effective treatments on the market. In recognition of this fact, FDA has put in place a mechanism for potentially life-saving drugs to gain accelerated approval. This talk will describe some of the recent experiences in accelerated approval and discuss some of the issues that would highlight certain drawbacks inherent in the current mechanism as implemented. The concept of a multi-stage design lends itself naturally to the design and implementation of accelerated approval. A proposal for an accelerated approval decision process is advanced that may alleviate this problem inherent in its
implementation.

e-mail: chig@cder.fda.gov

ON EFFECTIVENESS OF PHASE 2/3 COMBINATION DESIGNS FOR CLINICAL DEVELOPMENT

Qing Liu*, J&J Pharmaceutical Research and Development
Gordon Pledger, J&J Pharmaceutical Research and Development

In conventional drug development, phase 2 trials are conducted prior to initiation of large-scale phase 3 trials. The phase 2 results are used to determine if larger phase 3 trials are warranted. In situations where phase 3 trials are to proceed, the phase 2 results are incorporated in phase 3 trial designs. The conventional approach has the advantage of limiting the cost if it turns out the null hypothesis is true, but has the disadvantage of requiring a longer time and more resources if in fact the alternative hypothesis is true. As a result, in recent drug development, sponsors took the risky approach of proceeding to larger phase 3 trials without conducting phase 2 trials first. We propose a phase 2/3 combination design for clinical development. The approach has the advantage of limiting the cost under the null hypothesis and the advantage of maximizing the expected gain under the alternative hypothesis.

Key words:
Adaptive design; Benefit-risk analysis; Conditional error; Interim analysis; Type I error rate

e-mail: qliu2@prdus.jnj.com

ESTIMATION OF PARAMETER AND ITS EXACT CONFIDENCE INTERVAL FOLLOWING SEQUENTIAL SAMPLE SIZE RE-ESTIMATION TRIALS

Yu Shen*, M.D. Anderson Cancer Center
Yi Cheng, Indiana University at South Bend

We develop a new procedure for estimation following a sample size re-estimation design. The method for obtaining an exact confidence interval and point estimate is based on a general distribution property of the final test statistic of the Self-designing group sequential clinical trial by Shen and Fisher (1999). A more elaborate estimate is proposed to explicitly account for futility stopping boundary with reduced bias. We also provide a modified weight function to improve the power of the test. Extensive simulation studies show that the exact confidence intervals have accurate nominal probability of coverage, and the proposed point estimates work well with practical sample sizes.

e-mail: yshen@mdanderson.org

STATISTICAL AND PRACTICAL ASPECTS OF A NON-STOP DRUG DEVELOPMENT STRATEGY

Ronald W. Helms, Rho, Inc.
Karen L.Kesler*, Rho, Inc.

The traditional strategy for drug development Phases II through III is to plan a series of studies, each of which has a small number of relatively specific objectives. Recent advances in statistical methods may lead to more time efficient strategies. By designing the study with a wide range of treatment arms, we can prune unproductive arms, intensifing the focus on more promising arms. This novel approach is especially useful in Phase II trials, where the goal is to gather the efficacy and safety information to proceed “non-stop” to the confirmatory trials. Although this approach to study design holds the potential for more time and cost efficient trials, there are many logistic and statistical challenges that remain to be addressed. We will address several of these issues including: how many treatment arms to implement, how to select a primary outcome for the Phase III study, the role of centralized randomization, and efficient data management. Additionally, we will discuss regulatory issues including unblinding and FDA acceptance of this novel approach. Finally, we will discuss the shift from the treatment selecting paradigm in Phase II to the more rigid classical hypothesis testing design of Phase III.

e-mail: kkesler@rhoworld.com

ON THE USE OF THE COMPETING RISKS PROPORTIONAL HAZARDS MODEL IN THE CHOICE OF OPTIMAL TREATMENT

Richard Kay*, PAREXEL International, Navigation House, Sheffield

This presentation will detail the application of the Competing Risks Proportional Hazards Model to a clinical assessing the effects of diethylstilbestrol (DES) on survival time in stage 3 and 4 prostate cancer (Kay (1986)). These same data had previously been analysed using a standard proportional hazards approach incorporating complex treatment x covariate interactions in order to provide a model which investigates the choice of optimal treatment for specific patients (Green and Byar (1980)). Recognising the competing risks nature of the problem however and directly modelling that structure is clearly seen to provide a more direct answer to the question of optimal treatment and also greater modelling flexibility.

References:

Green, S. B. and Byar, D. P. (1980) The choice of treatment for cancer patients based on covariate information: Application to prostate cancer. Bulletin Cancer, Paris, 67, 477-488

Kay, R. (1986) Treatment effects in competing risks analysis of prostate cancer data. Biometrics, 42, 203-211

e-mail: richard.kay@parexel.com

CUMULATIVE INCIDENCE REGRESSION

Jason P. Fine*, Departments of Statistics and of Biostatistics &
Medical Informatics, University of Wisconsin-Madison

With covariates, the standard analysis for competing risks data models cause-specific hazards under a proportional hazards assumption. However, the cause-specific hazard is not directly interpretatable in terms of survival probabilities for particular failure types. Recently, clinicians have begun using the cumulative incidence function, the failure probabilities for a particular cause, which is intuitively appealing and easily explained to non-statisticians. One approach is to combine regression models for the cause-specific hazards to estimate the cumulative incidence. Such methods do not allow a direct assessment of the covariate effect on the cause specific failure probabilities. In this talk, we discuss an alternative modelling strategy in which the cumulative incidence is modelled directly, as one ordinarily models the survival function. Semiparametric estimation procedures are developed for finite-dimensional covariate effects and the baseline failure probabilities. Inferences for regression effects and for the cumulative incidence with certain covariates are provided. Breast cancer data is used to contrast the two frameworks.

e-mail: fine@biostat.wisc.edu

BOUNDS ON JOINT SURVIVAL PROBABILITIES WITH POSITIVELY DEPENDENT COMPETING RISKS

Kalyan Ghosh*, Merck & Co.
Sanat Sarkar, Temple University

In many competing risks problems, assuming positive dependence among the risks is often a realistic approach. In this talk we present upper and lower bounds on the potential joint survival probabilities based on the observable quantities when the potential failure times corresponding to the different risks are assumed to be positively dependent. The data from Nair (1993) will be used to illustrate how one could guess the potential joint as well as marginal survival probabilities in terms of empirical upper and lower bounds obtained by using appropriate nonparametric estimates of the corresponding theoretical bounds obtained here.

e-mail: ghoshk@merck.com

77. Statistical Aspects of Polygraphy

LIES, DAMNED LIES: THE STATISTICS OF NATIONAL SECURITY POLYGRAPH SCREENING

Stephen E. Fienberg, Dept. of Statistics, Carnegie-Mellon University
Patricia L.Grambsch, Dept. of Biostatistics, Univ. of Minnesota
Peter B. Imrey*, Dept. of Biostatistics & Epidemiology, The Cleveland Clinic Foundation
Aleksandra Slavkovic, Dept. of Statistics, Carnegie-Mellon University
For The Committee to Review the Scientific Evidence on the Polygraph, National Research Council

In 1999, concerned about security violations at Department of Energy (DoE) weapons laboratories, Congress required polygraph screening examinations of over 1,300 DoE employees involved in sensitive work. However, polygraph examinations are rarely admissible as evidence in court because of concerns about their accuracy and Congress, by the Employee Polygraph Protection Act of 1988, had previously prohibited polygraph screening of job applicants and current personnel in the private sector. In January 2001 the National Research Council's Committee on National Statistics convened a multidisciplinary committee, funded by DoE, charged to review relevant research and assess the validity of polygraph examinations for personnel security screening. The resulting report, 'The Polygraph and Lie Detection,' reviews the construct validity of polygraph testing from a psychophysiological perspective, and its criterion validity from an epidemiological viewpoint using receiver operating characteristics (ROCs) extracted from a structured literature review. We review issues of bias in the empirical literature on polygraph testing, statistical issues in summarizing the strength of polygraphy as a diagnostic technology, and the implications of existing data for national securiy polygraph screening.

e-mail: pimrey@bio.ri.ccf.org

META-ANALYSIS OF COARSELY CLASSIFIED DIAGNOSTIC DATA ON POLYGRAPH TESTING

Xiao-Hua Andrew Zhou, VA Puget Sound Health Care System and University of Washington

The scientific validity of polygraph examinations to detect deception is still controversial almost 100 years since their inception. In 2001-2, the National Research Councils Committee to Review the Scientific Evidence on the Polygraph extracted criterion validity data on polygraph testing from a structured literature review. One of limitations in the data is that it is reported in a coarse classification of polygraph test scores, typically in two categories such as significant response or no significant response, or these with a third, indeterminate, category. In this talk, I will discuss statistical issues regarding the meta-analysis of coarsely classified diagnostic data and illustrate some current meta-analytic techniques that may be used to analyze the polygraph data. I will also discuss potentially inherent biases in the polygraph data, and their implications for meta-analytic results.

e-mail:

