International Conference on
Advances in Interdisciplinary Statistics and Combinatorics
AISC - 2014

A Conference of
The North Carolina Chapter of the American Statistical Association

October 10–12, 2014

The University of North Carolina at Greensboro

Conference Sponsors

The University of North Carolina at Greensboro

The North Carolina Chapter of the American Statistical Association

The American Statistical Association

National Science Foundation, Grant # DMS-1212830

Pearson

RTI International

RHO

Taylor and Francis Group

International Conference on
Advances in Interdisciplinary Statistics and Combinatorics
AISC - 2014

Local Organizing Committee

Ratnasingham Shivaji (Advisor)
Department of Mathematics and Statistics, UNCG

Sat Gupta (Chair)
Department of Mathematics and Statistics, UNCG

Xiaoli Gao
Department of Mathematics and Statistics, UNCG

Scott Richter
Department of Mathematics and Statistics, UNCG

Jan Rychtář
Department of Mathematics and Statistics, UNCG

Shanmugathasan Suthaharan
Department of Computer Science, UNCG

Haimeng Zhang
Department of Mathematics and Statistics, UNCG

International Advisory Committee

Tasos Christofides
University of Cyprus, Cyprus

Frank Coolen
Durham University, UK

B. K. Dass
University of Delhi, India

Angela Dean
Ohio State University

Giancarlo Diana
The University of Padua, Italy

Jean-Franois Dupuy
Institut de Recherche Mathmatique de Rennes et INSA de Rennes, France

Sudhir Gupta
Northern Illinois University

Jonathan Hosking
IBM

Cem Kadilar
Hacettepe University, Turkey

Ben Kedem
University of Maryland

Simos Meintanis
National and Kapodistrian University of Athens, Athens, Greece

Prajneshu
Indian Agricultural Statistics Research Institute, New Delhi, India

John Rayner
The University of Newcastle, NSW, Australia

Jerry Reiter
Duke University

Bimal Roy
Director, Indian Statistical Institute, Kolkata, India

Kunio Shimizu
Keio University, Japan

K. F. Turkman
University of Lisbon, Portugal

Sergio Verdu
Princeton University

Abdus Wahed
University of Pittsburgh

Ricardas Zitikis
University of Western Ontario, Canada

Welcome from the Hosts

On behalf of the North Carolina Chapter of the American Statistical Association and The University of North Carolina at Greensboro, we are very pleased to welcome you to **AISC -2014, the International Conference on Advances in Interdisciplinary Statistics and Combinatorics**. We sincerely hope that your participation in the conference is productive and your stay in Greensboro is a pleasant one.

A project of this magnitude is not possible without liberal support from many quarters. We are very thankful to the NC-ASA executive committee, and the UNCG administration, particularly Dr. Timothy Johnston, Dean of the College of Arts & Sciences and Dr. Ratnasingham Shivaji, Head of the Mathematics and Statistics Department, for their whole-hearted support. We would also like to express our most sincere thanks to the conference secretary Haley Childers, for taking care of all of the logistical issues. We would also like to take this opportunity to thank all of the sponsors for this conference, particularly the **National Science Foundation** whose generous support made it possible for many of the young researchers to attend this conference. We would also like to thank all of the plenary speakers and the session organizers whose contributions made it possible for us to put together such a wonderful academic program.

Hopefully you will go back with very good memories of the conference and will continue to be part of future NC-ASA activities.

Best wishes,

Sat Gupta, UNC Greensboro
Conference Chair

Jerry Reiter, Duke University
NC-ASA President - 2014

Welcome from
Dr. Ratnasingham Shivaji, Head
Department of Mathematics and Statistics
UNCG

On behalf of the Department of Mathematics and Statistics at UNCG, I welcome you to this important event. Actually our Department is home to several other conferences of this type, all federally funded in part. I invite you to visit our department's website at `http://www.uncg.edu/mat/` to know us better.

I know Sat and his co-organizers have put together a rich academic program for the conference covering a wide range of topics. I am particularly impressed with the two workshops on Big Data/Machine Learning, and Mathematical Biology/Game Theory. Big Data/Machine Learning will clearly be the focus of research for many years. Similarly research in Mathematical Biology/Game Theory is a perfect example of how mathematicians, statisticians, biologists, computer scientists, and many others can go outside of their narrow domains and come together on one platform to do meaningful research. Our department is one of the leaders in this area and we have run several NSF funded math/bio programs for the past eight years.

I would also like to take this opportunity to thank the North Carolina Chapter of the American Statistical Association for bringing this conference to our beautiful campus, and for supporting the conference financially too. I assure NC-ASA that they will not be disappointed. I would like to thank all of the other sponsors as well.

We will make every effort to make your stay in Greensboro comfortable but if there is anything specific we can do for you, just let Sat or the conference secretaries know and we will make every effort to take care of it. Please do come back and visit our department for a colloquium talk, or just to say hello. We will be very happy to have you here.

With best wishes,

Ratnasingham Shivaji
Helen Barton Excellence Professor & Head
Department of Mathematics and Statistics, UNCG

Welcome from
Dr. Timothy Johnston, Dean
College of Arts & Sciences
UNCG

On behalf of the College of Arts and Sciences, home of the Department of Mathematics and Statistics at UNCG, I would like to extend a very warm welcome to all of the delegates, particularly those who have come from abroad. I would like to thank the North Carolina Chapter of the American Statistical Association for letting us host this important event. We are hopeful that UNCG will become the permanent home for future AISC conferences. I would also like to thank all of the sponsors of the conference for their financial contributions at a time when everyones budgets are under considerable strain.

I am very proud of our Mathematics and Statistics Department which now hosts several very important NSF-funded annual or biennial events of this type. I know that Professor Sat Gupta and his colleagues have put together a wonderful program that will encourage you to return to UNCG to enjoy this conference in future years also.

With best wishes,

Timothy D. Johnston, Dean
College of Arts & Sciences, UNCG

Welcome from
Dr. Terri Shelton
Vice Chancellor for Research and Economic Development
UNCG

On behalf of The University of North Carolina at Greensboro and Chancellor Linda Brady, I extend a very warm welcome to all of you to our campus. The main theme of the conference, Interdisciplinary Statistics, is very timely since quantitative methods are at the core of research and the ability to employ data analytics across disciplines is key to answering some of the major questions facing scholars today. Our Statistical Consulting Center, housed in the Department of Mathematics and Statistics, is a perfect example of this. SCS supports not only the research activities across campus, but also helps the community at large with statistical issues in a variety of settings. I am particularly pleased that the list of sponsors for this conference includes not just the traditional sponsors like NSF but includes several private entities as well. This public/private partnership leads to asking better questions as well as discovering better solutions. I know this year will bring the same invigorating discussion and exchange of ideas as in years past and invite you to steal a few moments from your conference schedule to explore our campus and our city.

With best wishes,

Terri L. Shelton
Vice Chancellor for Research and Economic Development &
Carol Jenkins Mattocks Distinguished Professor, UNCG

Plenary Speakers: Workshop on Big Data

Dr. Syed Ejaz Ahmed
Dean, Faculty of Mathematics and Science, Brock University, Canada

Ejaz Ahmed is professor of statistics and Dean of Faculty of Math and Science. Before joining Brock, he was a professor and Head of Department of Mathematics & Statistics at the University of Windsor, and University of Regina. Prior to that, he held faculty position at the University of Western Ontario. His area of expertise includes statistical inference, shrinkage estimation, and big data analysis. He has numerous published articles in scientific journals, both collaborative and methodological. Further, he has written several books, edited and co-edited several volumes and special issues of scientific journals. He organized several workshops and many invited sessions. He made numerous invited scholarly presentations in many countries around the globe. Dr. Ahmed serves on the editorial board of many statistical journals (SPL, CSDA, JSCS, JSTP, and others) and review editor for Technometrics. He served as a Board of Director and Chairman of the Education Committee of the Statistical Society of Canada. He was a VP communication for ISBIS. Recently, he served as a member of an Evaluation Group, Discovery Grants and the Grant Selection Committee, Natural Sciences and Engineering Research Council of Canada. He is an elected fellow of American Statistical Association.

Dr. David B. Dunson
Arts and Sciences Distinguished Professor
Department of Statistical Science, Duke University, Durham, NC, USA

David Dunson is Arts and Sciences Distinguished Professor of Statistical Science at Duke University. His research focuses on Bayesian statistical theory and methods motivated by high-dimensional and complex applications, with a particular focus on biomedical applications areas including epidemiology and neuroscience. A methods emphasis is on dimensionality reduction, latent factor models, and nonparametric approaches, particularly for high-dimensional, longitudinal and dynamic data, including images, functions, shapes, networks and other complex objects. Dr. Dunson is a fellow of the American Statistical Association and of the Institute of Mathematical Statistics. He is winner of the 2007 Mortimer Spiegelman Award for the top public health statistician, the 2010 Myrto Lefkopoulou Distinguished Lectureship at Harvard University, the 2010 COPSS Presidents' Award for the top statistician under 41, and the 2012 Youden Award for interlaboratory testing methods. He has over 300 peer reviewed publications.

Dr. Sudipto Guha
Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

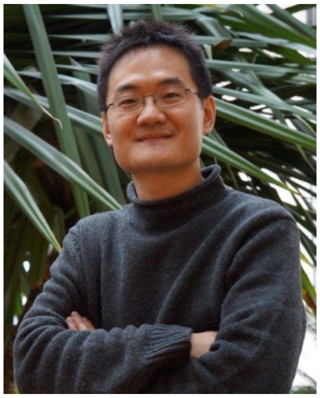
Sudipto Guha is an Associate Professor in the Department of Computer and Information Sciences at University of Pennsylvania since Fall 2001. He completed his Ph.D. in 2000 at Stanford University working on approximation algorithms and spent a year working as a senior member of technical staff in Network Optimizations and Analysis Research department in AT&T Shannon Labs Research. He is a recipient of the NSF CAREER award in 2007, and the Alfred P. Sloan Foundation fellowship.

Dr. Lexin Li
Division of Biostatistics, University of California, Berkeley, CA, USA

Lexin Li obtained B.E. in Electrical Engineering from Zhejiang University, China, in 1998, and Ph.D. in Statistics, from School of Statistics, University of Minnesota, in 2003. He then worked as a Postdoctoral Researcher at School of Medicine, University of California, Davis. Dr. Li joined Department of Statistics, North Carolina State University, in 2005, as an Assistant Professor in Statistics, and was promoted to an Associate Professor in 2011. He was a visiting faculty at Department of Statistics, Stanford University and Yahoo Research Labs from 2011 to 2013 and finally he joined Division of Biostatistics, University of California, Berkeley, as an Associate Professor, in 2014. His research interests include imaging analysis, networks data analysis, dimension reduction, variable selection, computational biology, data mining, and statistical machine learning.

Dr. Jerry Reiter
Mrs. Alexander Hehmeyer Professor of Statistical Science
Department of Statistical Science, Duke University, Durham, NC, USA

Jerome Reiter is the Mrs. Alexander Hehmeyer Professor of Statistical Science at Duke University, and the Associate Director of the Information Initiative at Duke. He received his PhD in statistics from Harvard University in 1999. His primary research interests include methods for protecting data confidentiality, for dealing with missing data, for combining information from multiple sources, and for analyzing complex data. He is the principal investigator of the Triangle Census Research Network (TCRN), one of the nodes of the NSF NCRN network. The TCRN is dedicated to improving the way that statistical agencies disseminate data for secondary analysis.



Dr. Ming Yuan

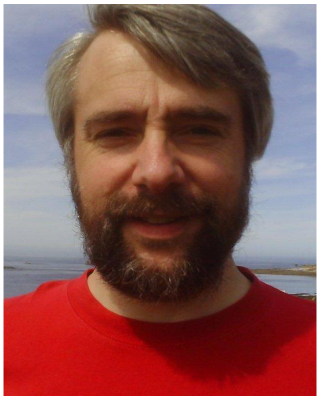
Department of Statistics, University of Wisconsin-Madison & Medical Sciences Center, Madison, WI, USA

Ming Yuan is Professor at Department of Statistics at University of Wisconsin-Madison and Senior Investigator at Morgridge Institute for Research. He was previously Coca-Cola Junior Professor in the H. Milton School of Industrial and Systems Engineering at Georgia Institute of Technology. He received his Ph.D. in Statistics and M.S. in Computer Science from University of Wisconsin-Madison. His main research interests lie in theory, methods and applications of data mining and statistical learning.

Plenary Speakers: Workshop on Mathematical Biology/Game Theory

Dr. Benjamin Allen
Department of Mathematics, Emanuel College, Boston, MA, USA

Benjamin Allen is an Assistant Professor of Mathematics at Emmanuel College in Boston, MA, and a Visiting Scholar at the Center for Mathematical Sciences and Applications at Harvard University. He holds a PhD in Mathematics from Boston University, and did postdoctoral research at the Program for Evolutionary Dynamics at Harvard University. His research is in the mathematical modeling of evolutionary processes. Specific research foci include evolution in spatially structured populations, the somatic evolution of cancer, and microbial evolution.



Dr. Mark Broom
Department of Mathematics, City University London, United Kingdom

Professor Mark Broom obtained a BA (Hons) in Mathematics from the University of Oxford in 1989, followed by an MSc in Statistics (1990) and a PhD in Mathematics (1993) at the University of Sheffield. He then held a postdoctoral research position again at Sheffield (1993-5) and a temporary lecturing position at the University of Glasgow (1995-6). He joined the University of Sussex as a Lecturer in 1996, where he worked until the end of 2009, most recently as Reader in Mathematics, and he was the Head of the Department of Mathematics from 2007 to 2009. Mark was appointed as Professor of Mathematics at City University in January 2010.

Mark's research interests are in Mathematical Biology, in particular Evolutionary Game Theory, and he has published just under 100 papers. This includes both theoretical work on general games and the modelling of specific animal behaviours. Main research areas include multiplayer game theory, models of food stealing (kleptoparasitism), the signalling behaviour of prey species, and evolutionary processes on graphs.

In 2013, together with Jan Rychtář, he completed the book Game-Theoretical Models in Biology published by Chapman and Hall.

He is a member of the following organisations: International Society of Dynamic Games, Higher Education Academy, European Society for Mathematical and Theoretical Biology, International Society for Behavioral Ecology, London Mathematical,Society and the Royal Statistical Society.

Dr. Ross Cressman

Department of Mathematics, Wilfrid Laurier University, Canada

Ross Cressman is Professor of Mathematics and Chair of the Department at Wilfrid Laurier University in Waterloo, Ontario, Canada. He earned his BSc in Mathematics and Physics at the University of Toronto and his PhD in Mathematics at the University of British Columbia. His research has focused on the theory and applications of evolutionary game theory, especially for models that include the effects of individual behavior in biological systems and of human decisions in social science areas such as economics.

He has published two books on the subject. The Solution Concept of Evolutionary Game Theory: A Dynamic Approach (Lecture Notes in Biomathematics, Vol 94, Springer, 1992) gives a dynamic characterization of the evolutionarily stable strategy (ESS) concept introduced for a single species by John Maynard Smith in terms of strong stability under corresponding evolutionary dynamics (e.g. the replicator equation). It also generalizes the ESS approach to biological systems with more than one species.

The second book, Evolutionary Dynamics and Extensive Form Games (MIT Press, 2003) extends evolutionary game theory to extensive form games that model sequential interactions among individuals whose strategy choices at later stages of the interaction may depend on earlier actions. Such games, which traditionally are used to predict rational human behavior through their static equilibrium properties (e.g. subgame perfect Nash equilibria), are now recognized as important predictors of dynamic stability for individual behavior in evolving populations.