78. Semi-and Nonparametric Methods for Longitudinal Data

KERNEL ESTIMATION IN LONGITUDINAL DATA WITH OUTCOME-RELATED OBSERVATION TIMES

Donglin Zeng*, Department of Biostatistics, University of North Carolina
Daohai Yu, Department of Biostatistics & Bioinformatics, Duke University

Kernel estimation is an important nonparametric approach and is often employed in a longitudinal study to characterize the time-pattern of a longitudinal response when no parametric models are appropriate for the data. When time points at which the response is observed are random and correlated with the longitudinal response, however, the classic kernel estimation is biased. The magnitude of such bias can be substantial, as shown under a joint model we considered for the longitudinal process and observation time process. Furthermore, we propose a bias-corrected kernel estimator under this joint model. The asymptotic distribution of the proposed estimator is provided. Numerical studies give evidence that our bias-corrected kernel estimator is superior to its classic counterpart in terms of both the coverage probability and mean squared error. Finally, our approach is illustrated through an application to an CD4 cell count analysis in a recent HIV study.

e-mail: dzeng@bios.unc.edu

SEMI-PARAMETRIC LONGITUDINAL MODELS SUBJECT TO MISSING DATA BY DESIGN

Jaroslaw Harezlak*, Department of Biostatistics, Harvard School of Public Health
Louise Ryan, Department of Biostatistics, Harvard School of Public Health
Nicholas Lange, Departments of Psychiatry and Biostatistics,
Harvard University Schools of Medicine and Public Health

We discuss cost effective study designs for settings where interest lies in characterizing growth rates over time, but where budgetary and practical constraints prohibit collection of data for the whole cohort at each time point of interest. A practical alternative is to recruit subjects at a variety of different timepoints and to observe each one over a shorter period of time. We show that so long as the combined set of Observation times spans the period of interest, non-parametric smoothing methods can be used to provide a consistent estimate of the mean population curve over the entire span of the cohort age range. While there are a number of options available for non-parametric curve estimation, we find tha penalized regression splines work well. The approach also facilitates estimation of confidence bands, along with estimation of the speed of growth. Furthermore, the penalized spline models can be easily fit using standard mixed models software (e.g. PROC MIXED in SAS). We describe simulation studies as well as application of our methods to a study of pediatric brain imaging, where each child has 3-5 MRI's during a short period of time (1-3 years).

e-mail: jharezla@hsph.harvard.edu

A SEMIPARAMETRIC LIKELIHOOD APPROACH TO JOINT MODELING OF LONGITUDINAL AND TIME-TO-EVENT DATA

Xiao Song*, University of Washington
Marie Davidian, North Carolina State University
Anastasios A. Tsiatis, North Carolina State University

Joint models for a time-to-event (e.g. survival) and a longitudinal response have generated considerable recent interest. The longitudinal data are assumed to follow a mixed effects model, and a proportional hazards model depending on the longitudinal random effects and other covariates is assumed for the survival endpoint. Interest may focus on inference on the longitudinal data process, which is informatively censored, or on the hazard relationship. Several methods for fitting such models have been proposed, most requiring a parametric distributional assumption (normality) on the random effects. A natural concern is sensitivity to violation of this assumption; moreover, a restrictive distributional assumption may obscure key features in the data. We investigate these issues through our proposal of a likelihood-based approach that requires only the assumption that the random effects have a "smooth" density. Implementation via the EM algorithm is described, and performance and the benefits for uncovering noteworthy features are illustrated by application to data from an HIV clinical trial and by simulation.

e-mail: songx@u.washington.edu

LOCAL POLYNOMIAL REGRESSION ANALYSIS OF CLUSTERED DATA

Kani Chen, Department of Mathematics, HKUST, Hong Kong
Zhezhen Jin*, Dept. of Biostatistics, Mailman School of Public Health, Columbia University

We present a nonparametric regression analysis method for clustered data by appropriately combines the methodologies of generalized estimating equations and local polynomial smoothing. The method issuperior to the existing generalized estimating equation type of curve estimation methods in its conceptual clarity, computational simplicity, efficiency and ease in potential application and theory. We illustrate the method by simulation and real examples.

e-mail: zjin@biostat.columbia.edu

BACKFITTING AND LOCAL LIKELIHOOD METHODS FOR NONPARAMETRIC MIXED-EFFECTS MODELS WITH LOGITUDINAL DATA

Jeong-gun Park*, Frontier Science & Technology Research Foundation
Hulin Wu, Frontier Science & Technology Research Foundation

We consider a nonparameteric mixed-effects model,$y_{i}(t_{ij})=\eta(t_{ij})+\upsilon_{i}(t_{ij})+\varepsilon_{i}(t_{ij}), j=1,...,n_{i}, i=1,...,m,$ for longitudinal data. The approach of combining a local polynomial kernel regression and a standard linear mixed-effects model techniques is used to estimate the parameters $\eta(t_{ij})$ and $\upsilon_{i}(t_{ij})$. We propose a local likelihood approach with a backfitting algorithm to estimate the population mean function, $\eta(t_{ij})$ which describes the overall changes over time, as our primary goal. The asymptotic properties are investigated and a simulation study compares the performance of our proposed method to two other approaches, local polynomial GEE method by Lin and Carroll (2000) and local polynomial LME approach by Wu and Zhang (2002) in terms of mean squared errors. Real application is also presented to illustrate our method for longitudinal data.

e-mail: park@fstrf.dfci.harvard.edu

SEMIPARAMETRIC ESTIMATION IN TRANSITION MEASUREMENT ERROR MODELS

Wenqin Pan*, Rho, Inc.
Donglin Zeng, University of North Carolina
Xihong Lin, University of Michigan

We propose a new class of models, transition measurement error models, for analyzing longitudinal data with covariate measurement error. Three approaches are considered to reduce the estimating bias due to measurement error and achieve valid inferences of parameter estimators. They are MLE approach, simulation extrapolation (SIMEX) approach, and semiparametric estimation approach. While the unbiasedness of the former two approaches both depend on some unobserved information, semiparametric approach can give consistent parameter estimators without any distribution assumption of the error-prone covariate. We first generalize a traditional conditional score approach to transition measurement error model, and apply the approach on both linear and logistic transition models to achieve consistent estimators. For the linear case, we then discuss the loss of efficiency in the conditional score approach. Finally, under the condition that a small set of validation data is available, we propose a one-step estimation approach to achieve semiparametric efficient estimators. Numerical calculation and simulation studies are performed to evaluate and illustrate the proposed approaches.

e-mail: wpan@rhoworld.com

A TECHNIQUE FOR IMPLEMENTING FIXED-KNOT REGRESSION SPLINES IN THE GENERAL LINEAR MIXED MODEL

Lloyd J. Edwards*, Department of Biostatistics, University of North Carolina at Chapel Hill
Paul W.Stewart, Department of Biostatistics, University of North Carolina at Chapel Hill
James E. MacDougall, Genzyme
Ronald W. Helms, Rho, Inc.

The use of regression or smoothing splines in the mixed model is applicable in many situations for the analysis of longitudinal data. Due to recent developments in fitting splines in mixed models, the data analyst has available a choice of techniques that offer different approaches depending on the properties of the data and scientific questions of interest. We propose a technique for implementing fixed-knot regression splines in the mixed model that has several attractive features and taken together, the features add up to a very flexible technique. The technique can be easily programmed using existing commercial software such as SAS Proc Mixed or Splus. A mixed model with fixed-knot regression splines for the fixed and/or random effects directly corresponds to a mixed model with linear equality constraints. By utilizing a reparameterization of the explicitly constrained mixed model, an implicitly constrained mixed model can be constructed that simplifies implementation of fixed-knot regression splines using non-iterative computations. The technique is illustrated by analyzing longitudinal viral load data from a study of subjects with acute HIV infection.

e-mail: Lloyd_Edwards@unc.edu

A COMPARISON OF THE DORFMAN-BERBAUM-METZ AND OBUCHOWSKI METHODS FOR RECEIVER OPERATING CHARACTERISTIC (ROC) DATA

Stephen L. Hillis*, Iowa City VA Medical Center

Several methods are currently used for analyzing multi-reader ROC studies, with the Dorfman-Berbaum-Metz (DBM) method used most frequently. The 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 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. I compare the two methods and show that they will usually give similar results.

e-mail: steve-hillis@uiowa.edu

PROPERTIES OF THE PARTIAL AREA UNDER THE CURVE (AUC) FOR SUMMARY ROC CURVES

Stephen D. Walter*, Clinical Epidemiology and Biostatistics, McMaster University

The area under the Summary ROC curve has been proposed as a summary measure to describe the performance of a diagnostic test in the context of meta-analysis. However, there may often be limited or no data in the region of low specificity for the test. The partial AUC has been suggested as an alternative summary measure that would reflect only clinically meaningful values of specificity.

This paper will present the results of a numerical investigation of the behaviour of the full and partial AUCs in data where the diagnostic odds ratio is homogeneous or heterogeneous. While the full AUC is robust to heterogeneity, the partial AUC and its standard error are strongly dependent on the amount of heterogeneity, and to the specification of the region of acceptable test specificity.

While the partial AUC may have some intuitive clinical advantages, its statistical properties imply caution is required in its use. This is particularly true if two or more diagnostic tests are being compared, and if their range of observed specificity varies.

e-mail: walter@mcmaster.ca

ASSESSING THE EFFICIENCY OF CT IMAGING FOR DETECTING AND CLASSIFYING LIVER LESIONS IN CANDIDATES FOR HEPATIC SURGERY

Richard E. Thompson*, Dept. of Biostatistics, Johns Hopkins Bloomberg School of Public Health
Ihab R.Kamel, Dept. of Radiology, Johns Hopkins School of Medicine

Receiver-operating curve (ROC) analysis is typically used to evaluate clinical tests that classify patients as 'sick' or 'well' against some gold standard. For tests assessing cancerous lesions, standard ROC analysis is not appropriate since the outcome is not binary. Typically, the radiologist must identify a lesion and correctly classify it as being benign or malignant with some confidence score. An alternative-free receiver-operating curve (AFROC) analysis can be used to evaluate data of this type. Adapting AFROC methods, we evaluated liver CT data on 77 patients with a total of 165 malignant and 71 benign lesions seen by pathology. Results from three independent readers were used. A true positive 'hit' was scored if the reader correctly identified and correctly classified a 'true' lesion. Sensitivity was calculated as the number of hits out of the total number of lesions. A false positive was scored if a benign lesion was classified as malignant, or if a reader identified a non-lesion object as being a lesion. Specificity was evaluated as the number of scans with no false positives out of the total number of scans evaluated. Area under the curve can then be calculated using ROC plots and corresponding variances assessed using boot-strapping methods.

e-mail: rthompso@jhsph.edu

DESIGN AND ANALYSIS OF COMPARATIVE DIAGNOSTIC ACCURACY STUDIES WITH MULTIPLE CORRELATED TEST RESULTS

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 (NICHD), National Institutes of Health
Madhu Mazumdar, Sloan-Kettering Cancer Center

Receiver operating characteristic (ROC) curves have been widely used to assess the comparative accuracy of diagnostic tests, with larger area under the curve indicating higher accuracy for that test. In diagnostic studies of advanced stage cancer, it is very common for the test to identify multiple tumors that come from the same patient. Analysis of an ROC curve needs to take into account the correlation between these multiple test outcomes. In this paper, we consider the design and analysis issues of such studies. We derive point and interval estimation for the area under the ROC curves. For testing the hypothesis of equal areas, we construct test statistics and derive power and sample size formulas. The methods are illustrated using an example of comparison of CT and PET scanner for detecting extra-hepatic disease for colorectal cancer.

e-mail: liua@mail.nih.gov

DECODING POOLS OF CHEMICAL COMPOUNDS IN THE PRESENCE OF COMPOUND INTERACTION AND DILUTION

Bingming Yi*, North Carolina State University
Jacqueline M.Hughes-Oliver, North Carolina State University
Stanley S. Young, GalxoSmithKline

During High Throughput Screening, large collections of chemical compounds are tested for potency with respect to one or more assays. In reality, only a very small fraction of the compounds in a collection will be potent enough to act as lead molecules in later drug discovery phases. Testing all compounds is neither cost-effective nor desirable. Pooling experiments have emerged as an attractive method for optimizing discovery of lead molecules, and many pharmaceutical companies are now engaged in designing and analyzing data on pooled chemical compounds. However, the existence of multiple potency mechanisms, compound-to-compound interactions within pools, and complex dependency relationships between pool and individual potencies all contribute to complicating the analysis of pooling experiments. This work proposes a modeling strategy that addresses many of the issues raised in pooling. Application is successfully made to real data originating from chemical libraries within GlaxoSmithKline, where hit rates are used to measure effectiveness of the screening process.

e-mail: byi@unity.ncsu.edu

80. Causal Inference and Missing Data

SEMIPARAMETRIC MARGINAL REGRESSION MODELS FOR REPEATED OUTCOMES IN THE PRESENCE OF INFORMATIVE FOLLOW-UP

Haiqun Lin*, Division of Biostatistics,Yale University
Daniel O.Scharfstein, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Heallth
Robert A. Rosenheck, North East Program Evaluation Center
Departments of Psychiatry & Public Health, Yale Univeristy, New Haven, CT

In longitudinal studies, subjects may miss scheduled visits or attend at unscheduled times. As a result, observed outcome data may be highly unbalanced creating a situation of intermittent missing data. Inference about marginal parameters for the longitudinal outcome maybe biased when using generalized estimating equatiton of Liang and Zeger (Biometrika, 1986). Building on the work of Robins, Rotnizky and Zhao (JASA, 1995), we propose a class of inverse intensity of visit process weighted estimators in marginal regression models for longitudinal response measured in continuous time. Large sample theory for our estimators is derived. Our method is illustrated using data from an experimental intervention study designed to reduce homelessness. The finite sample behavior of our estimators is investigated through simulation studies.

e-mail: haiqun.lin@yale.edu

LIKELIHOOD METHODS FOR COUNTERFACTUAL INFERENCE WITH CONFOUNDING VARIABLES

Abigail L. Jager*, Department of Statistics, University of Chicago
Paul J.Rathouz, Department of Health Studies, University of Chicago

In a non-randomized observational study, in order todetermine the counterfactual effect of a treatment variableon a response, one must adjust for all relevant confounding factors. This paper presents a likelihood-based approach to this problem by modeling the joint distribution of each counterfactual outcome and the confounders. The inferential targets of interest are the marginal distributions of the counterfactuals for each possible treatment value. This allows us to draw randomization type inferences regarding the effect of treatment on response, particularly in the case where some of the confounding variables are recorded concurrently with, or after the treatment is assigned. To demonstrate our method we estimate the effect of different cleansing regimens on burn infections and compare the results to other methods.

e-mail: jager@galton.uchicago.edu

CAUSAL LINEAR MODELS FOR NON-COMPLIANCE UNDER RANDOMIZED TREATMENT

Thomas R. Ten Have*, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania
Michael Elliott, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania
Marshall Joffe, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania
Elaine Zanutto, Wharton School, Wharton School of Business, University of Pennsylvania

In this paper, we compare two classes of causal methodology with respect to additive treatment effects on a univariate continuous outcome for two separate randomized encouragement trials in the context of treating depression in primary care practices. The two approaches are Imbens and Rubin's latent class model and Robins' structural mean model. Of particular interest is a comparison of these two causal approaches with respect to how they accommodate departures from assumptions that are typically made for causal inference. One specific assumption that may not hold in randomized encouragement trials is the exclusion restriction. We show that varying this assumption changes the relationship between these two approaches.

e-mail: ttenhave@cceb.upenn.edu

SEMIPARAMETRIC EFFICIENT ESTIMATION OF TREATMENT EFFECT IN A PRETEST-POSTTEST STUDY

Selene Leon*, North Carolina State University
Anastasios A.Tsiatis, North Carolina State University
Marie Davidian, North Carolina State University

Inference on treatment effects in a pretest-posttest study is a routine objective in medicine, public health, and a host of other fields, and a number of approaches have been advocated in the literature. We consider this problem from a semiparametric perspective, making no assumptions about the distributions of baseline and posttest responses. By representing the situation in terms of counterfactual random variables, we exploit recent developments in the literature on missing data and causal inference to derive the class of all treatment effect estimators, identify the most efficient estimator within this class, and outline a practical strategy for implementation of estimators that may improve on popular methods. We demonstrate the methods and their properties via simulation and by application to a data set from an HIV clinical trial.

e-mail: sleon@stat.ncsu.edu

MULTIPLE IMPUTATION FOR NONRANDOM MISSING OUTCOMES IN TRIALS WITH ANCILLARY RELIABILITY OR VALIDITY MEASURES

James B. Kampert*, The Cooper Institute,

In clinical trials with baseline and post-intervention outcome measures, missing outcomes may bias treatment comparisons. Unobserved outcomes of dropouts who receive little or no treatment are more likely to resemble the baseline measures than outcomes of participants who complete the trial. Imputing missing outcomes with the baseline measures (last observation carried forward) leads to inflated type-I error rates due to restricted variation in the imputed outcomes. Proper missing data methods require a model for the distribution of nonrandom missing data. Some clinical trials incorporate either test-retest replication or validation with a gold standard to determine the reliability or validity of the primary endpoint. When we believe missing outcomes would replicate the baseline measures plus random error, a reliability or validity ancillary study provides a model for the missing data. Random observations may be drawn from the Bayesian bootstrap distribution of the missing outcome given the observed baseline measure to impute missing outcomes. We illustrate with Project PRIME, a 24-month physical-activity intervention trial, where the primary outcome was evaluated for reliability.

e-mail: jkampert@cooperinst.org

FITTING A MIXED EFFECTS MARKOV MODEL FOR REPEATED BINARY OUTCOMES WITH NONIGNORABLE DROPOUT

Robert J. Gallop*, West Chester University
Thomas R.Ten Have, University of Pennsylvania