For the past ten years, Professor Cressman has expanded his research on evolutionary games in several directions. A series of papers (e.g. R. Cressman, V. Křivan, J. Garay (2004) "Ideal Free Distributions, Evolutionary Games, and Population Dynamics in Multiple-Species Environments", American Naturalist, 164, 473-489; V. Křivan, R. Cressman, C. Schneider (2008) "The Ideal Free Distribution: A Review and Synthesis of the Game-Theoretic Approach", Theoretical Population Biology 73, 403-425) developed the connection between the ESS and the earlier concept of the ideal free distribution (IFD). The IFD assumes species are distributed in a patchy environment in such a way that individuals in all occupied habitats have the same fitness that is at least as high as it would be in any unoccupied habitat. Other, more theoretical, directions analyzing games with continuous strategy spaces and the dynamical consequences of continuously stable strategies (CSS) and neighborhood invader strategies (NIS) are summarized in R. Cressman and Y. Tao ("The Replicator Equation and other Game Dynamics", Proceedings of the National Academy of Sciences USA (PNAS) 2014, 111, 10810-10817).

He is currently active in the worldwide research effort to understand, both in theory and through game experiments, the evolution of cooperation in social dilemmas such as the Prisoner's Dilemma Game and the Public Goods Game (e.g. R. Cressman, J. Song, B. Zhang, Y. Tao (2012) "Cooperation and Evolutionary Dynamics in the Public Goods Game with Institutional Incentives", Journal of Theoretical Biology 299, 144-15; J. Wu, B. Zhang, Z. Zhou, Q. He, X. Deng, R. Cressman, Y. Tao (2009) "Costly Punishment does not always increase Cooperation", PNAS 106, 17448-17451). A recent initiative (R. Cressman, V. Křivan, J. Brown, J. Garay (2014) "Game-Theoretic Methods for Functional Response and Optimal Foraging Behavior", PLoS ONE 9, e88773) shows classical models of optimal foraging can be solved by interpreting predator decisions in an extensive form game-like setting. This direction adds an exciting new dimension to game theory by considering the effect that interaction time has on the evolutionary outcome.

Dr. Vlastimil Křivan
Department of Mathematics and Biomathematics, University of South Bohemia, Czech Republic

Professor Vlastimil Křivan obtained an MSc in Mathematics from the Charles University in Prague in 1982, followed by a PhD in Mathematics (1989) at the Czech Academy of Sciences. Since his graduation he held a research position at the Biology Centre of the Czech Academy of Sciences at České Budějovice where he currently serves as the Head of Department of Ecology and Systematics. Since 2008 he also serves as the Head of the Department of Mathematics and Biomathematics at the Faculty of Science, University of South Bohemia.

Vlastimil's research interests are in applications of mathematics to biology and ecology. In particular, he is interested in links between animal behavior, population dynamics and evolutionary processes that preserve biodiversity on Earth. To this end, he combines game theoretical methods with differential equations describing population dynamics. These methodologies include theory of evolutionarily stable strategies, theory of differential equations with discontinuous vector fields and differential inclusions.

He serves on editorial boards of international journals such as American Naturalist, Dynamic Games and Applications, Theoretical Population Biology and was a member of the board of International Society of Dynamic Games and European Society for Mathematical and Theoretical Biology.

Dr. Alun Lloyd
Mathematics Department, North Carolina State University, Raleigh, NC

Professor Lloyd studied mathematics at Trinity College, Cambridge, before moving to the Department of Zoology in Oxford to do a Ph.D. with Robert May, which he completed in 1996. The following three years in Oxford saw him doing his first postdoc, on a Medical Research Council Non-Clinical Fellowship, and a lectureship at St. Hilda's College. In 1999 he moved to the US, for a four year stint as a Long-Term Member in the Institute for Advanced Study's Program in Theoretical Biology. In 2003 he moved to North Carolina State University to take up a faculty position in the Department of Mathematics and its Biomathematics Graduate Program, which he has been directing since 2007. He has broad interests in mathematical biology, but his main focus is on the transmission dynamics of infectious diseases, particularly mosquito-borne infections.

Dr. Mike Mesterton-Gibbons
Department of Mathematics, Florida State University, Tallahassee, FL, USA

Mike Mesterton-Gibbons grew up in Coventry in the United Kingdom. He graduated with a D.Phil. from the University of Oxford in 1977 before moving to the US in 1982 for a tenure-track position in the Department of Mathematics at Florida State University, where he has been a full professor since 1996. He conducts research on game-theoretic modelling of animal behavior and social structure, on which he has published numerous articles. He is also the author of three texts on modelling and optimization, and an associate editor for Journal of Theoretical Biology. Further biographical details can be found at `http://www.math.fsu.edu/DepartmentNews/Articles/Fac_MG.math`

Dr. Dmitri Zaykin
National Institute of Environmental Health Sciences

Dmitri Zaykin completed a master's-equivalent degree in biology and population genetics from Far Eastern State University, Vladivostok, Russia, and a doctorate in biomathematics from North Carolina State University. He held research positions at the Institute of Marine Biology, Vladivostok, North Carolina State University and GlaxoSmithKline before joining the National Institute of Environmental Health Sciences in 2004, where he is a tenured investigator. His research is at the interface between statistics, mathematics and genetics with current focus on uncovering relationships between genetic variation and phenotypic traits.

Plenary Speakers: AISC 2014

Dr. Ching-Shui Cheng
Institute of Statistical Science, Academia Sinica, and Department of Statistics, University of California, Berkeley, CA, USA

Ching-Shui Cheng is currently a Distinguished Research Fellow and Director of the Institute of Statistical Science, Academia Sinica, in Taiwan, and a Professor Emeritus at the University of California, Berkeley. He received his B.S. in mathematics from National Tsing Hua University and both his MS in mathematics and Ph.D. in mathematics from Cornell University. He joined the Department of Statistics at the University of California, Berkeley in 1977 and retired in 2013.

Dr. Cheng's research interest is mainly in experimental design and related combinatorial problems. He is a fellow of the Institute of Mathematical Statistics and the American Statistical Association and an elected member of the International Statistical Institute. He was an associate editor of the Journal of Statistical Planning and Inference, Annals of Statistics, Statistica Sinica, Biometrika, and Technometrics. He also served as the chair-editor of Statistica Sinica from 1996 to 1999.

Dr. Marie Davidian
William Neal Reynolds Professor
Department of Statistics, North Carolina State University, Raleigh, NC, USA
ASA President 2013

Marie Davidian is William Neal Reynolds Professor in the Department of Statistics at North Carolina State University (NCSU) and Adjunct Professor of Biostatistics and Bioinformatics at Duke University. She received her PhD in Statistics in 1987 from the University of North Carolina at Chapel Hill and is a Fellow of the American Statistical Association (ASA), the Institute of Mathematical Statistics (IMS), and the American Association for the Advancement of Science, and is an elected member of the International Statistical Institute. Marie has served as chair of grant review panels for the National Institutes of Health (NIH), including as chair of the Biostatistical Methods and Research Design study section; as Coordinating and Executive Editor of Biometrics; as a member of US Food and Drug Administration Advisory Committees; and as 2004 president of the Eastern North American Region of the International Biometric Society and 2013 president of the ASA. At NCSU, Marie served as founding director of the Center for Quantitative Sciences in Biomedicine from 2007-2013, and she is Cluster Coordinator for the Chancellor's Faculty Excellence Program's Personalized Medicine Discovery Faculty Cluster.

Marie's interests include analysis of longitudinal data, methods for design and analysis of clinical trials and observational studies, methods for making statistical inference in the presence of missing or mismeasured data, and causal inference and dynamic treatment regimes. She is co-author of the 1995 book ”Nonlinear Models for Repeated Measurement Data” and is co-editor of the 2009 book ”Longitudinal Data Analysis.” She received the 2007 Janet L. Norwood Award for Outstanding Achievement by a Woman in the Statistical Sciences; the 2009 George W. Snedecor and the 2011 F.N. David Awards presented by the Committee of Presidents of Statistical Societies (COPSS); the 2010 NCSU Alexander Quarles Holladay Medal for Excellence, the highest honor the University bestows on a faculty member; and has delivered several distinguished lectures, including a 2010 IMS Medallion Lecture.

Marie is PI or co-investigator for several NIH research grants supporting statistical methods research, and is one of three PIs of a joint P01 Program Project funded by the National Cancer Institute, ”Statistical Methods for Cancer Clinical Trials,” involving faculty from NC State, Duke, and UNC-CH. Since 2004, Marie has served as director of the annual joint NCSU-Duke Clinical Research Institute Summer Institute for Training In Biostatistics (SIBS), a six-week program to inspire US undergraduates to pursue graduate training in biostatistics, funded by the National Heart, Lung, and Blood Institute and the National Center for Advancing Translational Sciences.

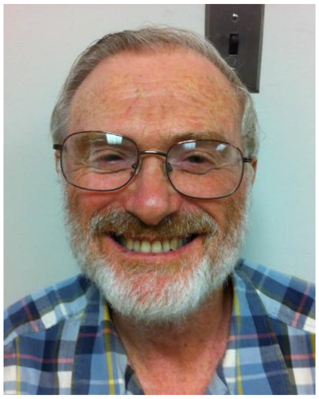
Dr. Amy Herring
Department of Biostatistics, UNC Chapel Hill, USA

Amy Herring is Professor and Associate Chair of Biostatistics in the Gillings School of Global Health at the University of North Carolina at Chapel Hill. Her research interests include statistical methods for studying complex environmental exposures and methods for multivariate and longitudinal data of mixed types. She enjoys collaborating with researchers in public health and was honored to receive the 2012 American Public Health Associations Mortimer Spiegelman Award for outstanding public health statistician under age 40.

Dr. Alan F. Karr
Director of the Center of Excellence for Complex Data Analysis (CoDA)
RTI International, Research Triangle Park, North Carolina, USA

Alan F. Karr is Director of the Center of Excellence for Complex Data Analysis (CoDA) at RTI International. Previously, he was Director of the National Institute of Statistical Sciences (NISS) (2000–14, and Associate Director from 1992 to 2000), Professor of Statistics & Operations Research and Biostatistics at the University of North Carolina at Chapel Hill (1993–2014), and Associate Director of the Statistical and Applied Mathematical Sciences Institute (SAMSI) (2002– 07 and 2011–14). Before coming to North Carolina, he was Professor of Mathematical Sciences and Associate Dean of the School of Engineering at Johns Hopkins University. His research activities are cross-disciplinary and cross-sector collaborations involving statistics and such other fields as data confidentiality, data integration, data quality, data availability, survey methodology, education statistics, healthcare, transportation, aging, the scientific workforce, software engineering, materials science and disease surveillance. He holds one patent, is the author of three books and more than 140 scientific papers, the majority of which have co-authors from disciplines other than statistics, and has been the PI or co-PI on more than $50 million of sponsored research funding.

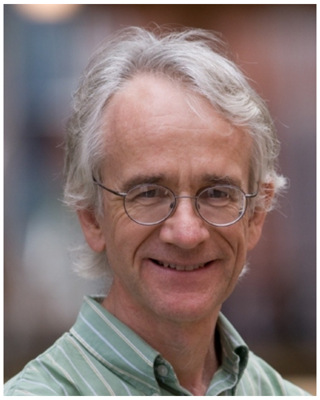
Karr is a Fellow of the American Statistical Association (ASA) and of the Institute of Mathematical Statistics (IMS), an elected member of the International Statistical Institute, and an elected member of the Johns Hopkins Society of Scholars. He currently serves on the Steering Committee of Section U of the American Association for the Advancement of Science, the Executive Advisory Board of the Institute for Advanced Analytics at North Carolina State University and the Committee on Privacy and Confidentiality of the ASA, as well as being Vice-Chair of the Transportation Statistics Interest Group of the ASA. He is an Associate Editor of Statistics and Public Policy.



Dr. Benjamin Kedem
Department of Mathematics, University of Maryland, College Park, MD, USA

Benjamin Kedem is a professor in the Mathematics Department, and an affiliate of the Institute for Systems Research, University of Maryland, College Park. He is a Fellow of the American Statistical Association. He served as the STAT Director, 2000-2004. His areas of expertise include spectral analysis, temporal/spatial prediction, generalized linear models for time series, and semiparametric statistical inference based on integrated data from multiple sources. Among his honors: *a*) 1986: Research on higher order crossings (HOC) selected as an accomplishment by AFOSR; *b*) 1988: IEEE W.R.G. Baker award, given for the most outstanding paper reporting original work in any of the Transactions, Journals, and Magazines of the IEEE Societies or in the PROCEEDINGS of the IEEE; *c*) 1997: NASA/Goddard Exceptional Achievement Award, for "theoretical development and optimization of statistical methods maximizing the accuracy of rainfall measurements for the Tropical Rainfall Measuring Mission(TRMM)"; *d*) 2006: Weston Visiting Professorship, Weizmann Institute of Science, Israel; *e*) 2006: IBM Faculty Award; and *f*) 2013: Fulbright Specialist, New University of Lisbon, Portugal. He served as PI for grants and contracts from AFOSR, ONR, NASA, NCHS, NSF, and USDA.

He is the author of: *a*) Binary Time Series, Dekker, 1980; *b*) Time Series Analysis by Higher Order Crossings, IEEE Press, 1994; *c*) Regression Models for Time Series Analysis (with K. Fokianos), Wiley, 2002.



Dr. Richard Smith
Mark L. Reed III Distinguished Professor
Director of SAMSI
Department of Statistics and Operations Research, UNC Chapel Hill, NC, USA

Richard L. Smith is Mark L. Reed III Distinguished Professor of Statistics and Professor of Biostatistics in the University of North Carolina, Chapel Hill. He is also Director of the **Statistical and Applied Mathematical Sciences Institute (SAMSI)**, a Mathematical Sciences Institute supported by the National Science Foundation. He obtained his PhD from Cornell University and previously held academic positions at Imperial College (London), the University of Surrey (Guildford, England) and Cambridge University. His main research interest is environmental statistics and associated areas of methodological research such as spatial statistics, time series analysis and extreme value theory. He is particularly interested in statistical aspects of climate change research, and in air pollution including its health effects. He is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics, an Elected Member of the International Statistical Institute, and has won the Guy Medal in Silver of the Royal Statistical Society, and the Distinguished Achievement Medal of the Section on Statistics and the Environment, American Statistical Association. In 2004 he was the J. Stuart Hunter Lecturer of The International Environmetrics Society (TIES). He is also a Chartered Statistician of the Royal Statistical Society.

Dr. John Stufken
Charles Wexler Professor in Statistics
School of Mathematical and Statistical Sciences, Arizona State University, USA

After receiving his BS and MS degrees from the University of Nijmegen in the Netherlands, John Stufken received his PhD in Statistics in 1986 from the University of Illinois at Chicago. Since that time, he has held positions as Assistant and Associate Professor at the University of Georgia (1986-1990), Assistant, Associate and Full Professor at Iowa State University (1988-2002), Program Director for Statistics at the National Science Foundation (2000-2003), and Professor and Head of the Department of Statistics at the University of Georgia (2003-2014). Since August 2014 he is the Charles Wexler Professor in Statistics in the School of Mathematical and Statistical Sciences at Arizona State University. His primary area of research interest is design and analysis of experiments. He has approximately 75 publications in this area, and is co-author (with A. Hedayat and N.J.A. Sloane) of the Springer Verlag book Orthogonal Arrays: Theory and Applications. He is also co-Editor (with D. Bingham, A.M. Dean and M.D. Morris) of the forthcoming Handbook of Design and Analysis of Experiments published by Chapman and Hall/CRC. He has served as Associate Editor for many journals, including currently for the Journal of the American Statistical Association and the Journal of Statistical Theory and Practice. From 2004-2006, he served a three-year term as Executive Editor for the Journal of Statistical Planning and Inference. This was followed by a three-year stint as Editor for The American Statistician (2009-2011). He has served in multiple capacities for the ASA, including on three Task Forces appointed by incoming presidents. He was co-organizer (with A. Mandal) of the conference Design and Analysis of Experiments 2012, which was held from October 17-20, 2012, in Athens, GA. John Stufken is an elected Fellow of the IMS (2000) and of the ASA (2001). He held a one-month appointment as the Rothschild Distinguished Visiting Fellow at the Isaac Newton Institute for Mathematical Sciences in Cambridge, UK, in 2011.