In many areas of research, repeated binary measures often represent a two-state stochastic process, where individuals can transition among two states. In a behavioral or physical disability setting, individuals can flow from susceptible or subthreshold state, to an infectious or symptomatic state, and back to a subthreshold state. Quite often the transition among the states happens in continuous time but are observed at discrete, irregularly spaced time points which may be unique to each individual. Methods for analyses are typically based on the Markov assumption. Cook (1999) introduced a Markov model that accommodates the subject to subject variation in the model parameters with random effects. We extend this model by adding a nonignorable dropout component to the model. Specification of the distribution of the random effects is made to guarantee a closed form expression of the marginal likelihood. This methodology is illustrated by applications to a data set from a parasitic field infection survey, a data set from a cocaine treatment study, and a data set from an aging study.

e-mail: rgallop@wcupa.edu

ON THE CONSTRUCTION OF BOUNDS IN PROSPECTIVE STUDIES WITH MISSING ORDINAL OUTCOMES: APPLICATION TO THE GOOD BEHAVIOR GAME TRIAL

Daniel O. Scharfstein*, Johns Hopkins Bloomberg School of Public Health
Charles F.Manski, Northwestern University
James C. Anthony, Johns Hopkins Bloomberg School of Public Health

This paper is concerned with drawing inference about aspects of the population distribution of ordinal outcome data measured on a cohort of individuals on two occasions, where some subjects are missing their second measurement. Our approach is based on the construction of bounds under plausible assumptions on the missing data mechanism. We develop our methodology within the context of a randomized prevention trial of an intervention called the "Good Behavior Game," which was designed to reduce aggressive misbehavior among children.

e-mail: dscharf@jhsph.edu

STATISTICAL ANALYSIS OF PSA PATTERNS FOLLOWING RADIATION THERAPY FOR PROSTATE CANCER

James A. Hanley*, McGill University
Elizabeth L.Turner, McGill University
Nandini Dendukuri, McGill University
Peter C. Albertsen, University of Connecticut

After radiation therapy, prostate specific antigen (PSA) values reach a nadir during the 1st 2 years, then rise slightly with age, or rapidly with increasing tumor volume. We describe methods to study the nature/frequency of these patterns, using serial values in >400 men, with varying numbers of observations at unequally spaced times. We fit a hierarchical model to allow for variations in the pattern for each man. The model for each man comprised a different 'checkmark'-type pattern i.e., a linear decrease of log2PSA values followed by a linear increase. The rates of increase were modeled as a mixture of 2 distributions, one for those 'cured' , and one for more rapid, tumor-related, increases. The model allowed us to estimate the frequency of each pattern, establish a boundary between the 2, relate doubling times to patient/tumor features, and compare doubling times with those after surgery. Model parameters were estimated using a Gibbs Sampler. Compared with fitting a different fixed effects model for each man, this approach allowed us to use data from more patients, particularly those with short data series.

e-mail: James.Hanley@McGill.CA

NEUROTOXICITY RISK ASSESSMENT: MODELING OF BEHAVIORAL CHANGES IN MULTIPLE ENDPOINTS

Yiliang Zhu*, Dept. of Epidemiology and Biostatistics, College of Public Health, Univ, of South Florida

Neurotoxic effects are an important non-cancer endpoint in human health risk assessment. As a first tier test neurobehavioral screenings are designed to identify chemicals with neurotoxical potential, and generate longitudinal dose-response data to profile neurological effects. Repeated measurements analysis of variance has been the standard in analyzing such data, but it fails to meet the requirement of dose-response assessment. Derivation of safety standard measures using methods such as the benchmark doses increasingly requires explicit dose-response modeling. In this paper we discuss the development of a family of toxicokinetic-toxicodynamic models and the application of mixed effects models to fit these models. The models allow for both transient and permanent neurobehavioral effects as well as identical or varying timing of peak effects. We also introduce a procedure to profile benchmark doses over time so that risk estimate is with respect to the time window where the exposure impact would be most severe. All illustrations are based on data generated from the EPA Superfund study and the study conducted by the International Programme in Chemical Safety.

e-mail: yzhu@hsc.usf.edu

PERFORMANCE OF SHRINKAGE ESTIMATORS FOR PREDICTION IN MULTIPLE REGRESSION

Xue Xin*, Dept. of Biostatistics, Tulane University School of Public Health and Tropical Medicine
Leann Myers,

Regression models in medical research are widely used for risk profile assessment, as diagnostic tools for disease, or for predicting future cases. When predicting the response at a new randomly chosen covariate vector x, problems can arise. The fit of a regression model to new data is nearly always worse than its fit to the original data, a deterioration known as shrinkage. Stein (or shrinkage) predictors, give a uniformly lower mean squared error for prediction (MSEP) than least squares estimators under certain assumptions. Three different currently used forms of the parametric shrinkage prediction in multiple regression were computed and compared in terms of empirical expectation of MSEP using re-sampling and cross-validation. A nonparametric estimator proposed by Copas was used to estimate shrinkage directly from sets of data without the usual parametric assumptions. Data were generated from models with both normally and nonnormally distributed error terms. Special cases for residual variances following particular patterns of heteroscedasticity were investigated. The parameters p (number of predictors), N, n and beta were varied. Suggestions for applied usage of shrinkage prediction using multiple regression are proposed.

e-mail: xxin@tulane.edu

MEASUREMENT ERROR ADJUSTMENT IN OCCUPATIONAL HEALTH STUDIES WHILE ACCOUNTING FOR NONDETECTABLE EXPOSURES

Kathleen A. Wannemuehler*, Emory University
Robert H.Lyles, Emory University

In occupational health studies, the goal is often to associate a worker's exposure to a known environmental pollutant with a continuous, adverse health effect variable. In many cases the true mean exposure is unknown and therefore a suitable surrogate, often measured with error, is utilized. A second source of measurement error can arise when the surrogate measure is based upon repeated exposure measurements that may not all be observable because they fall below a detectable limit. The motivating data set is from a study by Heederik et al. (1991) where repeated shift-long dust exposure measurements taken over a period of a year are related to a single response measure, forced expiratory volume (FEV1), in subjects working in the Dutch animal feed industry. Lyles and Kupper (1997) investigated several methods for addressing the first source of measurement error, assuming that a worker's true unknown mean dust exposure over the year was to be associated with FEV1. We propose to extend their approach in the context of full maximum likelihood analysis, while appropriately accounting for exposure non-detects that were previously ignored. The results from our re-analysis, together with a simulation study to illustrate the performance of the proposed approach, will be presented.

e-mail: kwannem@sph.emory.edu

AN ALTERNATIVE NONITERATIVE ESTIMATOR FOR THE HETEROGENEITY VARIANCE IN META-ANALYSIS

Kepher H. Makambi*, Howard University Cancer Center

The DerSimonian and Laird (1986) estimator of the heterogeneity variance can be regarded as the standard estimator in meta-analysis. Besides the possibility that this estimator can initially take on negative values, values that are not very useful and contradict practical understanding, it is also not possible to construct an explicit confidence interval on the parameter of interest. Here, a positive estimator, almost everywhere, of the heterogeneity variance is proposed. An explicit variance formula for the estimator is given and an approximate confidence interval for heterogeneity variance is constructed. By simulations, the bias and the MSE of the proposed estimator are compared with the DerSimonian-Laird estimator. Attained confidence coefficients for the constructed confidence interval are also presented.

e-mail: makambi1@hotmail.com

82. Analysis of Microarrays and Gene Expression III

STATISTICAL SIGNIFICANCE ANALYSIS OF LONGITUDINAL GENE EXPRESSION DATA

Xu Guo*, Division of Biostatistics, School of Public Health, University of Minnesota
Huilin Qi, Stem Cell Institute, Department of Medicine, University of Minnesota Medical School
Catherine M. Verfaillie, Stem Cell Institute, Dept. of Medicine, University of Minnesota Medical School
Wei Pan, Division of Biostatistics, School of Public Health, University of Minnesota

Longitudinal gene expression data arise from time-course microarray experiments, which are designed to study biological processes in a temporal fashion by taking samples from the same subject at different time points to measure gene expression levels. It is well known in other applications that valid statistical analyses have to appropriately account for possible correlation in longitudinal data. For this reason, we apply estimating equation techniques to construct a robust statistic, which is a variant of the robust Wald statistic, for longitudinal gene expression data to detect genes with temporal changes in expression. We associate significance levels to the proposed statistic by either incorporating the idea of the Significance Analysis of Microarrays (SAM) method (Tusher et al, 2001) or using the mixture model method (MMM) (Pan et al, 2001) to identify significant genes. The utility of the statistic is demonstrated through its application to an important study of osteoblast lineage-specific differentiation. Using simulated data, we also show pitfalls in drawing statistical inferences wh n the correlation in longitudinal gene expression data is ignored.

e-mail: weip@biostat.umn.edu

DETECTING DNA REGULATORY MOTIFS BY INCORPORATING POSITIONAL TRENDS IN INFORMATION CONTENT

Katherina Kechris*, Department of Statistics, U.C. Berkeley
Erik van Zwet, Department of Statistics, U.C. Berkeley
Peter Bickel, Department of Statistics, U.C. Berkeley
Michael Eisen, Lawrence Berkeley Natl. Lab, Molecular and Cell Biology, U.C., Berkeley