Dr. Anastasios A. (Butch) Tsiatis
Gertrude M. Cox Distinguished Professor of Statistics
Department of Statistics, North Carolina State University, Raleigh, NC, USA

Anastasios A. Tsiatis is the inaugural Gertrude M. Cox Distinguished Professor of Statistics at NCSU. Dr. Tsiatis is known internationally for his statistical methods research, particularly in the areas of survival analysis, semiparametric methods, cost and quality of life data analysis, and causal inference. He has published more than 150 papers, most in top statistical journals, and he is a Fellow of the American Statistical Association (ASA) and of the Institute of Mathematical Statistics (IMS) and an elected Ordinary Member of the International Statistical Institute. Dr. Tsiatis is currently co-editor of the journal Biostatistics, and he is a long-standing Editor for the Springer Series on Statistics for Biology and Health. He has received numerous honors and awards, including the 1986 Mortimer Spiegelman Award given by the American Public Health Association Statistics Section to recognize an outstanding statistician under age 40, the 2001 Committee of Presidents of Statistical Societies (COPSS) George W. Snedecor Award, and distinguished lectureships in numerous leading statistics and biostatistics departments. He is the recipient of a prestigious Method to Extend Research In Time (MERIT) award from the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH), which recognizes researchers who have demonstrated superior competence and outstanding productivity in research by awarding them long-term support to foster continued creativity; less than 5% investigators supported by NIH receive MERIT awards. Dr. Tsiatis is also the first statistician to receive the 2010-2011 Princess Lilian Visiting Professorship, which brings an outstanding scientific researcher to Belgium for a series of distinguished lectures; and he was awarded an honorary doctorate in 2013 by the University of Hasselt in Belgium. Dr. Tsiatis received his bachelor's degree in mathematics from the Massachusetts Institute of Technology in 1970 and his PhD in statistics from the University of California, Berkeley, in 1974 under the direction of Jerzy Neyman. He has been a member of the Department of Statistics faculty at NCSU since 1997. Prior to joining NCSU, he was a faculty member in the Department of Biostatistics at Harvard University, where he was promoted to Professor with tenure in 1990; a member of the Biostatistics Division at St. Jude Childrens Research Hospital; and a faculty member in the Department of Statistics at the University of Wisconsin-Madison.

Dr. Mike West
The Arts & Sciences Professor of Statistics & Decision Sciences
Department of Statistical Science, Duke University, Durham, NC, USA

Mike West holds a Duke University distinguished chair as the Arts & Sciences Professor of Statistics & Decision Sciences in the Department of Statistical Science, where he led the development of statistics from 1990-2002. A past president of the International Society for Bayesian Analysis (ISBA), West has served the international statistics profession broadly. This includes founding roles for ISBA and the Bayesian section of the American Statistical Association, among other areas of professional society development. He was a member of establishment committees and founding boards of the National Institute of Statistical Sciences and the Statistical & Applied Mathematical Sciences Institute in the USA, and has served on advisory boards of other national research institutes including the Institute for Mathematical Sciences in UK, and the Institute of Statistical Mathematics in Japan. West's expertise spans a wide range of areas in Bayesian statistical modelling and computation, focusing on methodology and applications of complex stochastic modelling in high-dimensional problems. His current research emphasizes emerging large-scale computation, multivariate time series and dynamic modelling, sparsity modelling in statistics, and spatio-dynamic modelling. Current areas of emphasis in applications include financial time series and decision analysis with increasingly complex, structured models, macro-econometrics and societally relevant big & fast data issues in global financial systems, monitoring and fast data analysis in dynamic networks, advanced imaging and statistical modelling in systems biology applications, large-scale environmental and atmospheric monitoring, and others.

Since his PhD in 1982, West has published around 180 papers in core statistics and interdisciplinary applications in business, econometrics and finance, signal processing, climatology, public health, genomics, immunology, neurophysiology, systems biology and other areas. He is co-author of three books on Bayesian time series analysis and dynamic modelling, several edited volumes, and multiple software packages. He has been a statistical consultant for multiple companies, banks, government agencies and academic centers in UK and USA. In the business world beyond consulting, West co-founded the biotechnology company Expressionanalysis (EA Inc.), now a wholly owned subsidiary of Quintiles, has been advisor or board member of financial investment firms, and is a current member of the board of directors of investment company BEAM Multi-Strategy Master Fund, Ltd., and of the IT energy solutions company Panton Inc.

West has received a number of awards for research and professional service, including the international Mitchell Prize for research in applied Bayesian statistics (in 1994, 1997 and again in 2012), outstanding statistical application paper award of the American Statistical Association (in 1997), the Zellner Medal of ISBA, the International Society for Bayesian Analysis (in 2014), and multiple distinguished speaking awards. West teaches broadly in Bayesian statistics, in academia and through short-courses. His professional kicks come from a variety of sources, but teaching and working with smart, young, emerging statisticians from all over the world is what he prizes most. In addition to working with and advising many undergraduates and Masters students, West has advised more than 50 primary PhD students and postdoctoral associates, most of whom are now in academic, industrial or governmental positions involving advanced statistical research.

Dr. Daniel Zelterman
Professor of Public Health (Biostatistics)
Yale University, New Haven, CT, USA

Dan Zelterman completed his PhD at Yale in the Department of Statistics and went on to academic positions at the University of New York at Albany and University at Minnesota before returning to Yale in the Department of Biostatistics in 1995. His research interests are centered in methods for the analysis of discrete-valued data, including the development of several discrete probability distributions. These have been applied to developing methods in genetics and environmental applications. He serves as Co-Director of the Biostatistics Core at the Yale Comprehensive Cancer Center. In this role he designs Phase I and II clinical trials in new therapies to treat cancer. In his spare time he is an amateur musician, playing oboe and bassoon in several orchestras in the New Haven area.

Conference Program
AISC 2014 - October 10-12, 2014
UNCG

Please Note: At UNCG Campus, shuttles will always drop off and pick up from Sterling Street between Elliott University Center (EUC: Conference Venue) and Walker Parking Deck

October 9, 2012, Thursday

5:00 - 8:00 pm	Conference Registration Desk at Sheraton Hotel

October 10, 2014, Friday

Plenary Sessions 1- 4 on Day 1 (Friday, October 10) are devoted exclusively to the workshops on Big Data/ Machine Learning, and Mathematical Biology/Game Theory.

7:30 - 7:45	Shuttles from Sheraton Hotel to UNCG Departures: 7:30, 7:45 Please don't wait for the last departure, you may have to take a cab or walk
7:30-	Registration Desk, Refreshments, EUC Auditorium Lobby
8:30 - 9:00	Inaugural Remarks - EUC Auditorium
9:00 - 10:30	**Plenary Session 1A: Workshop on Big Data/Machine Learning** EUC Auditorium Chair: **Ming Yuan** - University of Wisconsin-Madison **Jerry Reiter** - Duke University Protecting Confidentiality in an era with no Privacy **Sudipto Guha** - University of Pennsylvania Sampling, Sketching and Streaming: Designing Combinatorial Algorithms for Big Data
9:00 - 10:30	**Plenary Session 1B: Workshop on Mathematical Biology/Game Theory** Kirkland Room Chair: **Jan Rychtář** - UNCG **Mark Broom** - City University London Asymmetric Games in Monomorphic and Polymorphic Populations **Mike Mesterton-Gibbons** - Florida State University Variation between Self- and Mutual Assessment in Animal Contests
10:30 - 10:50	Coffee Break
10:50 - 12:20	**Plenary Session 2A: Workshop on Big Data/Machine Learning** EUC Auditorium Chair: **Lexin Li** - University of California, Berkeley **Ming Yuan** - University of Wisconsin-Madison Detection of Very Short Signal Segments **S. Ejaz Ahmed** - Brock University, Canada Big Data, Big Bias, Small Surprise!

10:50 - 12:20 **Plenary Session 2B: Workshop on Mathematical Biology/Game Theory**
Kirkland Room
Chair: **Mark Broom** - City University London

Vlastimil Křivan - University of South Bohemia
On Lotka-Volterra Population Games

Ross Cressman - Wilfrid Laurier University
Game-Theoretic Methods for Functional Response and Optimal Foraging Behavior

12:20 - 1:50 Lunch - Dining Commons (Two buildings down from EUC, the Conference Center)

1:50 - 3:20 **Plenary Session 3A: Workshop on Big Data/Machine Learning**
EUC Auditorium
Chair: **: J. S. Marron** - UNC Chapel Hill

David Dunson - Duke University
Bayesian Inference on Populations of Networks

Lexin Li - University of California, Berkeley
Two Years in Silicon Valley – Some experiences, stories, and thoughts about Big Data

1:50 - 3:20 **Plenary Session 3B: Workshop on Mathematical Biology/Game Theory**
Kirkland Room
Chair: **Ross Cressman** - Wilfrid Laurier University

Benjamin Allen - Emanuel College & Harvard University
Spatial Population Structure Affects the Rate of Neutral Substitution

Alun Lloyd - North Carolina State University
Modeling Novel Strategies for Controlling Mosquito-Borne Diseases

3:30 - 3:50 Coffee Break

3:50 - 4:20 **Plenary Session 4: Workshop on Mathematical Biology/Game Theory**
Kirkland Room
Chair: **Vlastimil Křivan** - University of South Bohemia

Dmitri Zaykin - National Institute of Environmental Health Sciences
Sorting out Genuine and Spurious Discoveries in Large Scale Genetic Studies

3:50 - 4:20 **NC-ASA Chapter Meeting (non-members also invited)**
EUC Auditorium
Chair: **Jerry Reiter** - Duke University, NC-ASA President

4:30 - 6:30 **Parallel Sessions 1A-2H**

6:45 - 8:30 Reception/Dinner in the Dining Commons

8:30 - 9:00 Shuttles Back to Sheraton
Departures: 8:30, 8:45

October 11, 2014, Saturday

7:30 - 7:45	Shuttles from Sheraton Hotel to UNCG Departures: 7:30, 7:45 Please don't wait for the last departure, you may have to take a cab or walk
7:30 - 8:30	Registration/Refreshments, EUC Auditorium Lobby
8:30 - 10:00	**Plenary Session 5** EUC Auditorium Chair: **Ben Kedem** - University of Maryland **Mike West** - Duke University Sparsity in Multivariate Dynamic Models **Muni Srivastava** - University of Toronto Tests for Covariance Matrices in High Dimension with Less Sample Size
10:00 - 10:20	Coffee Break
10:20 - 12:20	**Parallel Sessions 2A-3H**
12:20 - 1:50	Lunch - Dining Commons
1:50 - 3:20	**Plenary Session 6** EUC Auditorium Chair: **Sujit Ghosh** - SAMSI & North Carolina State University **Richard Smith** - SAMSI & UNC Chapel Hill Influence of Climate Change on Extreme Weather Events **John Stufken** - Arizona State University On Optimal Designs for Nonlinear Mixed Effects Models
3:20 - 3:40	Coffee Break
3:40 - 5:10	**Plenary Session 7** EUC Auditorium Chair: **Dan Zelterman** - Yale University **Marie Davidian** - North Carolina State University Implementing Personalized Medicine: Estimation of Optimal Dynamic Treatment Regimes. Part I: Statistical Framework and Regression-based Estimators **Anastasios Tsiatis** - North Carolina State University Implementing Personalized Medicine: Estimation of Optimal Dynamic Treatment Regimes. Part II: Value Search Estimators and Classification Perspective
5:20 - 7:20	**Parallel Sessions 3A-3G**
7:30 - 9:30	Conference Banquet & Chapter Awards Ceremony, Cone Ball Rooms, EUC
9:30 - 10:00	Shuttles Back to Sheraton Departures: 9:30, 9:45

October 12, 2014, Sunday

7:30 - 7:45	Shuttles from Sheraton Hotel to UNCG Departures: 7:30, 7:45 Please don't wait for the last departure, you may have to take a cab or walk
7:30 - 8:30	Registration/Refreshments, EUC Auditorium Lobby
8:30 - 10:00	**Plenary Session 8** EUC Auditorium Chair: **John Stufken** - Arizona State University **Dan Zelterman** - Yale University A Two-Stage, Phase II Clinical Trial Design with Nested Criteria for Early Stopping and Efficacy **Ben Kedem** - University of Maryland Repeated Out of Sample Fusion in Interval Estimation of Small Tail Probabilities in Food Safety
10:00 - 10:20	Coffee Break
10:20 - 11:50	**Plenary Session 9** EUC Auditorium Chair: **Jerry Reiter** - Duke University **Alan Karr** - RTI International Effects of Method Parameters and Ground Truth in the OMOP Results Database **C. S. Cheng** - Institute of Statistical Science, Academia Sinica & UC Berkeley Minimal Second-Order Saturated Designs
12:20 - 11:50	Lunch -Dining Commons
1:50 - 3:50	**Parallel Sessions 4A-4G**
3:50 - 4:30	Coffee Break
4:30 - 5:00	Best Paper Awards, EUC Auditorium
5:00 - 5:30	Shuttles Back to Sheraton Departures: 5:00, 5:15

Parallel Session 1A: EUC Joyner Room
Recent Advances in Multivariate Theory and Its Application
Chair: Dietrich von Rosen - Swedish University of Agricultural Sciences
Friday, October 10, 4:30 - 6:30

Shinpei Imori, Hiroshima University
Consistent Selection of the True Covariance Components in High-Dimensional Growth Curve Model with Random Coefficients

Francesca Greselin - University of Milano-Bicocca, Italy
Robust Mixtures of Factor Analyzers via the Trimming Approach

Dietrich von Rosen - Swedish University of Agricultural Sciences
Wilks Lambda and the Growth Curve Model

Martin Singull, Linköping University
Test for the Mean in a Growth Curve Model for High Dimensions

Parallel Session 1B: EUC Claxton Room
Foundations of Objective Inference
Chair: Jan Hannig - UNC Chapel Hill
Friday, October 10, 4:30 - 6:30

Kai Zhang - UNC Chapel Hill
Inference after Model Selection Valid Post-Selection Inference

Dungang Liu - Yale University
A Nonparametric Approach to Combining Information from Multiple Sources

Ryan Martin - University of Illinois at Chicago
On Optimal Inferential Models

Abhishek Pal Majumear - UNC Chapel Hill
Higher Order Asymptotics for Generalized Fiducial Distribution

Parallel Session 1C: EUC Maple Room
Statistical Consulting
Chair: Jason Brinkley - East Carolina University
Friday, October 10, 4:30 - 6:30

Jason Brinkley - East Carolina University
Consulting: The Good, The Bad, and The Ugly

Emily Griffith - North Carolina State University
Playing in everyone's Backyard: Statistical Consulting in the Sciences

Eric Vance - Virginia Tech
Training Statistical Consultants to Become Collaborative Statisticians

Julia Sharp- Clemson University
Perceptions of Statistical Consulting in a University Setting

Panel Discussion:
Jason Brinkley (East Carolina University), Emily Griffith (North Carolina State University), Eric Vance (Virginia Tech), Julia Sharp (Clemson University), Scott Richter (UNCG) and Hrishikesh Chakraborty (University of South Carolina)

Parallel Session 1D: EUC Alexander Room
New Techniques for Analyzing High Dimensional and Complex Data
Organizer: Yichao Wu - North Carolina State University
Chair: Donald Martin - North Carolina State University
Friday, October 10, 4:30 - 6:30

Yuexiao Dong - Temple University
Sufficient Dimension Reduction with Multivariate Response

Donald Martin - North Carolina State University
Efficient Computation of the Distribution of Spaced Seed Coverage