Model-based motif discovery methods, such as MEME and Gibbs Motif Sampler, are important tools for screening DNA regulatory regions to predict binding sites of proteins involved in transcription. Although these methods are very successful, they can be too general, having been developed for both DNA and protein sequences. These algorithms can be distracted by noisy signals in the data that are not characteristic of true transcription factor binding sites. We propose a simple extension to the underlying model of these methods to improve the prediction of real sites. Our method is based on the observation that examples of real sites show positional trends in the information content. We assign prior distributions to the frequency parameters of the model, penalizing deviations from the specified information content (e.g. only one base is conserved). Simulations studies show that these changes
improve the algorithm's ability to discover motifs with information content patterns typical of real binding sites.

e-mail: kechris@stat.berkeley.edu

A Generalized Additive Model for Microarray Gene Expression Data Analysis

Chenan Tsai*, National Center for Toxicological Research, Food and Drug Administration
Hueymiin Hsueh, National Chengchi University, Taipei, Taiwan
James J. Chen, National Center for Toxicological Research, Food and Drug Administration

Microarray technology allows the measurement of expression levels of a large number of genes simultaneously. Various statistical methods have been proposed for data normalization and data analysis. This paper proposes a generalized additive model for the analysis of gene expression data from different types of microarray experiments. This model consists of a two-step normalization procedure for the measured intensity data. The first step involves a non-parametric regression using lowess fits to adjust for non-linear systematic biases. The second step uses a linear ANOVA model to estimate the remaining effects including the interaction effect of genes and treatments, the effect of interest in a study. The proposed model is a general ANOVA model for microarray data analysis. We show correspondences between the lowess fit and the ANOVA model methods. The procedure can be applied to one channel or two channel data from the experiments with multiple treatments or multiple nuisance factors. Two toxicogenomic experiment data sets and a simulated data set are used to contrast the proposed method
with the commonly known lowess fit and ANOVA methods.

e-mail: CTsai@nctr.fda.gov

A BAYESIAN CLASSIFICATION METHOD FOR TREATMENTS USING MICROARRAY GENE EXPRESSION DATA

Yuan Ji*, UW-MADISON
Kam-Wah Tsui, UW-MADISON
KyungMann Tsui, UW-MADISON

An important application of microarray gene expression data is classification of treatments. One pioneering work in this area is by Golub et al. (1999) \nocite{Golub:1999} who classified samples into three types of leukemia using oligonucleotide microarrays. Recently, with more applications appearing in the literature, researchers began to realize the importance of constructing accurate and efficient classification methods in order to obtain desirable classification results. In this article, we propose an empirical Bayesian classification method that accommodates data structures emerging from microarray experiments such as correlations among gene expressions and high dimensionality with more genes (as predictor variables) than treatments (as data points). Specifically, we build a latent cluster structure for genes into a Bayesian classification process to achieve two goals: to account for the correlation among gene expressions and to reduce the number of nuisance parameters in the Bayesian model. We demonstrate our method via simulation as well as a case study using a real data set.

e-mail: yuanj@stat.wisc.edu

BAYESIAN ERROR-IN-VARIABLES SURVIVAL MODEL FOR THE ANALYSIS OF GENECHIP ARRAYS

Mahlet G. Tadesse*, Texas A&M University, Department of Statistics
Joseph G.Ibrahim, University of North Carolina, Chapel Hill, Department of Biostatistics
Robert Gentleman, Harvard School of Public Health and Dana-Farber Cancer InstituteSabina Chiaretti, Dana-Farber Cancer Institute, Cell manipulation and gene transfer laboratory

The analysis of microarray experiments is complicated by the significant amount of assay noise leading to expression readings that are often imprecise and can substantially differ from the true levels. Failure to accomodate this error and drawing inference based on the observed covariates can seriously affect the assessment of gene effects. We present a Bayesian semi-parametric survival model that adjusts for measurement error in the predictors. A piecewise constant hazards model is used for the time-to-event data and the gene expression estimates are assumed to fluctuate around the true unobserved transcript levels. The model is illustrated using microarray data from an adult acute lymphoblastic leukemia trial.

e-mail: mtadesse@stat.tamu.edu

LINEAR REGRESSION AND TWO-CLASS CLASSIFICATION WITH GENE EXPRESSION DATA

Xiaohong Huang*, University of Minnesota, Division of Biostatistics
Wei Pan, University of Minnesota, Division of Biostatistics

Using gene expression data to classify (or predict) tumor types has received much research attention recently. Due to special features of gene expression data, several new methods have been proposed, including the weighted voting scheme of Golub et al (1999), the compound covariate method of Hedenfalk et al (2001) (originally proposed by Tukey (1993)), and the shrunken centroids method of Tibshirani et al (2002). These methods look different and are more or less ad hoc. Here we point out a close connection of the three methods with a linear regression model. Casting the classification problem in the general framework of linear regression naturally leads to new alternatives, such as modified partial least squares (PLS) methods. Using two real data sets, we show the competitive performance of our new methods when compared with the other three methods.

e-mail: xiaohong@biostat.umn.edu

83. Generalized Linear Models and Regression

A SEMIPARAMETRIC EMPIRICAL LIKELIHOOD METHOD FOR BIASED SAMPLING SCHEMES IN EPIDEMIOLOGIC STUDIES WITH AUXILIARY COVARIATES

Xiaofei Wang*, Department of Biostatistics, University of North Carolina at Chapel Hill
Haibo Zhou, Department of Biostatistics, University of North Carolina at Chapel Hill

We consider statistical inference for epidemiologic studies conducted with both a simple random subsample and multiple outcome-dependent subsamples. The subsamples may be further stratified on a discrete auxiliary covariate. The advantage of such studies is that, while providing overall information about the population through the simple random subsample, it allows the investigators to oversample certain subpopulations that are believed to be more informative. We propose an efficient semiparametric method to fit a generalized linear model to the categorical outcome data from such studies.The novelty of the proposed method is that the conditional distribution of the model covariates is estimated by the empirical likelihood method. Simulation studies show that the proposed estimator outperforms alternative methods. The proposed method is illustrated with a dataset from the Collaborative Prenatal Project (CPP) to assess the association between maternalpolychlorinated biphenyl (PCB) levels and children's neurodevelopmental status.

e-mail: xfwang@bios.unc.edu

FUCTIONAL REGRESSION MODELS AND TEMPORAL PROCESSES

Jun Yan*, Department of Statistics, University of Wisconsin

We consider regression for response and covariates which are temporal processes observed over intervals. A functional generalized linear model is proposed which includes extensions of standard models in multistate survival analysis. Simple nonparametric estimators of time-indexed parameters are presented and shown to be uniformlyconsistent and to converge weakly to Gaussian processes. The procedure does not require smoothing or a Markov assumption, unlike approaches based on transition intensities. The estimators are the basis for new tests of the covariate effects and for the estimation of models in which greater structure is imposed on the parameters. The methodology enables goodness-of-fit testing and permits predictions involving estimated components from both the functional model and the submodels. Its practical utility is illustrated in simulations and a data analysis of the prevalence of Chronic Graft Versus Host Disease (CGVHD).

e-mail: jyan@stat.wisc.edu

CENSORED REGRESSION MODELS FOR PREDICTING SUICIDE ATTEMPT LETHALITY

Hanga C. Galfalvy*, New York State Psychiatric Institute
Maria A.Oquendo, New York State Psychiatric Institute, Columbia University
Michael F. Grunebaum, New York State Psychiatric Institute, Columbia University
John J. Mann, New York State Psychiatric Institute, Columbia University

Psychiatrists assess suicide attempt lethality (a measure of the medical damage) on a 9-point scale. The maximum lethality of all the attempts committed by a psychiatric patient is often used as a measure of disease severity. Ordinal regression models can be used to assess the relationship between the maximum lethality of past suicide attempts and clinical and demographic variables in a retrospective study. In a prospective study setting, however, the maximum lethality attempt may be unobserved, thus the lethality value will be censored. We will use parametric censored regression models to predict maximum attempt lethality for future suicide attempts, and study the connection with survival analysis.

e-mail: hanga@neuron.cpmc.columbia.edu

REGRESSIONS WITH SINGULAR DESIGN

Wenjiang Fu*, Michigan State University, Department of Epidemiology

We consider an ill-conditioned problem - regressions with singular design. This topic not only has many practical applications in engineering, health studies and social research, but is also of great interest in statistical theory. Although it has been studied theoretically and a number of methods have been proposed, the available methods are far less than enough to provide a positive answer to many questions the topic raises. For example, the identifiability problem has been unsettled, which refers to the full determination of parameter estimation and is difficult due to the existence of multiple estimators. I will review some methods in the literature and present some novel approaches, which can be very useful in studying this kind of problems.