Jichun Xie - Duke University
High Dimensional Tests for Functional Brain Networks

Jiajia Zhang - University of South Carolina
Accelerated Intensity Frailty Model for Recurrent Events Data

Parallel Session 1E: EUC Sharpe Room
Data Dimension Reduction Approaches
Chair: Kumer Pial Das- Lamar University
Friday, October 10, 4:30 - 6:30

J. S. Marron - UNC Chapel Hill
Nonnegative Nested Cone Analysis

S. Stanley Young - National Institute of Statistical Sciences
Two-Block Non-Negative Matrix Factorization for Unknown Chemical Identification

Karthik Devarajan - Fox Chase Cancer Center, Temple University
Non-Negative Matrix Factorization Algorithms for Heteroscedastic Data

Kumer Pial Das - Lamar University
Using Text Mining Methods to Analyze Biomedical Data

Parallel Session 1F: EUC Kirkland Room
Spatial Statistics and the Environment: Theory and Applications
Chair: Haimeng Zhang - UNCG
Friday, October 10, 4:30 - 6:30

Haimeng Zhang - UNCG
Covariance Structures of Axially Symmetric Spatial Processes on the Sphere

Yang Li - University of Minnesota, Duluth
Modeling Nonstationary Covariance Function with Convolution on Sphere

Elizabeth Mannshardt - North Carolina State University
Dynamic Spatial Evaluation of Deterministic Air Quality Models Using Monitoring Station Observations

Guoyi Zhang - University of New Mexico
Smoothing Splines Using Compactly Supported, Positive Definite, Radial Basis Functions

Janet S. Kim -North Carolina State University
Generalized Functional Concurrent Model

Parallel Session 1G: EUC Dogwood Room
Financial and Insurance Models
Chair: J. Beleza Sousa - New University of Lisbon
Friday, October 10, 4:30 - 6:30

J. Beleza Sousa - New University of Lisbon
Bonds Historical Simulation Value at Risk

Goncalo dos Reis - University of Edinburg
Securitization and Equilibrium Pricing Under Relative Performance Concerns

Manuel L. Esquvel - New University of Lisbon
Market Liquidity as a Probability or as a Stochastic Process

Ricardas Zitikis - University of Western Ontario
Trends in Disguise

Oleh Stefanets - Kyiv National University of Trade and Economics
Economic and Mathematical Models of Credit Risk Assessment in Commercial Bank

Parallel Session 1H: EUC Auditorium
NoSQL Systems for Big Data Analytics
Chair: Venkat Gudivada - Marshall University
Friday, October 10, 4:30 - 6:30

Jordan Paris - Marshall University
Big Data Scale Content Based Image Retrieval

S. Jothilakshmi - Annamalai University, India
Information Security in NoSQL Systems for Big Data

Dhana Rao - Marshall University
Canonical Discriminate Analysis of Functional Metagenomes

Akhil M. Gudivada - Marshall University
HCUP Data Analysis with MongoDB

Quan Do - New Mexico State University
Modeling Multimodal Healthcare Big Data

Parallel Session 2A: EUC Alexander Room
Clinical Trials Methods: Novel Methods or Applications
Organizer: Heejung Bang - UC Davis
Chair: Sonia Davis - UNC Chapel Hill
Saturday, October 11, 10:20 - 12:20

Zhiwei Zhang - FDA
New Analytical Methods for Non-Inferiority Trials: Covariate Adjustment and Sensitivity Analysis

Hrishikesh Chakraborty - University of South Carolina
Markov Chain Models for Clinical Trials: Intervention to Reduce Human Papillomavirus Infection

Bo Zhang - FDA/CDRH
Estimating Biomarker-Adjusted Treatment Effects in Randomized Clinical Trials Using Auxiliary Covariates: Application to a Phase 3 HIV-1 Trial

Jared Lunceford - Merck & Co
Clinical Utility Estimation for Assay Cut-offs in Early Phase Oncology Enrichment Trials

Bo Jin - Pfizer
Longitudinal Data Analysis to Assess Clinical Efficacy Equivalence in Biosimilar Clinical Trials

Parallel Session 2B: EUC Kirkland Room
Game Theory/Mathematical Biology
Chair: Jan Rychtář - UNCG
Saturday, October 11, 10:20 - 12:20

David Sykes - UNCG
Toxoplasmosis Vaccination Strategies

Ana Paula Tognasoli - UNCG
Game-Theoretical Model of Oviposition Site-Selection of Gravid Female Mosquitoes

David Suarez - UNCG
Greater Mobility and Larger Neighborhood Size Inhibit the Evolution of Cooperation

Aida Briceno - UNCG
The Dynamics of Offensive Messages in the World of Social Media

Jan Rychtar - UNCG
Ideal Cost-Free Distributions in Structured Populations for General Payoff Functions

Fei Xu - Wilfrid Laurier University
Evolution of Mobility in Predator-Prey Systems

Parallel Session 2C: EUC Claxton Room
Applications of Multiple Imputation via Bayesian Mixture Modeling
Chair: Jerry Reiter - Duke University
Saturday, October 11, 10:20 - 12:20

Lan Wei - Duke University
Releasing Synthetic Magnitude Microdata with Fixed Marginal

Nicole Dalzell - Duke University
A Bayesian Approach to Incorporating Uncertainty in Record Linkage

Jingchen Hu - Duke University
Nested Dirichlet Process Model for Household Data Synthesis

Jerry Reiter - Duke University
Imputation of Missing Data via Flexible, Mixture Modeling

Parallel Session 2D: EUC Auditorium
Design and Analysis of Experiments
Organizer: Manohar Aggarwal, University of Memphis
Chair: John Stufken, Arizona State University
Saturday, October 11, 10:20 - 12:20

Angela Dean - The Ohio State University
A-optimal and A-efficient Designs for Discrete Choice Experiments

Pallavi Chitturi - Temple University
Some Results on the Optimality of Choice Sets for 2^n Factorial Designs

Tena Katsaounis - The Ohio State University at Mansfield
Two-Level Screening Designs Derived From Binary Non-Linear Codes

John Morgan - Virginia Tech
Multi-factor Youden Rectangles

Jyotirmoy Sarkar - Indiana University Purdue University Indianapolis
Sudokus as Experimental Designs

Parallel Session 2E: EUC Maple Room
Recent Developments in Statistical Machine Learning
Organizer: Yufeng Liu - UNC Chapel Hill
Chair: Yuexiao Dong - Temple University
Saturday, October 11, 10:20 - 12:20

Xiaoli Gao - UNCG
Robust Feature Selection with Fussy Group Information

Yuying Xie - UNC-Chapel Hill
Estimation of Graphical Model from Noisy Data

Guan Yu - UNC-Chapel Hill
Sparse Learning Incorporating the Graphical Structure among Features

Yin Xia - UNC-Chapel Hill
High Dimensional Sparse MANOVA

Parallel Session 2F: EUC Sharpe Room
Quantitative Methods
Chair: Prajneshu, Indian Agricultural Statistical Research Institute, New Delhi, India
Saturday, October 11, 10:20 - 12:20

Eddy Kwessi - Trinity University
Generalized Signed-Rank Estimation for Nonlinear Models with Multidimensional Indices

B. K. Dass - University of Delhi
Construction of M-Repeated Burst Error Detecting and Correcting Non-Binary Linear Codes

Prithwish Bhaumik - North Carolina State University
Bayesian Estimation in Differential Equation Models

Benedicto Kazuzuru- Sokoine University of Agriculture, Tanzania
Challenges of Statistical Consultancy for Developing Countries

Parallel Session 2G: EUC Joyner Room
Risk Analytics
Chair: Ricardas Zitikis - University of Western Ontario
Saturday, October 11, 10:20 - 12:20

Vytaras Brazauskas - University of Wisconsin-Milwaukee
CATL Methods and Robust Credibility: A Study Using Hachemeister's Data

Bruce Jones - University of Western Ontario
Physiological Age, Health Costs and their Relationship

Danang Teguh Qoyyimi - University of Western Ontario
Measuring the Lack of Monotonicity in Functions

Jaehwan Kim - Korea University Business School
A Direct Utility Model for Economies of Scope

Ricardas Zitikis - University of Western Ontario
Extremal Risks: Paths and Indices of Maximal Tail Dependence

Parallel Session 2H: EUC Dogwood Room
Healthcare Challenges
Chair: Jennifer Fencl - Cone Health System
Saturday, October 11, 10:20 - 12:20

Dawn Engels - Cone Health System
Prevention of Hospital Acquired Pressure Ulcers in an Operating Room Setting

Jennifer Fencl - Cone Health System
Examining Neurosurgical Surgical Site Infections

Prashant Mittal - Muskie School of Public Service, University of Southern Maine
Measuring health outcomes in America using a multivariate approach

Jennifer Fencl - Cone Health System
Intra-Operative Skin Prep: Is There a Difference? A Retrospective Review

Parallel Session 3A: EUC Alexander Room
Clinical Trials: Less Discussed Issues
Chair: Heejung Bang UC Davis
Saturday, October 11, 5:20 - 7:20

Russell Reeve Quintiles
Critique of Trends in Phase II Trials

Guochen Song- Quintiles
Assess the Uncertainty of Probability of Success in Phase III Trial after Phase II Data

Sonia Davis UNC Chapel Hill
Time to Treatment Discontinuation as a Clinical Trial Outcome

Martin Ho FDA
Simulation for Designing Medical Device Trials

Gina-Maria Pomann - North Carolina State University
Case-Control Sampling for Brain Imaging and Enhancement Prediction

Parallel Session 3B: EUC Joyner Room
Mining Text Networks
Chair: David Banks Duke University
Saturday, October 11, 5:20 - 7:20

Christine Chai - Duke University
Pre-analysis of Text Data

Christoph Hellmayr Duke University
Bayesian Models for Text Networks

Shaobo Han Duke University
Discovering Dynamic Topics in Time Stamped Documents

David Banks Duke University
Analysis of a 2012 Political Blog Network

Parallel Session 3C: EUC Auditorium
Design and Analysis of Experiments
Organizer: Manohar Aggarwal, University of Memphis
Chair: John Morgan, Virginia Tech
Saturday, October 11, 5:20 - 7:20

Sanjoy Sinha - Carleton University, Canada
Optimal Designs for Generalized Linear Mixed Models

Miriam Tamm - University Hospital RWTH Aachen, Germany
Chronological Bias Caused By Unobserved Time Trends in Randomized Clinical Trials

Linwei Hu - University of Georgia
Optimal Designs for Multi-Covariate Generalized Liner Models

Hai Ying Wang - University of New Hampshire
Optimal Algorithms for Logistic Regression with Big Data

Ramesh K. Budhraja - S. V. College, University of Delhi, India
Fractional Factorial Designs with some Linear Trend Free Effects using Linear Codes

Parallel Session 3D: EUC Kirkland room
Advances in Mathematical and Statistical Ecology
Chair: Richard Smith, SAMSI & UNC Chapel Hill
Saturday, October 11, 5:20 - 7:20

Christopher Strickland SAMSI & UNC Chapel Hill
Modeling Savanna Water Resource Dynamics with Stochastic Daily Rainfall

Kimberly Kaufeld SAMSI & North Carolina State University
A Spatio-Temporal Model for Predicting Bark Beetle Damage

Daniel Taylor-Rodriguez SAMSI & Duke University
Intrinsic-objective Bayes Variable Selection for Site-Occupancy Models

Richard Smith SAMSI & UNC Chapel Hill
Mathematical and Statistical Ecology Program at SAMSI - Research Opportunities

Parallel Session 3E: EUC Dogwood Room
Econometric Methods in Statistics
Chair: Manuel L. Esquível - New University of Lisbon
Saturday, October 11, 5:20 - 7:20

Sujit Ghosh - SAMSI & North Carolina State University
A Computationally Efficient Flexible Observed Factor Model with Separate Dynamics for the Factor Volatilities and Correlation Matrix

Barry Goodwin - North Carolina State University
Modeling Dependence in Pricing Risk

Prajneshu Indian Agricultural Statistics Research Institute, New Delhi
Autoregressive-Stochastic Volatility with Threshold Nonlinear Time-Series Model

Priya Kohli Connecticut College
Clustering Financial Time Series: A Polyspectral SLEX Approach

Parallel Session 3F: EUC Maple Room
Big Data Challenges

Co-Chairs: Rajeev Agrawal - North Carolina A&T State University
Ashok Kumar, Kentucky State University
Saturday, October 11, 5:20 - 7:20

Rajeev Agrawal - North Carolina A&T State University
A Layer Based Architecture for Provenance in Big Data

Anirudh Kadadi - North Carolina A&T State University
Challenges of Data Integration and Interoperability in Big Data

Christopher Nyamful - North Carolina A&T State University
Big Data Storage and Management

Chitra Reddy Musku UNCG
Evaluation of Low-Dimensional Structures in High-Dimensional Twitter Data

Parallel Session 3G: EUC Sharpe Room
Sampling Methods
Chair: Breda Munoz RTI International
Saturday, October 11, 5:20 - 7:20

Rita Sousa New University of Lisbon, Portugal
Improved Mean Estimation Using Auxiliary Information in Stratified Sampling

Anu Chhabra University of Delhi
Modified Optional Unrelated Question RRT Models

Zaheen Khan - Federal Urdu University of Arts, Science and Technology, Islamabad, Pakistan
Circular and Diagonal Circular Systematic Sampling in the Presence of Linear Trend

Parallel Session 4A: EUC Alexander Room
Dynamic Treatment Regimes and Sequentially Randomized Trials
Chair: Abdus Wahed, University of Pittsburgh
October 12, Sunday, 1:50 - 3:50

Eric Laber - North Carolina State University
Functional Feature Construction for Personalized Treatment Regimes

Yichi Zhang - North Carolina State University
Interpretable Optimal Treatment Regime Using Decision Lists

Jonathan Hibbard - University of North Carolina - Chapel Hill
The Liberti Trial for Discovering A Dynamic Treatment Regimen in Burn Scar Repair

Abdus S. Wahed - University of Pittsburgh
SMART Estimators for SMART Studies

Jerry Q. Cheng - Rutgers University
Using Bootstrap Methods to Estimate Mean Survival Time in Clinical Trials

Parallel Session 4B: EUC Auditorium
Design and Analysis of Experiments
Organizer: Manohar Aggarwal, University of Memphis
Chair: C. S. Cheng - Institute of Statistical Science, Academia Sinica & UC Berkeley
October 12, Sunday, 1:50 - 3:50

Wei Zheng - Indiana University Purdue University Indianapolis

Universally optimal designs for two Interference Models

Nutan Mishra - University of South Alabama
Optimality of Variance Balanced Designs and Their Constructions

Dursun Bulutoglu - Air Force Institute of Technology, Dayton, OH
Finding the Symmetry Group of a Linear Program for Enumerating Orthogonal Arrays

Weng Kee Wong, UCLA
Multiple-Objective Optimal Designs for the 4-Parameter Logistic Model

Satoshi Aoki - Kagoshima University, Japan
Sampling Fractional Factorial Designs by Markov Basis

Parallel Session 4C: EUC Kirkland Room
Clinical Trials/Biostatistics
Organizer: Sudhir Gupta Northern Illinois University
Chair: Xiao Qiang Xue, Quintiles
October 12, Sunday, 1:50 - 3:50

Xiao Qiang Xue, Quintiles
Optimal Design of Clinical Trials with Randomly Censored "Time-To-Event" End Points

Yingqi Zhao - University of Wisconsin Madison
Efficient Augmentation and Relaxation Learning for Treatment Regimes Using Observational Data

Nicholas Meyer - North Carolina State University
An Adaptive Decision Making Algorithm To Manage the Spread of a Disease over a Network

Yunro Chung - University of North Carolina Chapel Hill
Statistical Challenges in Investigating the Effect of Busulfan Delivered by Targeted Pharmacokinetics in Phase I Oncology Trial

Sudip Roy - University of Texas at San Antonio
Autoregressive Methodology in Detection of Irregular Spiking Activity in Neurons

Parallel Session 4D: EUC Maple room
High Dimensional Time Series Data Analysis
Chair: Xiaoli Gao UNCG
October 12, Sunday, 1:50 - 3:50

Kei Hirose - Osaka University, Japan
Penalized Likelihood Factor Analysis

Ying-Ju Chen - Bowling Green State University
Jackknife Empirical Likelihood Based Detection Procedure for Change-Point in Mean Residual Life Functions

Kun Chen - University of Connecticut
On Some Low-Rank Models in Multivariate Time Series Analysis

Elizabeth M. Sweeney - Johns Hopkins Bloomberg School of Public Health
Analysis of Multi-sequence Time Series Data from MS Lesions on Structural MRI

Parallel Session 4E: EUC Sharpe Room
Sampling Theory and Methods
Chair: Javid Shabbir Quaid-I-Azam University, Islamabad, Pakistan
October 12, Sunday, 1:50 - 3:50

Javid Shabbir Quaid-I-Azam University
On a Generalized Class of Mean Estimators in Survey Sampling

G. N. Singh Indian School of Mines, Dhanbad
Improved Exponential Methods Of Imputation In Estimation Of Current Population Mean In Two-Occasion Successive Sampling

Geeta Kalucha University of Delhi, India
A Regression Estimator for Finite Population Mean of a Sensitive Variable

Jeong Sihm UNCG
A Modified Binary Optional RRT Model

Parallel Session 4F: EUC Joyner Room
Teaching & Research Innovations for Undergraduates

Organizers: Jan Rychtář (UNCG),
Kumer Pial Das (Lamar University), and
Meenakshi Prajneshu (University of Delhi)
October 12, Sunday, 1:50 - 3:50

Chad Awtrey- Elon University
One Approach to Researching, Presenting, and Publishing with Undergraduate Pure Math Majors

Jan Rychtář UNCG
Mentoring Interdisciplinary Research Projects

Meenakshi Prajneshu Deshbandhu College, University of Delhi
Undergraduate Research at Delhi University, India: An Overview

Kumer Pial Das Lamar University
Student Views on the Value of Undergraduate Research

Hyunju Oh Bennett College, Greensboro
Co-Directing from Two Institutions for the National Research Experience for Undergraduates Program

Ashok Kumar - Kentcuky State University
Quantitative Analysis of Effectiveness of Teaching an Online Vs In-Class Course

Parallel Session 4G: EUC Dogwood Room
Cluster-Based Statistical Methodologies in Community-Based Interventions

Organizer: Roy T. Sabo - Virginia Commonwealth University
October 12, Sunday, 1:50 - 3:50

Robert E. Johnson - Vanderbilt University
Cluster-Based Methodologies in Community-Based Interventions: An Overview

Scarlett Bellamy - University of Pennsylvania
Randomization Challenges in Couple-Based Randomization Trials: An Application to a Couple-Based, Clustered HIV Behavioral Modification Intervention Study

Roy T. Sabo - Virginia Commonwealth University
Using Meta-Analysis to Obtain Desired Random Effect Structures in Cluster-Randomized Studies

N. Rao Chaganty Old Dominion University
Likelihood Model and Parameter Estimation for Clustered Ordinal Data

Abstracts of the talks

A Layer Based Architecture for Provenance in Big Data

Rajeev Agrawal
North Carolina A & T State University
`ragrawal@ncat.edu`
Coauthors: Ashiq Imran

Big data is a new technology wave that makes the world awash in data. Various organizations accumulate data that are difficult to exploit. Government databases, social media, healthcare databases etc. are the examples of that Big data. Big data covers absorbing and analyzing huge amount of data that may have originated or processed outside of the organization. Data provenance can be defined as origin and process of data. It carries significant information of a system. It can be useful for debugging, auditing, measuring performance and trust in data. Data provenance in big data is relatively unexplored topic. It is necessary to appropriately track the creation and collection process of the data to provide context and reproducibility. In this paper, we propose an intuitive layer based architecture of data provenance and visualization. In addition, we show a complete workflow of tracking provenance information of big data.

Big Data, Big Bias, Small Surprise!

S. Ejaz Ahmed
Brock University, Canada
`sahmed5@brocku.ca`
Coauthors: Xiaoli Gao - UNC Greensboro

In high-dimensional statistics settings where number of variables is greater than observations, or when number of variables are increasing with the sample size, many penalized regularization strategies were studied for simultaneous variable selection and post-estimation. However, a model may have sparse signals as well as with number predictors with weak signals. In this scenario variable selection methods may not distinguish predictors with weak signals and sparse signals. The prediction based on a selected submodel may not be preferable in such cases. For this reason, we propose a high-dimensional shrinkage estimation strategy to improve the prediction performance of a submodel. Such a high-dimensional shrinkage estimator (HDSE) is constructed by shrinking a ridge estimator in the direction of a candidate submodel. We demonstrate that the proposed HDSE performs uniformly better than the ridge estimator. Interestingly, it improves the prediction performance of given candidate submodel generated from most existing variable selection methods. The relative performance of the proposed HDSE strategy is appraised by both simulation studies and the real data analysis.

A New Estimator of The Population Mean: An Application to Bioleaching Studies

Amer Ibrahim Al-Omari
Al al-Bayt University, Jordan
`alomari_amer@yahoo.com`
Coauthors: Carlos N. Bouza, Dante Covarrubias, Roma Pal

In this paper, multistage balanced groups ranked set samples (MBGRSS) method is considered for estimating the population mean for samples of size where is a real integer. The MBGRSS is compared with the simple random sampling (SRS) and ranked set sampling (RSS) schemes. It turnout that for the symmetric distributions considered in this study, the MBGRSS estimator is unbiased estimator of the population mean and it is more efficient than SRS and RSS methods based on the same number of measured units. In addition, the efficiency of MBGRSS is increasing in s for fixed value of the sample size, where s is the number of stages. For non symmetric distributions considered in this paper, the MBGRSS estimator is biased. The methods is applied in a study of bioleaching. It appeared as more efficient than RSS and SRS.

Spatial Population Structure Affects the Rate of Neutral Substitution

Benjamin Allen
Emmanuel College
`allenb@emmanuel.edu`
Coauthors: Christine Sample, Yulia A. Dementieva, Ruben Medeiros, Christopher Paoletti, and Martin A. Nowak

Over time, a population acquires neutral genetic substitutions as a consequence of random drift. A famous result in population genetics asserts that the rate, K, at which these substitutions accumulate in the population coincides with the mutation rate, u, at which they arise in individuals: $K = u$. This identity enables genetic sequence data to be used as a "molecular clock" to estimate the timing of evolutionary events. While the molecular clock is known to be perturbed by selection, it is thought that $K = u$ holds very generally for neutral evolution. Here we show that asymmetric spatial population structure can alter the molecular clock rate for neutral mutations, leading to either $K < u$ or $K > u$. Deviations from $K = u$ occur because mutations arise unequally at different sites and have different probabilities of fixation depending on where they arise. If birth rates are uniform across sites, then $K \leq u$. In general, K can take any value between 0 and Nu. Our model can be applied to a variety of population structures. In one example, we investigate the accumulation of genetic mutations in the small intestine. In another application, we analyze over 900 Twitter networks to study the effect of network topology on the fixation of neutral innovations in social evolution.

Sampling Fractional Factorial Designs by Markov Basis

Satoshi Aoki
Kagoshima University, Japan
`aoki@sci.kagoshima-u.ac.jp`

Fractional factorial designs of several controllable factors are also expressed as binary contingency tables, i.e., the contingency tables whose entries are the indicator functions of the designs. For given marginal totals of the binary contingency tables, Markov chain can be constructed by Markov basis. This approach is first introduced by Fontana, Rapallo and Rogantin (2013) for the saturated designs. In this paper, we consider sampling methods of fractional factorial designs by Markov basis of the binary contingency tables. Because the entries of the binary contingency tables are 0 or 1, the results on the Markov basis of the bounded tables is important. In this paper, we show the fundamental results on the relations between the marginal totals, confoundings and regular fractional factorial designs.

One Approach to Researching, Presenting, and Publishing with Undergraduate Pure Math Majors

Chad Awtrey
Elon University
`cawtrey@elon.edu`

The goal of this talk is to describe the speaker's recent collaborations with undergraduates on projects related to computational Galois theory. Included are discussions of the following: recruiting students, funding sources, dissemination outlets, a description of a recent project, and a summary of students' achievements and post-baccalaureate plans.

Text Mining in Dynamic Blog Networks

David Banks
Duke University
`banks@stat.duke.edu`

Many applications (the Internet, Wikipedia) concern networks of documents. We mine the corpus that consists of all U.S. political blog posts in 2012. The intent is to use recent advances in dynamic network modeling to improve the topic discovery, and recent research on text mining to improve the network modeling. We describe an analysis based on the subset of blog posts that concern the shooting of Trayvon Martin, and then a full analysis of the entire corpus, at a lower level of resolution.

Randomization Challenges in Couple-Based Randomization Trials: An Application to a Couple-Based, Clustered HIV Behavioral Modification Intervention Study.

Scarlett Bellamy
University of Pennsylvania
`bellamys@mail.med.upenn.edu`

We present an algorithm for randomizing units in blocks/clusters for controlled trials when the composition of blocking factors is not known in advance. For example, suppose the desired goal of an intervention study is to randomize units to one of two interventions while blocking on a dichotomous factor (e.g., gender), but the total number of units, and therefore number or composition, of males and females among those units assembled for randomization cannot be determined in advance. This situation arises in randomized trials when subjects are scheduled to be randomized as a group, but not all of the subjects show up for the visit. Since investigators do not know which of the scheduled subjects will or will not show up, a dynamic randomization scheme is required to accommodate the unknown composition of the blocking factor once a group of subjects (units) is assembled for randomization. These settings are further complicated when there is more than one blocking factor. In this paper, we present an algorithm that ensures the integrity of the randomization process in these settings.

Bayesian Estimation in Differential Equation Models

Prithwish Bhaumik
North Carolina State University
`pbhaumi@ncsu.edu`
Coauthors: Subhashis Ghosal (North Carolina State University)

Ordinary differential equations (ODEs) are used to model dynamic systems appearing in engineering, physics, biomedical sciences and many other fields. These equations contain unknown parameters, say θ of physical significance which have to be estimated from the noisy data. Often there is no closed form analytic solution of the equations and hence we cannot use the usual non-linear least squares technique to estimate the unknown parameters. There is a two step approach to solve this problem, where the first step involves fitting the data nonparametrically. In the second step the parameter is estimated by minimizing the distance between the nonparametrically estimated derivative and the derivative suggested by the system of ODEs. The statistical aspects of this approach have been studied under the frequentist framework. We consider this two step estimation under the Bayesian framework. The response variable is allowed to be multidimensional and the true mean function of it is not assumed to be in the model. We induce a prior on the regression function using a random series based on the B-spline basis functions. We establish the Bernstein-von Mises theorem for the posterior distribution of the parameter of interest. Interestingly, even though the posterior distribution of the regression function based on splines converges at a rate slower than $n^{-1/2}$, the parameter vector θ is nevertheless estimated at $n^{-1/2}$ rate.

CATL Methods and Robust Credibility: A Study Using Hachemeister's Data

Vytaras Brazauskas
University of Wisconsin-Milwaukee
`vytaras@uwm.edu`
Coauthors: Harald Dornheim (KPMG Switzerland)

Two recent papers by Dornheim and Brazauskas (2011a,b) introduced a new likelihood-based approach for robust-efficient fitting of mixed linear models and showed that it possesses favorable large- and small-sample properties which yield more accurate premiums when extreme outcomes are present in the data. In particular, they studied regression-type credibility models that can be embedded within the framework of mixed linear models for which heavy-tailed insurance data are approximately log-location-scale distributed. The new methods were called corrected adaptively truncated likelihood (CATL) methods. In this talk, we build upon that work and further explore how CATL methods can be used for pricing risks. For illustrative purposes, we use a well-studied Hachemeister's data of bodily injury claims. The process of outlier identification, the ensuing model inference, and related issues are thoroughly investigated on the featured data set. Performance of CATL methods is compared to that of other robust regression credibility procedures.

The Dynamics of Offensive Messages in the World of Social Media

Aida Briceno
UNCG
`anbricen@uncg.edu`
Coauthors: Krystal Blanco, Javier Tapia, Andrea Steele, John McKay, Sherry Towers, Kamuela Yong

The 21st century has redefined the way we communicate, our concept of individual and group privacy, and the dynamics of acceptable behavioral norms. The messaging dynamics on Twitter, an internet social network, has opened new ways/modes of spreading information. As a result, cyberbullying or in general, the spread of offensive messages, is a prevalent problem. The aim of this report is to identify and evaluate conditions that would dampen the role of cyberbullying dynamics on Twitter. We present a discrete-time non-linear compartmental model to explore how the introduction of a Quarantine class may help to hinder the spread of offensive messages. We based the parameters of this model on recent Twitter data related to a topic that communities would deem most offensive, and found that for Twitter a level of quarantine can always be achieved that will immediately suppress the spread of offensive messages, and that this level of quarantine is independent of the number of offenders spreading the message. We hope that the analysis of this dynamic model will shed some insights into the viability of new models of methods for reducing cyberbullying in public social networks.

Consulting: The Good, The Bad, and The Ugly

Jason Brinkley
East Carolina University
`brinkleyj@ecu.edu`

It is often the case that statisticians find large deviations from statistical theory and unique problems in the world of statistical consulting. Consultants routinely deal with a wide range of clients who have varying degrees of statistical acumen and mathematical capacity. This talk will highlight a series of case studies of real life incidents where consulting clients have clearly deviated from the traditional textbook data collection and analysis. What happens when physicians decide upon themselves to violate the protocol of a clinical study? How does a analyst deal with survey data when seemingly unnecessary questions are answered with surprising results? How do you do analysis when the outcome of interest is in fact missing data? These are just a some of the examples of daily issues that present to the consulting statistician who must manage both the use and misuse of statistics.

Statistical Consulting - Panel Discussion

Jason Brinkley
East Carolina University
`brinkleyj@ecu.edu`
Coauthors: Emily Griffith (NCSU) Eric Vance (Virginia Tech) Julia Sharp (Clemson) Scott Richter (UNCG) Rishi Chakraborty (USC)

Consulting in the Academic Setting - Panel discussion among 6 universities.

Asymmetric Games in Monomorphic and Polymorphic Populations

Mark Broom
City University London, UK
`Maek.Broom@city.ac.uk`
Coauthors: Jan Rychtář, University of North Carolina at Greensboro

Evolutionary game theory is an increasingly important way to model the evolution of biological populations. Many early models were in the form of matrix games, or bi-matrix games in asymmetric situations when individuals occupy distinct roles within the contest, where rewards are accrued through independent contests against random members of the population. More recent models have not had the simple linear properties of matrix games, and more general analysis has been required. In this talk we consider a general analysis of asymmetric games, comparing monomorphic and polymorphic populations. We are particularly interested in situations where the strategies that individuals play influence which role that they occupy, for example in a more realistic variant of the classical Owner-Intruder game. We both show general results and consider specific examples to illustrate the difficulties of these more complex games.

Fractional Factorial Designs with some Linear Trend Free Effects using Linear Codes

Ramesh K. Budhraja
University of Delhi, India
`rkbudhraja@yahoo.com`

The paper presents a simple method to construct fractional factorial designs with s levels($s \gg 2$) having some linear trend free effects (main and two factor interactions) using the parity check matrices of linear codes. Further, the method is used to generate the blocked fractional factorial and mixed fractional factorial designs with some linear trend free effects.