e-mail: fuw@msu.edu

METHODS OF LEARNING IN STATISTICAL EDUCATION: DESIGN OF A RANDOMIZED TRIAL AND ANALYSIS USING MARGINAL STRUCTURAL MODELS

Felicity T. Boyd*, Johns Hopkins University, Bloomberg School of Public Health, Dept. of Biostatistics
Marie Diener-West, Johns Hopkins University, Bloomberg School of Public Health, Dept. of Biostatistics

Active learning methods may influence understanding of statistical concepts. A randomized trial of 265 consenting students was conducted within an introductory graduate biostatistics course: 69 received eight small group cooperative learning sessions; 100 accessed internet activities; and 96 received no intervention. Analytic methods for assessing intervention effects on examination score included intent-to-treat analysis, generalized linear models (GLMs), and marginal structural models (MSMs). There were no differences in examination score by intent-to-treat; however, 51% dropped out by the third session. The GLMs, using reported participation, yielded an adjusted average improvement of 3.9 points for one cooperative session (95% CI: 1.5-6.3) and 3.2 points for one internet session (95% CI: 0.8-5.6) per 100 point examination. The MSMs, incorporating the probability of session participation, suggested improvement for both interventions (2.6 points (95% CI: 0.3-4.9) and 2.4 points (95% CI: 0.0-4.7), respectively). Participation in active learning may improve understanding of statistical concepts and enhance opportunities for instruction beyond the traditional classroom.

e-mail: fboyd@jhsph.edu

MODELING EXCESS ZEROS IN ORDINAL DATA

Mary E. Kelley*, VA Pittsburgh Healthcare System
Stewart J.Anderson, Dept of Biostatistics, University of Pittsburgh Graduate School of Public Health

One of the most common issues surrounding the analysis of clinical and utilization outcomes is what statisticians refer to as the excess zero problem. This occurs when an outcome such as symptom severity or counts of service utilization is measured and the resulting data collected contain many observations that are zero, i.e., did not possess the symptom or did not use the service being measured. The case of excess zeros has previously been handled using mixture models for Poisson, normal, truncated normal and log-gamma distributions among others. We develop a zero inflated ordinal model that is similar in derivation to previous zero inflated models, but with the distribution component consisting of a multinomial distribution. Our formulation is based on the proportional odds model suggested by McCullagh. However, this methodology can be applied to any situation in which a multinomial model is appropriate. We assess the usefulness of the zero inflated ordinal model as compared to the standard ordinal model using data simulated from a mixture distribution. We also give an example using alcohol consumption data obtained in large primary care sample.

e-mail: chaoschic@aol.com

POSTER SESSION

TESTS FOR EFFECTS ON TUMOR FREQUENCY AND LATENCY IN MULTIPLE DOSING PHOTOCARCINOGENICITY EXPERIMENTS

Daniel F. Molefe*, National Center for Toxicological Research
Ralph L.Kodell, National Center for Toxicological Research
James J. Chen, National Center for Toxicological Research

The single induction test procedure for modeling multiple tumor data by Kodell and Chen (2001) is extended to a test for multiplicity in multiple induction experiments. The new test can detect overall differences in the distribution of the observed tumors between two groups as well as isolate whether the detected overall differences is due to the difference in the distribution of time to tumor observation and/or the difference in the frequency of tumors. The behavior of the proposed test was studied via Monte Carlo Simulations studies. Data was simulated to resemble some of the experiments conducted at the National Center for Toxicological Research. Using the simulation studies we were able to show that parameters were estimated reasonably well. Furthermore, simulated size and power of this test was fairly close to the desired nominal values. Results of the proposed test were compared to the results of the Logrank test, the negative binomial test, and Latency Multiplicity test.

e-mail: dmolefe@nctr.fda.gov

LIFETIME RISK OF DIABETES MELLITUS IN THE UNITED STATES

Kabayam M. Venkat Narayan, Centers for Disease Control and Prevention
James P.Boyle*, Centers for Disease Control and Prevention
Theodore J. Thompson, Centers for Disease Control and Prevention
Stephen W. Sorensen, Centers for Disease Control and Prevention
David W. Williamson, Centers for Disease Control and Prevention

National Health Interview Survey data were used to estimate age-, race-, and sex-specific prevalence and incidence of diagnosed diabetes in year 2000. U.S. Census Bureau data, and a previous study of diabetes as a cause of death, were used to estimate mortality rates for the diabetic and non-diabetic populations. These estimates were then entered into a Markov model to estimate age at diagnosis, residual lifetime risk of diagnosed diabetes, duration with diabetes, and life-years lost from diabetes. The estimated lifetime risk of developing diabetes for people born in 2000 was 33% for males and 39% for females, with females having higher lifetime risks at all ages throughout the lifespan. The highest estimated lifetime risks of diabetes were in Hispanics, with males having a 45% risk and females a 53% risk. The mean age at diagnosis ranged from 56 years to 59 years for both sexes, depending on race. The youngest age at diagnosis was observed in blacks, with a mean age of 56 years for both sexes. People diagnosed with diabetes had large reductions in life expectancy. For people diagnosed at age 60, males lost 7.3 years and females lost 9.5 years of remaining life.

e-mail: jboyle@cdc.gov

A BAYSIAN APPROACH TO COST-EFFECTIVENESS WHEN THE OUTCOME DEPENDS ON A PREDICTIVE MODEL

Raymond G. Hoffmann*, Medical College of Wisconsin
Laud W.Purushottam, Medical College of Wisconsin
Hayat Matthew, HIH/NIDCD
Ana Johnson-Masotti, McMaster University

Because it is a ratio of two random quantities, the cost-effectiveness ratio is a difficult quantity to obtain a confidence interval. It is even more difficult when the effectiveness must be obtained with a predictive model based on changes in behavior. This study presents a Bayes model for the incremental cost-effectiveness ratio, the ICER, when the effectiveness is estimated with a Bernoulli model that translates behavioral change into changes in the probability of HIV infection. The model uses a multivariate approach to produce posterior distributions that are usually handled univariately with sensitivity analysis or with probabalistic sensitivity analysis.

e-mail: hoffmann@mcw.edu

APPROACHES FOR HAPLOTYPE-PHENOTYPE ASSOCIATION STUDIES IN UNRELATED SUBJECTS

Serkalem Demissie*, Department of Biostatistics, Boston University School of Public Health
Jose M.Ordovas, Jean Mayer USDA Human Nutrition Research Center on Aging at Tufts University
L. Adrienne Cupples, Department of Biostatistics, Boston University School of Public Health

We evaluated the association between plasma low-density lipoprotein cholesterol (LDL-C) and haplotypes in the scavenger receptor class B type I (SR-BI) gene in 1607 subjects from the Framingham Heart Study. First, individual haplotypes were inferred from haplotype frequencies estimated by the EM algorithm from 3 markers. Regression analysis was then performed with LDL-C as an outcome and
haplotypes as predictors, using two approaches: 1) for each subject we assigned multiple observations, one for each haplotype, and weighted the observations by haplotype probabilities; and 2) for each subject we assigned k variables (one for each haplotype) with values equal to the haplotype probabilities, and used k-1 variables as predictors. Results from these approaches were compared with haplo.score (http://www.mayo.edu/statgen/software/) results; haplo.score computes score statistics and p-values. The p-values computed using the second approach were similar to those from haplo.score. Results from the first and second approaches were notably different and simulation studies are needed to understand the differences. In association studies it is useful to evaluate effect estimates in addition to p-values. In this study, we have demonstrated that standard computer programs can be used to obtain estimates of haplotype-effects and their p-values.

e-mail: demissie@bu.edu

COMPARING NONPARAMETRIC CORRELATION COEFFICIENTS

Leann Myers*, Dept. of Biostatistics, Tulane University
Maria J.Sirois, Dept. of Biostatistics, Tulane University

Nonparametric correlation coefficients are often used in place of Pearson correlation coefficients when data are not bivariate normal. The rank correlation coefficient Spearman?s rho is probably the most common choice although Kendall?s tau is another option. The standard errors of these correlation coefficients are easy to compute under the null distribution (population correlation = 0) but they are less tractable under the assumption of nonzero population correlation, and the distributions of the coefficients are highly skewed. This makes traditional hypothesis testing difficult. The current study investigated possible methods of testing the equality of two nonparametric correlation coefficients when the data were not bivariate normal and the assumed population correlations were nonzero. Bootstrapping methods, approximate solutions, and transformations were compared in terms of Type 1 error rates, ease of usage, and computational time.

e-mail: myersl@tulane.edu

GEOMETRY OF THE BIVARIATE NORMAL DISTRIBUTION

Timothy D. Curtsinger*, Biostatistics Division, University of Lousville
Linda J.Goldsmith, Dept of Bioinformatics and Biostatistics, University of Louisville,

This poster presentation demonstrates the relationship between the random variables Y and X in a bivariate normal distribution. Various combinations of population correlation, means, and variances are used to demonstrate relationships. The geometry, trigonometry, and distribution theory of the bivariate normal distribution are used to generate high-resolution graphs depicting equiprobable ellipsoids, regressions and principal component axes of the population data. Scattergrams of simulated data and estimated parameters and lines will be superimposed in contrasting colors in some cases of interest.