Likelihood Model and Parameter Estimation for Clustered Ordinal Data

Rao N. Chaganty
Old Dominion University
`rchagant@odu.edu`
Coauthors: Dr. Raghavendra Kurada (SAS Institute) Dr. Roy Sabo (Virginia Commonwealth University)

In this paper we will discuss a latent variable likelihood methodology for clustered ordinal data using Gaussian copula with probit and logit link functions. We derive the score functions and simplified expressions for the Hessian matrices, which allow easy computation of the standard errors for the marginal regression parameter estimates as well as the dependence parameters. Through asymptotic relative efficiency calculations we demonstrate that these likelihood estimators are superior as compared to estimators arising from previously established estimating equation approaches. We apply this likelihood-based methodology in an analysis of two real-life data examples using an R package developed specifically for the likelihood estimation.

Pre-analysis of Text Data

Christine Chai
Duke University
`christine.chai@duke.edu`

It is now becoming common to analyze text as data. But text requires different kinds of preparation than is needed for numerical data. This talk describes issues in scraping data from on-line sources, then tokenizing it, and finally n-gramming it. Popular methods for text analysis rely upon bag-of-words models, and this loses semantic information, especially negation, but proper processing can recover some of this. We also describe methods for reducing the number of tokens, to expedite computation. These ideas are illustrated in the context of mining a corpus of political blogs.

Markov Chain Models for Clinical Trials: Intervention to Reduce Human Papillomavirus Infection

Hrishikesh Chakraborty
University of South Carolina
`rishic@mailbox.sc.edu`

In randomized clinical trials and comparative studies of recurrent infection, researchers often collect longitudinal data on disease status of subjects from different groups to compare them and determine the intervention/drug effect. Usually, the comparisons are made with respect to infection rate during follow-up, time to new infection, clearance rate of existing infection, and time to clearance of existing infection. At any follow-up visit, an individual may be infected or uninfected, and may maintain or change infection status in subsequent visits. Researchers often use different survival type models for estimation and comparison using a subset of the data. By not using information from all enrolled subjects, they are then prohibited from claiming that the analysis follows the intention-to-treat property. To address this problem, we used a discrete-time two-state homogenous Markov chain model and a time homogeneous multistate continuous Markov model with related asymptotic distribution to analyze recurrent infection data for an intention-to-treat type analysis. This methodology was used to compare intervention and control groups in a longitudinal Human Papillomavirus (HPV) trial in Kenya and HPV trial at the University of South Carolina.

On Some Low-Rank Models in Multivariate Time Series Analysis

Kun Chen
University of Connecticut
`kun.chen@uconn.edu`
Coauthors: Kung-Sik Chan

Reduced-rank vector autoregressive (VAR) models are commonly used in analyzing multivariate time series data. In practice, finding a subset VAR model that parsimoniously approximates the underlying process could be more important than only determining the VAR order. Moreover, the variable association characterized by the low-rank structure can be refined by incorporating sparse estimation, assuming that each latent association only involves a subset of the variables and their lags. We thus propose to use a sparse and low-rank regularized regression approach to conduct the VAR model analysis, for simultaneous order selection, rank reduction and variable selection. It is shown that, in the context of stationary time-series data, the proposed approach is able to correctly identify both the rank and the sparse structure with probability one asymptotically. We also discuss several extensions of low-rank models, regarding robust estimation/outlier detection, functional data and high dimensional inference.

Jackknife Empirical Likelihood Based Detection Procedure for Change-Point in Mean Residual Life Functions

Ying-Ju Chen
Bowling Green State University
`yingc@bgsu.edu`
Coauthors: Dr. Wei Ning and Dr. Arjun K. Gupta

Mean residual life function is an important function in survival analysis which describes the expected remaining lifetime given survival time up to a certain age. In this talk, I will introduce the nonparametric method based on Jackknife empirical likelihood through U-statistic to test the change point of mean residual life functions of independent random variables. The test statistic and its asymptotic distribution are investigated. Monte Carlo Simulations under different lifetime settings are carried out to show the power performance and the control of Type I error. The real data example is analyzed to illustrate the testing procedure.

Minimal Second-Order Saturated Designs

Ching-Shui Cheng
Institute of Statistical Science, Academia Sinica, and University of California, Berkeley
`cheng@stat.berkeley.edu`

This conference is about advances in statistics and combinatorics. I shall discuss two related topics, minimal second-order saturated designs and minimal 1-saturating sets in statistical design of experiments and finite projective geometry, respectively. Under a 2^{n-p} fractional factorial design, in addition to the main effects, there are $2^{n-p} - 1 - n$ degrees of freedom available for estimating interactions. A design is said to be second-order saturated (SOS) if all these degrees of freedom can be used to estimate two-factor interactions. An SOS design is minimal if it is no longer SOS whenever an arbitrary factor is deleted. All SOS designs can be obtained by adding factors to minimal ones. Minimal SOS designs are important for studying the structures and construction of factorial designs and are also useful for constructing space-filing designs for computer experiments.

Using Bootstrap Methods to Estimate Mean Survival Time in Clinical Trials

Jerry Q. Cheng
Rutgers University
`jcheng1@rwjms.rutgers.edu`
Coauthors: Javier Cabrera and John B. Kostis

In clinical trials, we often need to assess performance of several types of treatments by comparing their outcomes. The gain in life expectancy is one of the performance measures of interest because it has a simple interpretation. This quantity can be easily estimated by the area between estimated survival curves given by the treatment and placebo or for two comparative treatments. However, it is not less straightforward to quantify the variability of such an estimate. In this presentation, we use Bootstrap methods to obtain confidence intervals and conduct statistical testing. Simulation studies are conducted to evaluate the effectiveness of the methods against the asymptotic results under different scenarios (i.e., sample sizes, censoring rates, bootstrap resampling methods, confidence intervals). We apply the methods to a real clinical trial data set.

Modified Optional Unrelated Question RRT Models

Anu Chhabra
University of Delhi, india
`a.chhabra02@gmail.com`
Coauthors: B. K. Dass, University of Delhi

Sihm et al. (2014) introduced modified Unrelated Question Optional RRT Models for simultaneously estimating the mean of a sensitive variable and the sensitivity level of the underlying sensitive question without using a split-sample approach. In this study, we carry that work forward in the context of two-stage RRT models the T-Model proposed by Mangat and Singh (1990) and the F-Model proposed by Mehta et al. (2012). Extensive computer simulations are used to compare efficiencies of the one-stage and two-stage optional unrelated question RRT models

Some Results on the Optimality of Choice Sets for 2^n Factorial Designs

Pallavi Chitturi
Temple University
`chitturi@temple.edu`
Coauthors: Jing Xiao

Choice-based conjoint experiments are used when choice alternatives can be described in terms of attributes. The objective is to infer the value that respondents attach to attribute levels. This method involves the design of profiles on the basis of attributes specified at certain levels. Respondents are presented sets of profiles called choice sets, and asked to select the one they consider best. Information Per Profile (IPP) is used as an optimality criteria to compare designs with different numbers of profiles. The optimality of connected main effects plans based on two consecutive choice sets, Sl, Sl+1 has been examined in the literature. However, the optimality of non-consecutive choice sets has not been examined. In this paper we examine the IPP of both consecutive and non-consecutive choice sets and show that IPP can be maximized under certain conditions. Further, we show that non-consecutive choice sets have higher IPP than consecutive choice sets for $n > 4$. Finally, we examine the optimality of connected first-order-interaction designs based on three choice sets and show that non-consecutive choice sets have higher IPP than consecutive choice sets under certain conditions.

Statistical Challenges in Investigating the Effect of Busulfan Delivered by Targeted Pharmacokinetics in Phase I Oncology Trial

Yunro Chung
University of North Carolina at Chapel Hill
`yunro@live.unc.edu`
Coauthors: Anastasia Ivanova, Jason P. Fine

For some oncology agents dosing patients according to mg/m^2 might lead to high toxicity rates or lower efficacy and dosing with respect to the area under the curve (AUC) estimated by pharmacokinetic analysis is preferred. We described challenges we encountered in analyzing data from a Phase I trial investigating Busulfan in patients undergoing allogeneic hematopoietic cell transplantation. The first challenge was modeling the relationship between relapse free survival, death from relapse and death from other causes and AUC. The Cox regression cannot be used because assumption of independent censoring is violated. Therefore we used a competing risk regression model. The second challenge was to have a flexible model for time to even as a function of AUC. A natural assumption is that the failure rate is decreasing with AUC. We developed a nonparametric method of estimating the shape restricted hazard function. We illustrate our approaches by analyzing the Phase I Busulfan data.

Game-Theoretic Methods for Functional Response and Optimal Foraging Behavior

Ross Cressman
Wilfrid Laurier University, Canada
`rcressman@wlu.ca`
Coauthors: Vlastimil Krivan, Joel Brown and Jozsef Garay

The number of prey of a given type consumed by a predator per unit time (i.e. the functional response) depends on the spatial distribution of the prey as well as on the decisions taken by the predator (e.g. where to search for prey, which type of prey to pursue, whether to attack, etc.) and on the times taken for these activities relative to their expected nutritional value gained. I will discuss (i) how this information can be conveniently represented in a tree diagram similar to an extensive form game; (ii) how standard functional responses such as Holling II emerge naturally from this setting; (iii) how game-theoretic methods can be used to analyze the optimal foraging strategy of the predator. The methods will be illustrated for classical diet choice and patch choice models, including those that involve the effects of simultaneously encountering different types of prey and of prey recognition effects. They will also be applied to predict equilibrium foraging behavior of a predator population with intraspecific strategies taken from a Hawk-Dove game that includes interaction times.

A Bayesian Approach to Incorporating Uncertainty in Record Linkage

Nicole Dalzell
Duke University
`nmd16@stat.du`
Coauthors: Jerry Reiter, Ph.D

Data on a single individual or entity are often available from many different sources. Researchers seeking to combine information from such diverse sources must identify which records in one file correspond to the same individual in another. The difficulty of this task is compounded when confidentiality concerns mandate that records are stored without a unique identifier. Record linkage is the process of comparing information in such de-identified records in order to determine groups of records which correspond to the same entity and create a final linked data set to be used for inference. The linking process is often hindered by the possibility of errors in the record fields used for comparison. This work presents a Bayesian method for record linkage which accounts for potential errors in the linking fields while incorporating the uncertainty of

the matching process into inference conducted on the final linked data set. An example with social science data is presented to illustrate the method.

Using Text Mining Methods to Analyze Biomedical Data

Kumer Pial Das
Lamar University
`kumer.das@lamar.edu`
Coauthors: Chelsea Boling

The primary objective of this study is to use two statistical techniques in text mining to discover patterns and trends in biomedical literature. Latent Semantic Analysis (LSA) discovers a data representation that has a lower dimension than the original semantic space by singular value decomposition. Similarly, Non-negative Matrix Factorization (NMF) discovers two non-negative matrices whose product can approximate the original matrix. Essentially, both techniques reduce the dimensionality of given data to preserve the most crucial aspects. Selecting an appropriate dimension for a reduced representation maximizes the performance of an information retrieval system. We illustrate the effectiveness of each method and analyze the latent relationships of the reduced representation of 3,398 full-text articles in the Open Access subset of PubMed Central data. By revealing meaningful relationships between terms and documents, the goal is to overcome the problem of unsatisfactory indexed results and to ultimately discover new knowledge based on the diversification these techniques entail.

Student Views on the Value of Undergraduate Research

Kumer Pial Das
Lamar University
`kumer.das@lamar.edu`

Research experience is crucial for an undergraduate students professional and personal development, academic success, and career preparation. Educational Institutions around the country are encouraging and stimulating undergraduate research programs. This study examines student views on the value of undergraduate research. To understand student views several groups of undergraduate researchers were surveyed and this talk will provide an insight of the findings.

Construction Of m-Repeated Burst Error Detecting and Correcting Non-Binary Linear Codes

B. K. Dass
University of Delhi, India
`dassbk@rediffmail.com`
Coauthors: Rashmi Verma University of Delhi, India E-mail: rashmiv710@gmail.com

The paper proposes a simple algorithm for constructing a parity-check matrix for any linear code over GF(q) capable of detecting and correcting a new kind of burst error called ‘m-repeated burst error of length b or less’ recently introduced by the authors. Codes based on the proposed algorithm have been illustrated.

Implementing Personalized Medicine: Estimation of Optimal Dynamic Treatment Regimes. Part I: Statistical Framework and Regression-Based Estimators

Marie Davidian
North Carolina State University
`davidian@ncsu.edu`

Personalized medicine is focused on making treatment decisions for an individual patient based on his/her genetic/genomic, clinical, and other characteristics. A popular perspective on achieving this goal is to focus on identifying subgroups of patients sharing certain characteristics who may benefit from a certain treatment and for whom new, targeted treatments may be developed. Another is based on operationalizing clinical practice. Clinicians make a series of treatment decisions over the course of a patient's disease or disorder based on accruing information on the patient with the goal of achieving the "best" possible outcome for the patient. A dynamic treatment regime is a list of sequential decision rules that formalizes this process. Each rule corresponds to a key decision point in the disease/disorder progression and takes as input the information on the patient to that point and outputs the treatment that s/he should receive from among the available options. A key step toward personalized medicine from the perspective of clinical practice is thus identifying the optimal dynamic treatment regime, that which would lead to the most favorable outcomes given the information available. We describe a statistical framework based on the notion of potential outcomes in which the definition of an optimal treatment regime may be formalized. Given data from a clinical trial or observational study, the statistical problem is to estimate an optimal treatment regime, a regime that leads to the most favorable outcome if followed by all patients in the population. We discuss methods for estimating an optimal treatment regime based on regression modeling. This talk is based on joint work with Baqun Zhang, Phillip Schulte, Eric B. Laber, and Anastasios A. Tsiatis.

Time to Discontinuation as a Primary Outcome in Clinical Trials

Sonia Davis
University of North Carolina
`sonia.davis@unc.edu`

Clinical trials of neuroscience indications often have markedly high rates of early patient discontinuation. Statistical methods to compare treatment groups for traditional outcomes in the face of missing data have become more developed in the past decade, yet high rates of patient discontinuation pose substantial complications for addressing bias and interpreting results. As an alternate strategy to address high dropout rates, we present time to all-cause treatment discontinuation (TTD) as a primary clinical trial outcome. We evaluate the benefits and drawbacks of TDD, and identify indications where is it well suited. Results from the CATIE study of antipsychotics for schizophrenia are used as an illustrative example. Time to discontinuation for lack of efficacy, intolerability, and patient decision are also explored. All-cause discontinuation is further contrasted with Treatment Failure.

A-Optimal and A-Efficient Designs for Discrete Choice Experiments

Angela Dean
The Ohio State University
`amd@stat.osu.edu`
Coauthors: Fangfang Sun

Discrete choice experiments are used widely in areas such as marketing, city planning, and medical care, etc., for the purpose of determining consumer preference to products and services. In a forced choice experiment, respondents are shown choice sets of options or profiles. The profiles represent a selection of products or services and are made up of a number of attributes (factors) at various levels and are similar to treatment combinations in a factorial experiment. Each respondent is asked to select their most preferred profile from each choice set. In- formation about the choices made can be used to infer the effects of the

individual attribute levels on consumer preference as well as interactions. In turn, this attribute information can be used in the future to design new products more tailored to the population requirements. In this talk, the multinomial logit model (MNL) is used to model the probability that a particular option is chosen from a given finite choice set. The systematic component of the utility of each option is modelled in terms of parameters representing the main effects and interactions of the attributes. An A-optimal design is one which minimizes the variance of the contrasts of interest of these parameters. Since the MNL model is a non-linear model, the optimal design despends upon the parameter values. A common approach is to construct designs for the particular case of equally attractive options. For this case, a lower bound is derived for the A-value of a discrete choice design. Some examples of designs acheiving the lower bound are given. For situations when the bound cannot be achieved, and when a complete search for the optimal design is not feasible, the model is linearized and a factor loading approach is taken to suggest options that should be included in a choice set. It is shown that this approach leads to highly A-efficient designs. This approach can be extended to construction of designs under the assumption of unequally attractive options.

Non-Negative Matrix Factorization Algorithms for Heteroscedastic Data

Karthik Devarajan
Temple University Health System
`karthik.devarajan@fccc.edu`
Coauthors: Vincent C.K. Cheung, McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA

Non-negative matrix factorization (NMF) is a powerful machine learning method for decomposing a high-dimensional non-negative matrix V into two non-negative matrices W and H such that V = WH + e where e is error. It has been success fully applied in the analysis and interpretation of large-scale data arising in computational biology, neuroscience and natural language processing, among other areas. A distinctive feature of this approach is its non-negativity constraints that allow only additive linear combinations of the data, thus enabling it to learn parts that have distinct physical representations in reality. Most of the methodological developments in NMF have ignored the heteroscedastic nature of the data observed in these diverse fields. We present an information-theoretic approach to NMF for modeling data exhibiting signal-dependent noise and discuss its connections to generalized linear models. We develop algorithm-specific measures for evaluating the performance of the proposed methods. Our methods are demonstrated using data from high-throughput genomics, text-mining and electromyography studies in the context of dimension reduction, and compared with existing methods.

Modeling Multimodal Healthcare Big Data

Quan Do
New Mexico State University
`quando@nmsu`
Coauthors: Son Tran - New Mexico State University and Venkat Gudivada - Marshall University

Healthcare data satisfies all the five attributes of Big Data. It is voluminous, heterogeneous, every day adds data in the order of gigabytes to terabytes, data undergoes transformation due to various businesses processes, and hospital and healthcare providers depend on this data for operational and strategic decision making. In the talk, we provide a critical evaluation of existing data representation schemes for modeling healthcare data. Our evaluation includes Relational, Column Relational, RDF Stores, Native XML, Object-oriented, Search Engine-based, Key-Value, Column-oriented, Document-oriented, Graph, NewSQL, CMS, Multivalue, and Navigational models. We will also discuss our preliminary work on data representation scheme that is specifically suited to address the needs of healthcare Big Data.

Sufficient Dimension Reduction with Multivariate Responses

Yuexiao Dong
Temple University
`ydong@temple.edu`
Coauthors: Zhou Yu, Liping Zhu

Existing methods for sufficient dimension reduction (SDR) with multivariate responses, such as projection resampling sliced inverse regression, have two limitations. First, the slicing based methods rely on selecting the slicing scheme, and different slicing could lead to different estimation results. Second, it can be computationally intensive as a large number of random projections of the multivariate response is needed in practice. To address such limitations, we propose a novel method for SDR with multivariate response which utilizes the distance between the distinct responses. In the challenging case of $p > n$, our proposal naturally leads to a novel feature screening method for multivariate response.

Bayesian Inference on Populations of Networks

David Dunson
Duke University
`dunson@duke.edu`
Coauthors: Daniele Durante and Joshua Vogelstein

In connectomic applications, a network of inter-connections in the brain is collected for each patient in the study. Interest focuses on characterizing variability among individuals, and how individuals in different groups (eg., according to mental health status) vary in their connection structure. Current methods focus on reducing network data to features prior to statistical analysis, while we propose a fully generative and flexible Bayesian nonparametric approach for modeling the population distribution of network-valued data. The approach is based on mapping the binary adjacency matrix within each brain to a latent space, conducting a flexible low-rank matrix factorization in this space, and then using Dirichlet process priors for the distribution of individual-specific components in this factorization. The formulation is provably flexible, and leads to a simple and computationally efficient Gibbs sampler. We discuss properties, illustrate the performance in simulations, and consider an application to brain network data in relation to covariates.

Prevention of Hospital Acquired Pressure Ulcers in an Operating Room Setting

Dawn Engels
Cone Health
`dawn.engels@conehealth.com`
Coauthors: Melody Austin, Laurie McNichol, Jennifer L. Fencl, Sat Gupta, and Haseeb Kazi

Prevention of hospital acquired pressure ulcers (HAPU) is an important quality measure since this is considered a never event. Literature suggests pressure ulcer prevalence occurs at a rate of 8.5% or higher among patients who undergo surgical procedures that last longer than three hours. The purpose of this study was to determine what factors might contribute towards the development of pressure ulcers in patients that have been to the Operating Room. (OR) This project entailed a retrospective chart review after surgical procedures were performed. The sample population included patients who obtained a pressure ulcer that was not present upon admission and which was noted during their post-operative hospital stay. The study revealed consistent risk factors which may contribute towards the development of pressure ulcers in patients that have been to the OR. These findings can drive implementation of preventative interventions to reduce the occurrence of HAPU associated with surgical procedures.

Market Liquidity as a Probability or as a Stochastic Process

Manuel L. Esquvel
New University of Lisbon, FCT, Department of Mathematics, Quinta da Torre
`mle@fct.unl.pt`
Coauthors: P. P. Mota

Motivated by the needs of risk analysis we introduce and study some alternative notions of liquidity for financial products.

Intra-Operative Skin Prep: Is There A Difference? A Retrospective Review

Jennifer L. Fencl
Cone Health
`jennifer.fencl@conehealth.com`

Surgical Site Infections (SSI) are the most common type hospital acquired infections in the surgical population with nearly 750,000 occurring annually. An individuals risk for dying is 11 times greater with the occurrence of an SSI. Additionally, SSIs increase length of stay by 7-31 days per event resulting in 1.6 billion dollars in cost for treatment. Hospitals are challenged to evaluate current practices to decreases SSI. The purpose of this project was to identify which intra-operative prepping solution utilized in our health network for adult open abdominal surgical procedures had the lowest incidence of SSI 30-days post-operatively. We performed a retrospective medical record of consecutive patients age 18 years and older undergoing elective open abdominal procedures at four acute care, community hospitals within an integrated tertiary health network in the Southeastern US from December 2008-December 2010. What we found out, procedures using PCMX intra-operative solution (n = 71) resulted in no SSI, while procedures that used other prepping solutions (CHG, povidone-iodine, and iodine-base with alcohol) (n=91) resulted in five SSIs. Chi-square test indicated a significant difference in the incidence of SSI among PCMX intra-operative prep solution when compared to the other solutions in open abdominal surgical procedures (2 = 4.025 (1, 162), p = 0.045). Research on this subject acknowledges an appropriate and effective surgical skin antiseptic plays a vital role in the prevention of devastating post-operative infections. By staying abreast of research and understanding the advantages and disadvantages of available surgical skin antiseptics, peri-Operative registered nurses are able to make informed recommendations

Examining Neurosurgical Surgical Site Infections

Jennifer L. Fencl
Cone Health
`jennifer.fencl@conehealth.com`
Coauthors: Felecia G. Wood, Debbie Green, Sat Gupta, Vangela Swofford, Melissa Morgan

The objective of this study was to analyze surgical site infection (SSI) risk factors for spinal surgery, an then develop an SSI risk assessment tool considering both pre- and intra-operative risk factors to identify patients at highest risk to develop an infection. We did this by utilizing a retrospective cohort study applying similar methodology as discussed in previous studies focusing on neurosurgery spinal SSI. The setting was an acute care hospital in the southeastern United States performing approximately 13,217 surgical cases per year with approximately 2579 of those being neurosurgical spinal cases. Study subjects included patients who developed a SSI after a neurosurgery spinal procedure as defined by the Center for Disease Control and Prevention (case patient) and randomly selected non-infected patients (match control patient) who underwent a neurosurgery spinal procedure during the same timeframe as the case patient (June 2012-June 2013). Patients were matched based on type of surgery and Anesthesiologists (ASA) score. We used a retrospective chart review using an investigator-developed drill down tool reflecting specific risk factors for SSI identified in the literature for the neurosurgery spinal population. We found that the strongest predictor of SSI was the type of intra-operative prepping solution utilized (p= .009) and how skin closure occurred (p=

.006). In addition, the placement of drains (p= .001), misappropriate antibiotic administration (p=.026), and if glue utilized to repair dural tears (p = .041) also demonstrated an increased risk of developing a SSI. This study suggests that healthcare organizations must proactively take measures to identify patients at highest risk for poor outcomes and thoughtfully implement standardization to improve care, which needs to include assessment of intraoperative risk factors.

Robust Feature Selection with Fussy Group Information

Xiaoli Gao
University of North Carolina at Greensboro
`x_gao2@uncg.edu`

In high-dimensional data analysis, grouping information may behave as a double-edged sword and have the potentiality of damaging the feature selection efficiency if fussy grouping information is incorporated. In this paper, we propose a robust alternative to existing group LASSO approach when the assumed grouping structure is flawed. Some small sample simulation studies and real data analysis will be conducted to demonstrate the advantage of the proposed approach.

Finding the Symmetry Group of a Linear Program for Enumerating Orthogonal Arrays

Dursun Bulutoglu and Andrew Geyer
AAir Force Institute of Technology
`dursun.bulutoglu@gmail.com`

For a given LP (linear program) the set of all permutation of its variables that send feasible points to feasible points was defined by Margot (2010) to be the symmetry group of the LP denoted by G^{LP}. Margot (2010) described a method for computing the G^{LP} of an LP with only inequality constraints. His method finds a subgroup of the G^{LP} when the LP has equality constraints. We will discuss a method for finding the G^{LP} of an LP that also has equality constraints. We will also discuss two different (integer linear programming) ILP formulations of the orthogonal array problem and the G^{LP}s of their corresponding LP relaxations.

A Computationally Efficient Flexible Observed Factor Model with Separate Dynamics for the Factor Volatilities and Correlation Matrix

Sujit Ghosh
North Carolina State University & SAMSI
`sujit.ghosh@ncsu.edu`
Coauthors: Yu-Cheng Ku and Peter Bloomfield

Multivariate stochastic volatility (MSV) models have become important tools in financial econometrics, largely due to the successful utilization of Markov chain Monte Carlo (MCMC) methods. Recent developments in the MSV literature focus on dimension reduction via factor analysis, given that the complexity of computation and the difficulty in model interpretation drastically increase as the dimension of data increases. In order to obtain a flexible yet computationally efficient procedure, we propose a MSV structure with separate dynamics for the volatilities and the correlation matrix. The correlation matrix of the factors is allowed to be time varying, and its evolution is described by an inverse Wishart process. The proposed MCMC method is shown to have some major advantages compared to existing methods for similar models in the literature. For numerical illustrations, we compare the proposed model with other similar multivariate volatility models using Fama-French factors and portfolio weighted return data. The result shows that our model has better predictive performance and computationally more stable.

Modeling Dependence in Pricing Risk

Barry K. Goodwin
North Carolina State University
`bkgoodwi@ncsu.edu`

As empirical scientists, we have come to define dependence largely in linear terms, typically represented by correlation or linear regression parameters. Galton introduced the notion of linear correlation in 1885 and this narrow idea of dependence has dominated our thinking since. Dependence is a much broader concept than this, however, and it is key to understanding multivariate ordering and modeling. The proper representation of dependence has significant implications in economic models of risk. Insurance contracts typically consider multiple dependent sources of risk. Reinsurance portfolios and the pricing of derivative instruments also depend heavily on the relationships among multiple dependent sources of risk. This paper reviews the implications of dependence for the pricing of risks. Tail-dependence, state-dependence, and nonlinear functional relationships are considered within the general context of copulas. Examples taken from the pricing of revenue insurance contracts are discussed. The implications of small samples and sampling variability in the precision of model estimates are also developed.

Robust Mixtures of Factor Analyzers via the Trimming Approach

Francesca Greselin
University of Milano-Bicocca, Italy
`francesca.greselin@unimib.it`
Coauthors: Luis Angel Garcia-Escudero, Alfonso Gordaliza, Salvatore Ingrassia and Agustin Mayo-Iscar

Mixtures of Gaussian factors are powerful tools for modeling an unobserved heterogeneous population, offering - at the same time - dimension reduction and model-based clustering. Unfortunately, the high prevalence of spurious solutions and the disturbing effects of outlying observations, along maximum likelihood estimation, open serious issues. In this paper we complement model estimation with restrictions for the component covariances and trimming, to provide robustness to violations of normality assumptions of the underlying latent factors. A detailed AECM algorithm, which enforces constraints on eigenvalues and tentatively discards outliers at each step, is also presented. In the new setting, under mild conditions on the underlying distribution, the existence of the solution for the population and sample problems can be obtained, as well as the consistency of the sample estimator to the population parameter. Properties of the estimators are evaluated through MonteCarlo experiments and compared to the estimators from the classical model. Results from the analysis a a real dataset are presented.

Playing in Everyones Backyard: Statistical Consulting in the Sciences

Emily H Griffith
North Carolina State University
`emily_griffith@ncsu.edu`

Playing in everyones backyard (apologies to John Tukey) is an exciting and challenging part of being a statistical consultant. If you work with scientists in a broad array of fields, however, it can quickly become overwhelming. Is it better to limit your expertise to the statistical part of the study, or should you immerse yourself in the science? How do you keep from being overwhelmed by working with many different specialties? When do you become a collaborator instead of a consultant? This talk will cover how to evaluate the amount of involvement in the science that is appropriate using specific examples from consulting projects.

HCUP Data Analysis with MongoDB

Akhil M Gudivada
Marshall University
`gudivada2@marshall.edu`
Coauthors: Dhana Rao

The HCUP is a publicly available dataset and includes the largest collection of longitudinal hospital care data in the U.S. from 1988 to present. In this talk, we describe our approach to analyzing HCUP data for various trends using MongoDB, a NoSQL database management system. Our analysis encompasses over 24 million records corresponding to patient discharge data from over 1,000 hospitals. The unit of analysis is the discharge record, not individual patients. Our analysis focuses on gaining insights into access to health care, cost and quality of health services, medical practice patterns, and outcomes of treatments at the national, state, and local levels.

Sampling, Sketching and Streaming: Designing Combinatorial Algorithms for BigData

Sudipto Guha
University of Pennsylvania
`sudipto@cis.upenn.edu`

The explosion of data over the last two decades has mandated that space and access constraints are as important to algorithms as accuracy and running time. This has led to the investigation of a variety of models of computation such as streaming, compressed sensing, and map-reduce, to name a few. Significant recent progress has been made and for several statistical estimation problems, the tradeoffs between the four parameters are well understood. Yet similar progress has been elusive in large combinatorial or graph problems, even though graphs constitute one of the most classic representations of data.

In this talk we will review how the different models relate to each other and a few of the recent advances in parameter estimation problems. We will particularly review the notion of sketching computing inner products of the input with suitably designed pseudo-random vectors. The majority of the talk will focus on developing graph algorithms in these models that go beyond parameter estimation and output extended structures, for example, a spanning tree or an approximately optimal matching.

Discovering Dynamic Topics in Time Stamped Documents

Shaobo Han
Duke University
`shaobohan@gmail.com`
Coauthors: Lin Du, Esther Salazar, and Lawrence Carin

We consider the problem of discovering latent structures in a text streams. We propose a dynamic sparse factor model based on the extended rank likelihood, capable of (1) discovering the time-evolving importance of topics, (2) learning contemporary topic and word correlations. The high-dimensional, time-evolving word counts, after an arbitrary monotone transformation, are well accommodated through an underlying dynamic sparse factor model. The framework naturally admits heavy-tailed innovations, capable of inferring abrupt temporal jumps in the importance of topics. Locally conjugate posterior inference is performed through Gibbs sampling, based on the forward-filtering backward-sampling algorithm. The model is demonstrated by analyzing the US Presidential State of the Union addresses from 1790 to 2014.

Dynamic Network-Topic Models

Christoph Hellmayr
Duke University
`ch.hellmayr@gmail.com`
Coauthors: David Banks

In many situations in which Topic Modelling is appropriate there exists additional data regarding the network structure between the document generating agents. We attempt to include this information in a dynamic Network-Topic Model. We model both the time evolution of and the effects of the network structure on topics. We present an algorithm for generating data from this model, and an outline of a MCMC-sampler to fit it.