e-mail: TCurtsinger@aol.com

THE ASSOCIATION OF ESTROGEN RECEPTOR POLYMORPHISM AND CHANGE IN OSTEOCALCIN, A BONE TURNOVER MARKER

Wei Liang*, Dept of Biostatistics, Univ of Michigan
Mary L.Jannausch, Dept of Epidemiology, Univ of Michigan
Maryfran R. Sowers, Dept of Epidemiology, Univ of Michigan

Increasing evidence suggests that maintenance of peak bone mass, bone turnover and bone loss have strong genetic determinants. Osteocalcin (OCN), a skeletal extracellular mineral binding protein, is used as a bone formation marker because it is associated with bone mineral density (BMD) loss and increased osteoporotic fracture risk. We examined whether (1) the association of OCN level with femoral neck and spine BMD is different according to selected genotypes for the estrogen receptor, vitamin D receptor, type I collagen, OCN gene promoter region; and (2) change in OCN level overtime is determined by genotype. Ten-year longitudinal data was collected on 664 women aged 24-50 in the Michigan Bone Health Study. To address those OCN measurements below detection, we used survival analysis of left-censored data and logistic regression. Proc Mixed (SAS) was used for correlated repeated measurements. We found that the association of OCN with femoral neck BMD was significantly different according to the genotypes (+/+, +/-, -/-) of the estrogen receptor alpha defined by the restriction endonucleases PUV_II and XBA_I. However, these genotypes were not associated with OCN change overtime.

e-mail: wliangz@umich.edu

COMPARISON OF RESULTS FROM TWO METHODS FOR EXAMINING THE REPRODUCIBILITY OF BREAST FLUID BIOMARKERS

Peter H. Gann, Feinberg School of Medicine, Northwestern University
Robert T.Chatterton, Jr., Feinberg School of Medicine, Northwestern University
Borko D. Jovanovic, Feinberg School of Medicine, Northwestern Univ.
Irene B. Helenowski*, Feinberg School of Medicine, Northwestern University
Angela S. Geiger, Feinberg School of Medicine, Northwestern University

An important facet in cancer research involves biomarker assays, detecting levels of particular hormones and other proteins in high-risk patients. For example, changes in concentration of female hormones such as estradiol and progesterone in women at high risk for breast cancer particularly interest researchers (Gann et al. 2001). For this presentation, we examine between breast reproducibility of various nipple aspirate fluid proteins using permutation methods involving the Spearman correlation, suggested by Gann et al. (1997), and the intraclass correlation coefficient (ICC), based on a model with factors important to the design of a clinical trial from where the data was taken. Both methods have their advantages and limits. The Spearman correlation is non-parametric, but the permutation method does not account for differences in intercept, as the ICC does, based on the variance components of random effects (Deyo et al., 1991). The ICC, however, relies on the assumption that errors and random effects are normally distributed. Future endeavors involve development of a semi-parametric ICC.

e-mail: i-helenowski@northwestern.edu

REVIEW AND EVALUATION OF METHODS FOR COMPUTING CONFIDENCE INTERVALS FOR THE RATIO OF TWO PROPORTIONS AND IMPLICATIONS TO NON-INFERIORITY CLINICAL TRIALS

Rebekkah S. Dann*, University of North Carolina at Chapel Hill
Gary G.Koch, University of North Carolina at Chapel Hill

This paper reviews several methods for forming confidence intervals for ratios of two independent binomial proportions and evaluates their statistical performance. These methods include use of a Taylor Series expansion to estimate variance, solutions to a quadratic equation, and maximum likelihood methods. These methods were initially compared by computing the confidence limits and judging similarity of results. Secondly, simulations were used to identify the better methods for controlling the type I error rate while maintaining appropriate power. Lastly, relationships between more complex, computer-intensive methods and simpler methods were evaluated through correlations. Findings support the use of the simpler group of methods using a Taylor Series expansion in variance estimation. This method meets criteria for determining better methods including favorable results obtained from a review of the literature, reduction of type I error rate while maintaining power, ease of computation, and similarities with more complex methods. Applications of these findings are discussed for ratios which arise in randomized clinical trials that are conducted to show non-inferiority.

e-mail: rdann@bios.unc.edu

AN EVALUATION OF METHODS FOR MISSING DATA IN A LARGE-SCALE EPIDEMIOLOGICAL STUDY

Jamie L. Bigelow*, University of North Carolina at Chapel Hill
Amy Herring, University of North Carolina at Chapel Hill
Nancy Dole, University of North Carolina at Chapel Hill
Diane Kaczor, University of North Carolina at Chapel Hill
David Savitz, University of North Carolina at Chapel Hill

The Pregnancy, Infection, and Nutrition (PIN) study is a prospective cohort study examining potential risk factors for preterm birth. These risk factors (which include diet quality, drug use, medical history, genetic factors, stress and anxiety levels, and socioeconomic characteristics) are measured using self-administered questionnaires, telephone interviews, and biological samples. Given the scope of this study, missing data is nearly unavoidable.

Concentrating on stigmatized behaviors (such as smoking and drug use) for which missing data are suspected to be nonignorable, we conduct a sensitivity analysis and explore whether a standard complete case analysis leads to substantial bias or inefficiency for these data. In particular, we are sensitive to investigator desires to find an appropriate analysis that is most accessible in terms of currently-available software.

e-mail: jbigelow@bios.unc.edu

SPOT QUALITY MEAUSRES IN MICROARRAY EXPERIMENT

Taesung Park, University of Pittsburgh andSeoul National University
Ki Woong Kim, Seoul National University
Seung Yeoun Lee*, Sejong University
Jin Hyuk Kim, Hanyang University

Microarray technology allows us to monitor expression levels of thousands of genes simultaneously and helps us to understand genome-wide gene regulations and interactions. However, in microarray experiments many undesirable systematic variations are commonly observed. These variations can affect the quality of spots and slides. Recently, many researches have been proposed regarding quality control in microarray data. For example, the signal to noise ratio and the local background coefficients of variation (CV) are commonly used. However, previous studies focus on measures based on spot intensity and some of them are highly dependent on the specific experimental conditions. In this paper, new morphological quality measures are proposed using spot shapes and noise patterns. Through real microarrary data and artificially generated microarray data, we show that the proposed measures are more effective than other measures in detecting spots with poor quality.

e-mail: leesy@sejong.ac.kr

ESTIMATING RISK FACTORS UNDER CLUSTERED SURVIVAL DATA WITH TIME DEPENDENT COVARIATES

Leila D. Amorim*, Department of Biostatistics, UNC-Chapel Hill
Shrikant Bangdiwala, Department of Biostatistics, UNC-Chapel Hill
Joanne S. Harrell, School of Nursing, UNC-Chapel Hill

The Cardiovascular Health in Children (CHIC) study is a randomized controlled trial of behavior modification interventions in young adolescents in North Carolina for impacting risk factors for future cardiovascular diseases. Schools, rather than individual subjects, were randomized to three different intervention programs. Individual subjects were evaluated almost every year up to 7 years of follow-up and physiological information, such as weight, height, blood pressure, age and physical activity was collected. We are interested in evaluating the effect of age, physical activity, intervention program, gender and race on time to development of the composite outcome of obesity (defined as BMI - body mass index – over 85%tile) and hypertension (defined by age-gender specific national cut points). The survival data is clustered as subjects are grouped into schools. In this paper we propose the use of a discrete proportional hazards model, including age and physical activity as time-dependent covariates. We are using SUDAAN (RTC, Inc.) to perform this analysis. The model parameters are estimated using a standard partial likelihood, and a robust variance estimator is applied. The intracluster correlation was also estimated. For comparison purposes, we also include results based on the assumption of complete independence among the subjects at the same school.

e-mail: lamorim@email.unc.edu

GROUP LEVEL EFFECTS IN HIERARCHICAL POISSON MODELS

Leslie H. Morgan*, Tulane University
Leann Myers, Tulane University
Frances J. Mather, Tulane University

One application of hierarchical Poisson models is to analyze the spatial distribution of cancer cases. Individual level data often consist of age, race, sex, and year of diagnosis. Individuals are nested within political/geographic areas such as counties. An empirical study of the hierarchical Poisson model's performance with this type of data was conducted. The SEER population structure for Louisiana was used to generate cases that follow a Poisson process with three fixed effects, age, race, and year of diagnosis. 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. Four county group sizes were used, 8, 16, 32, and 64. In addition, each condition was simulated with a cluster effect. That is, a random county and its neighbors were given an elevated risk of disease. This resulted in 24 models. For the 24 models, the power and bias of the estimates were examined as well as the distribution of the scale parameter.

e-mail: lmorgan1@tulane.edu

MODELING INFLUENZA EPIDEMICS

Paola Sebastiani, University of Massachusetts, Amherst
John Tsimikas, University of Massachusetts, Amherst
Ling Wang*, University of Massachusetts, Amherst
Kenneth Mandl, Children's Hospital Informatics Program, Harvard Medical School

Influenza is a contagious disease that affects between 10% and 20% of U.S. residents every year. It is estimated that an average of about 20,000 people per year in the U.S. die from the flu, and 114,000 per year have to be admitted to the hospital as a result of influenza. Early detection of the severity of the epidemics is important as well as the development of monitoring systems to measure the impact that influenza is having on deaths in the U.S. In this work we analyze the time series of deaths for pneumonia and influenza in New England between 1996 and 2002 and we develop a predictive model able to forecast the severity of the epidemics and give early warnings of influenza outbreaks and changing of regime.

e-mail: sebas@math.umass.edu

USE OF THE KAPPA WITH ASYMMETRIC WEIGHTS WHEN EVALUATING AGREEMENT OF CERVICAL CYTOLOGY

Todd G. Nick*, University of Mississippi Medical Center
Zelma Cason, University of Mississippi Medical Center

The kappa index is very popular in the medical literature. During a two-year period from 2000 to 2001, over 6,000 articles were found when searched for the keyword 'kappa' in PubMed. In contrast, the weighted kappa is infrequently presented. During the same two-year period, less than 200 articles were found when searched for the keyword 'weighted kappa'. Even when agreement is evaluated on an ordinal scale, such as with cervical cytologic interpretations, few articles used the weighted kappa index. However, the kappa index treats all disagreements as equally serious. The weighted kappa index is a more appropriate measure since the seriousness of each disagreement could be specified. Various weights have been suggested, including quadratic weights and Cicchetti weights, but most schemes assume a symmetric agreement weight matrix. We consider and discuss the utility of various weighting matrices that are asymmetric. For example, a greater penalty could be given to undercalling a smear than to overcalling a smear. All examples given are from cytologic interpretations of Pap smears.

e-mail: tnick@shrp.umsmed.edu

THE EFFECT OF MISSING DATA IN THE STUDY OF SNPS-TO-PHENOTYPE ASSOCIATION

Min Qin*, Covance Laboratories Inc.
Paola Sebastiani, University of Massachusetts, Amherst

This work uses a new approach to study the effect of missing data in discovering associations between SNPs and a phenotype. We describe this approach on a publicly available database that consists of about 250 families with children affected by type 2 diabetes. For each family member, the SNPs in 13 loci were recorded and the association between the transmission of each SNP and the diseases prevalence were analyzed by means of the TDT test, by disregarding the missing data. The results were published by Altshuler, et al in Nature Genetics 2000. By modeling the pattern of incomplete data for all 13 SNPs, we found two missing data mechanisms, which were used to develop imputation models to fill in the missing data. We then analyzed the completed data by modeling the SNPs-to-phenotype associations by Bayesian networks, and no strong association was found in the completed database. This result allows us to conclude that the associations between the PPAR Pro12Ala polymorphism and type 2 diabetes
hypothesized by the authors of the original paper could actually be a spurious association due to the missing data pattern.

e-mail: min.qin@covance.com

SPECIAL NUMBERS IN SAMPLE SIZE ANALYSIS

Linda J. Goldsmith*, University of Louisville

Statisticians are often called upon for sample size calculations. Certain sample sizes occur often in efforts to achieve statistical power. Sometimes adequate numbers are required to ascertain that asymptotic results will be approximately correct. There are rules of thumb regarding sample sizes for some methods. Some research strategies can involve small sample sizes. This paper reviews some important sample sizes and their justifications.

e-mail: jane.goldsmith@louisville.edu

SAMPLE SIZE CALCULATIONS IN PHASE II CLINICAL TRIALS USING INFORMATIVE CONJUGATE PRIORS

Matthew S. Mayo, University of Kansas Medical Center
Byron J.Gajewski*, University of Kansas Medical Center

There are a number of papers that discuss Phase II clinical trials from a Bayesian perspective. A recent article by Tan and Machin focuses on sample size calculations, in particular, the determination of sample size, using diffuse prior specification, based on the probability that the true response proportion exceeds a pre-specified target. In our article we extend these sample size calculations to include informative prior distributions using various strategies that allow both optimistic and pessimistic researchers direct involvement in the sample size decision making. We select the informative priors via multiple methods determined by the mean, median, or mode of the conjugate prior. These cases can result in varying sample sizes.

e-mail: bgajewski@kumc.edu

LINMOD 4: A PROGRAM FOR GENERAL LINEAR MULTIVARIATE MODELS WITH MISSING DATA

Diane J. Catellier*, University of North Carolina
Keith E.Muller, University of North Carolina

MISSMOD provides accurate test size in small samples for many kinds of Gaussian repeated measures and multivariate data with missing values. The software computes approximate tests (Catellier and Muller, 2000) for General Linear Multivariate Models. Simulations support the conclusion that, in contrast to current mixed model competitors, the methods control test size even with as few as 12 observations for 6 repeated measures and 5% missing data. Assuming data missing at random (MAR), the EM algorithm provides maximum likelihood estimates using all of the available data. The tests generalize standard multivariate and "univariate approach" to repeated measures tests by reducing the error degrees of freedom by replacing the number of independent sampling units by various functions of the numbers of non-missing pairs of responses. The program is based closely on LINMOD (version 3). Source code, an extensive user's guide, and example programs may be obtained free of charge via the Internet.

e-mail: diane_catellier@unc.edu

LINKAGE TESTING OF AFFECTED SUBJECTS WHEN PEDIGREE INFORMATION IS UNCERTAIN OR IGNORED: A MIXTURE MODEL APPROACH

Bobby L. Jones*, University of Pittsburgh, Department of Human Genetics
Daniel E.Weeks, University of Pittsburgh, Department of Human Genetics

We consider the problem when subjects or families have been recruited from a population where it is difficult to measure historic inbreeding or where there are errors in the reported relationships (e.g. half sibs reported as full sibs.) These scenarios reduce the power for detecting linkage. However, marker information from genome-wide scans can replace pedigree structure to test for linkage. This approach also supports an alternate design strategy where pools of typed, affected individuals are analyzed by
ignoring or supplementing pedigree structures.

The linkage analysis is performed using groupings of estimated identity by descent (IBD) sharing. Pairwise IBD probabilities based on the genome-wide scan are calculated. Pairs with similar IBD estimates are grouped using a bivariate normal mixture likelihood to form clusters of similarly related pairs. Our statistic contrasts the likelihood of the observed marker data as a function of two different sharing estimates: in the numerator, the average within-group IBD sharing at the marker of interest; in the
denominator, the within group IBD sharing based on the genome-wide marker information. The statistic is calculated for all the groups taking into account uncertainty in pairwise group membership. In the conventional test, denominator sharing estimates are based on known relationships.

e-mail: bjones@helix.hgen.pitt.edu

SELECTING THE OPTIMAL TRANSFORMATION OF A CONTINUOUS COVARIATE IN COX'S REGRESSION: IMPLICATIONS FOR HYPOTHESIS TESTING

Mamun Mahmud*, Dept of Epidemiology and Biostatistics, McGill University
Michal Abrahamowicz, PhD, Dept of Epidemiology and Biostatistics, McGill University, Division of Clinical Epidemiology, Montreal General Hospital
Karen Leffondré, Division of Clinical Epidemiology, Montreal General Hospital

Many exposures in epidemiological studies have non-linear effects and the problem is to choose an appropriate functional relationship between such exposures and the outcome. One common approach is to investigate several parametric transformations of the covariate of interest, and to select a posteriori the function that fits the data best. However, such approach may result in an inflated type I error. We generated data from Cox's models with different transformations of a continuous covariate and investigated type I error rate and the power of the likelihood ratio test (LRT) corresponding to three different procedures. The type I error rate of the second approach was two times higher than the nominal size. For simple monotone dose-response, the corrected test had similar power as the unconditional approach, while for non-monotone dose-response, it had a higher power. Our results confirm that a posteriori selecting the functional form of the dose-response induces an inflation of type I error. The corrected procedure, which can be applied in a wide range of situations, may provide a good trade-off between type I error and power.

e-mail: mmahmu@p-box.mcgill.ca

Modeling Chronic Stress in Longitudinal Data and the Relation to Heart Disease Development

Imke Janssen*, Department of Preventive Medicine, Rush University
Kelly Karavolos, Department of Preventive Medicine, Rush University
Lynda H. Powell, Department of Preventive Medicine, Rush University

In studies of stress and disease outcome, predictor variables, such as depression, are most frequently measured at baseline and used to predict disease outcome several years later. This approach does not reflect the possible chronicity of the stressor. We take advantage of a longitudinal design with annual repeat assessments to model the pattern of predictors over the follow-up period to account for chronicity, variability and change. We illustrate the efficacy of these methods on a study of 428 women, both African-American and Caucasian, undergoing menopausal transition. In this study, psycho-social variables were assessed yearly over a 5 year period, and outcome variables are sub-clinical disease indicators of heart disease development. We illustrate how the modeling of the enduring effects of a stressor can result in more meaningful associations with disease outcome rather than baseline data only.

e-mail: Imke_Janssen@rush.edu

Index of Participants

INSERT AD FOR DUXBURY


I
B
S
ENAR

Tampa Marriott Waterside Floor Plan