Nonparametric Bayes Models for Modeling Longitudinal Change in Association among Categorical Variables

Amy Herring
The University of North Carolina at Chapel Hill
`amy_herring@unc.edu`
Coauthors: Tsuyoshi Kunihama, Duke University David Dunson, Duke University Carolyn Halpern, The University of North Carolina at Chapel Hill

Modeling and computation for multivariate longitudinal data has proven challenging, particularly when data are not all continuous but contain discrete measurements. Approaches based on generalized linear mixed modeling, and related exponential family hierarchical models, have been criticized due to a lack of robustness. In particular, problems arise due to the dual role of the random effects structure in controlling the dependence and shape of the marginal distributions. Motivated by an interesting application to sexual preference data, we propose a novel approach based on a Dirichlet process mixture of Gaussian latent factor models. The proposed model uses a rounded kernel method to allow data to be mixed scale, with a longitudinal factor structure incorporating dependence within-subjects in their repeated measurements. Survey weights are incorporated into the model to facilitate generalizability. Parameter interpretation is considered, and an efficient Markov chain Monte Carlo algorithm is proposed. The methods are assessed through simulation studies and applied to the National Longitudinal Study of Adolescent Health.

The Liberti Trial for Discovering a Dynamic Treatment Regimen in Burn Scar Repair.

Jonathan Hibbard
UNC Chapel Hill
`jhibbard@email.unc.edu`
Coauthors: Michael R. Kosorok

In this talk we describe the design and analysis plan for the Liberti (Laser Induced, Biologically Engineered Remodeling of Thermally Injured) Skin Trial. This is a SMART (Sequential Multiple Assignment Randomized Trial) design to discover the best sequence of treatments over three time intervals to improve outcomes for patients with severe burn scaring as a function of baseline and historical tailoring variables. In addition, a simple randomized comparison of the three treatments being compared (standard of care plus two different laser treatments) using a surrogate outcome is embedded within the SMART design.

Extension of Rotation Technique via Penalization in Factor Anlaysis Model

Kei Hirose
Osaka University
`hirose@sigmath.es.osaka-u.ac.jp`
Coauthors: Michio Yamamoto

We consider the problem of sparse estimation in a factor analysis model. A traditional estimation procedure in use is the following two-step approach: the model is estimated by maximum likelihood method and then a rotation technique is utilized to find sparse factor loadings. However, the maximum likelihood estimates cannot be obtained when the number of variables is much larger than the number of observations. Furthermore, even if the maximum likelihood estimates are available, the rotation technique does not often produce a sufficiently sparse solution. In order to handle these problems, this paper introduces a penalized likelihood procedure that imposes a nonconvex penalty on the factor loadings. We show that the penalized likelihood procedure is an extension of the traditional two-step approach. In addition, the proposed methodology can produce sparser solutions than the rotation technique. The model is estimated by EM algorithm along with coordinate descent.

Simulation for Designing Medical Device Trials

Martin Ho
FDA/CDRH
`martin.ho@fda.hhs.gov`
Coauthors: Xuefeng Li (FDA/CDRH) Telba Irony (FDA/CDRH)

Division of Biostatistics at the Center for Devices and Radiological Health (CDRH) of Food and Drug Administration has reviewed and accepted various types of adaptive clinical trial designs for confirmatory pivotal medical device trials for a decade. Although these designs operating characteristics often do not have closed form solutions, CDRH requires the sponsors to provide sufficient simulations under different scenarios for study design evaluations. In this presentation, I will discuss the experience of CDRH using simulations to evaluate clinical study designs and analyze study results. I will first explain why simulation importance is important for clinical study designs. Next, I will describe what and how we evaluate the simulations submitted by the sponsors. After going through a hypothetical example of Bayesian adaptive design trial, I will conclude the presentation with some remarks from a regulatory perspective.

Nested Dirichlet Process Model for Household Data Synthesis

Jingchen Hu
Duke University
`jh309@stat.duke.edu`
Coauthors: Jerome P. Reiter

Surveys on household data may comprise only nominal responses (sex, race, marital status etc.) for each household member (such as the Decennial Census Data on Families). In order to protect respondents privacy, suitable model for generating synthetic dataset is needed. In this study we are developing a fully Bayesian, joint modeling approach for categorical data based on the nested Dirichlet process developed by Rodriguez et al. (2008), for household synthetic data simulation. The model aims to preserve dependence structures among variables at both the individual level and the household level. Extensions of this model can deal with relational structural zeros in household dataset.

Optimal Designs for Multi-Covariate Generalized Liner Models

Linwei Hu
University of Georgia
`pangaea@uga.edu`
Coauthors: John Stufken

Optimal design results for multi-covariate generalized linear models are scarce. In this paper, we will derive some complete class results for generalized linear models with an arbitrary number of covariates and possible interactions between the covariates. The design space can also be any hypercube. Based on the complete class results, we identify a small class of designs in which the optimal designs can be found, thus greatly simplifies the computational cost for searching the optimal design. The saving in computation increases exponentially as the number of covariates increases, and is shown in a few numerical studies.

Consistent Selection of the True Covariance Components in High-Dimensional Growth Curve Model with Random Coefficients

Shinpei Imori
Osaka University
`imori@sigmath.es.osaka-u.ac.jp`
Coauthors: Dietrich von Rosen

We try to select the true covariance components in the growth curve model with random coefficients when the dimension of response variables is large. The generalized information criterion type model selector is proposed. Under mild conditions, the proposed criterion has a consistency property, i.e., the selection probability of the true covariance components by the proposed criterion goes to 1 with the sample size tending to infinity.

Longitudinal Data Analysis on the Assessment of Equivalence in Biosimilar Clinical Trials

Bo Jin
Pfizer Inc
`bo.jin@pfizer.com`
Coauthors: Kerry Barker, Ron Menton, Jim Bolognese, Mikhail Salganik, Lingyun Liu

The assessment of equivalence between biosimilar and reference products in clinical pharmacodynamics (PD) and/or efficacy endpoints is essential to Biosimilar clinical development after the equivalence in clinical pharmacokinetics is demonstrated. Unlike innovator non-inferiority trials where non-inferiority is usually targeted on a primary endpoint (a single time point), biosimilar equivalence to reference products often involves the evaluation on multiple time points in PD and/or efficacy endpoints. It is therefore of particular interest how to evaluate equivalence over multiple time points in an efficient and effective way. In this paper, we study the application of longitudinal data analysis models on the evaluation of equivalence on multiple time points. Specifically, several longitudinal data analysis (LDA) models are compared in the analysis of DAS28, a continuous composite endpoint which is commonly used to measure severity of rheumatoid arthritis. Longitudinal Emax model, linear hockey-stick (HS) model, longitudinal linear mixed model and robust linear mixed model will be compared under the various simulated true time-response scenarios. The simulations indicated a tradeoff for the Emax model, which had the smallest standard errors and the greatest power, but had the greatest potential at the same time for yielding biased estimates and inflating the alpha level. The traditional linear mixed models and robust linear models can yield unbiased estimates and keep conservative alpha levels, but are less powerful than Emax and HS models. The linear hockey-stick model appeared to be most robust to true response curves and seemed to control alpha levels while still having satisfactory power.

Cluster-Based Methodologies in Community-Based Interventions: An Overview

Robert E. Johnson
Vanderbilt University
`robert.e.johnson@vanderbilt.edu`

The randomized control trial is the prominent methodological tool used to establish evidence in medical and other research. Community-based interventions may demand implementation at the community level rather than at the subject level. This requires all or a sample of subjects within the community to be randomized to intervention or control conditions. Such clustering must be accounted for in the determination of sample size and statistical analysis. A few short years ago it was too common for analysts to ignore the cluster-base of the sample, causing over-stated significance of interventions. The proper use of cluster-randomized trials has thankfully become prevalent, mainly through education. With this, an array of methodological issues has come to the forefront. Sample size strategies and related software, mixed models and related hierarchical models for analysis, adapted sampling and randomization strategies, and approaches to accommodate cohorts, repeated values, nesting, etc., are among the various cluster-based methodologies that have and continue to be developed. A brief overview of the history of cluster-based trials and current developments will be provided.

Physiological Age, Health Costs and their Relationship

Bruce Jones
University of Western Ontario
`jones@stats.uwo.ca`
Coauthors: Maria Govorun, Xiaoming Liu, David Stanford

We use a phase-type approach to demonstrate the impact of observable annual health cost information on the evaluation of an individual's physiological age.

In 2007, Lin and Liu proposed a model for human aging which is a particular Markov chain with states being unobservable physiological ages of an individual. In this phase-type aging model, an individual of a given age has a defined distribution of his/her physiological age. Many studies indicate that it is important for health cost projections and for underwriting purposes to take into account the physiological condition of an individual rather than just his/her chronological age. In this work, we extend the Lin and Liu model in order to use information on annual health costs to infer the distribution of the physiological age for an individual of a given chronological age. We also show the impact of the cost information on the net present value health care costs.

Challenges of Data Integration and Interoperability in Big Data.

Anirudh Kadadi
North Carolina A&T State University
`akadadi@aggies.ncat.edu`
Coauthors: Dr. Rajeev Agrawal

The enormous volumes of data created and maintained by industries, research institutions are on the verge of outgrowing its infrastructure. The advancements in the organizations work flow include Data Storage, Data Management, Data Maintenance, Data Integration and Data Interoperability. Among these levels, Data Integration and Data Interoperability can be the two major focus areas for the organizations which tend to implement advancements in their workflow. Overall, Data Integration and Data Interoperability influence the data performance. The Data Integration and Data Interoperability is complex for the Big Data thereby requires a comprehensive path to negotiate the challenges. Given a scenario where two big organizations merge and tend to operate centrally, data integration and interoperability could be the major area of focus as the two organizations might have had respective data management techniques before the merger. This presentation focuses on the challenges of Data Integration and Data Interoperability, emphasizing on the

different techniques to integrate and operate the Big Data, also focusing on ideas for Data Integration and Interoperability in Big Data.

A Regression Estimator for Finite Population Mean of a Sensitive Variable

Geeta Kalucha
University of Delhi, India
`geetakalucha@gmail.com`
Coauthors: Sat Gupta & Javid Shabbir

Kalucha et al. (2014) introduced an additive ratio estimator for finite population mean of a sensitive variable in simple random sampling without replacement and showed that this estimator performs better than the ordinary mean estimator based on an optional randomized response technique (RRT). In this article, we introduce a regression estimator that performs better than the ratio estimator even for modest correlation between the primary and the auxiliary variable. The primary study variable is sensitive in nature and the corresponding auxiliary variable that is positively correlated with the primary variable is non-sensitive. In the present work, we improve on the results of Gupta et al. (2013) further by using an optional RRT model. Expressions for the Bias and MSE (Mean Square error) for the proposed estimator are derived based on first order of approximation. A comparison of the proposed estimator with the corresponding ratio estimator and the ordinary RRT mean estimator is carried out both theoretically and numerically.

Effects of Method Parameters and Ground Truth in the OMOP Results Database

Alan Karr
RTI International
`karr@niss.org`

The OMOP results database is a unique dataset containing the results of more than 6 million statistical analyses of patient-level medical data. The analyses are meant to identify drugs associated with particular adverse outcomes. The data are drawn from five large provider databases, and cover 181 drugs and 19 outcomes. Seven classes of statistical methods were applied to each (drug,outcome) pair, with 1246 variants arising from parameter settings within each method class. The dataset also contains a binary ground truth for each (drug,outcome) pair.

This presentation will describe two sets of condition-specific analyses of the OMOP results database that illuminate the roles of modeling choices and ground truth in determining log relative risk values. The first are partition models that show concretely the effects and interactions of model parameters and ground truth, as well as reveal clearly the bias resulting from particular choices of parameters. The second set of analyses are "efficient frontier" models that identify non-dominated methods on the basis of misclassification rates.

An simplified encapsulation of the results is that often how the data are analyzed affects the results more than does ground truth.

Two-Level Screening Designs Derived from Binary Non-Linear Codes

Tena I. Katsaounis
Ohio State University at Mansfield
`katsaounis.1@osu.edu`
Coauthors: Manohar Aggarwal, University of Memphis

Two-level non-regular fractional factorial designs can provide economical designs in screening experiments because of their flexible run sizes and good projection properties. Two criteria are given for evaluating the overall aliasing and uniformity properties of projections of a non-regular fractional factorial design onto the active factors. They are used to rank two-level non-regular fractional designs which correspond to binary non-linear codes. Examples of such non-regular fractional factorial designs are compared to the Plackett-Burman designs with 12, 24, 32 and 40 runs and various numbers of factors.

A Spatio-Temporal Model for Predicting Bark Beetle Damage

Kimberly Kaufeld
SAMSI/North Carolina State University
`kimberly.kaufeld@gmail.com`
Coauthors: Matthew Heaton, Stephen Sain

Bark beetle damage is one of the primary causes for the dramatic change of the forest composition in the Rocky Mountain region over the last few years. In order to mitigate the bark beetle epidemic, statistical modeling is needed to understand and predict the spread and occurrence of damage. Using data in the Rocky Mountains between the years 2001-2011, a zero-augmented spatio-temporal beta regression model is used to predict both the occurrence of beetle damage and, given damage occurred, the percent of the region infected. A stick-breaking representation accounts for the proportion of cumulative damage occurred over time. Temporal evolution of the bark beetle damage is captured using a dynamic linear model where both the probability and extent of damage depend on the amount of damage incurred in neighboring regions in the previous time period. A sparse conditional autoregressive model captures any spatial information not modeled by spatially varying covariates. We find that the occurrence and extent of beetle damage are negatively associated with elevation and damage in previous time periods. Additionally, we demonstrate the utility of predicting bark beetle damage by highlighting areas that display a higher probability of outbreak in future time periods.

The Challenges of Statistical Consultancy in Developing Countries Tanzanian Case Study

Benedicto Kazuzuru
Sokoine University of Agriculture-Tanzania & Virginia Tech
`kazuzuru@vt.edu`

Laboratory for Interdisciplinary Statistical Analysis 2020 is a program at Virginia Tech University to train statisticians from developing countries to become good statistical collaborators and thereafter be able to establish statistical collaboration laboratories in their countries. LISA 2020 program has a plan to establish 20 statistical collaboration laboratories in 20 developing countries by the year 2020. LISA has so far earmarked some developing countries including Nigeria (OAU University) and Tanzania (Sokoine University of Agriculture) as countries from Africa where such a laboratory can be established. Unfortunate the developing countries social economic environments may not be as good as the environments in developed countries and particular in USA where we have such a statistical laboratory (LISA). This scenario may somehow slow down the functioning and progress of Statistical laboratories in developing countries. For example, the issue of African universities funding LISA may be among the critical challenges in the functioning of LISA in developing countries. This talk intend to highlight some key challenges which statistical laboratories in developing countries may face and propose where possible alternatives solutions to such challenges

Repeated Out of Sample Fusion in Interval Estimation of Small Tail Probabilities in Food Safety

Benjamin Kedem
University of Maryland
`bnk@math.umd.edu`

In food safety and bio-surveillance in many cases it is desired to estimate the probability that a contaminant such as some insecticide or pesticide exceeds unsafe very high thresholds. The probability or chance in question is then very small. To estimate such a probability we need information about large values. However, in many cases the data do not contain information about exceedingly large contamination levels. A solution is suggested whereby more information about small tail probabilities is obtained by REPEATED FUSION of the real data with computer generated data. The method provides short but reliable interval estimates from moderately large samples. An illustration is provided using exposure data from the National Health and Nutrition Examination Survey (NHANES).

Circular and Diagonal Circular Systematic Sampling in the Presence of Linear Trend

Zaheen Khan
Federal Urdu University of Arts, Science and Technology, Pakistan.
`zkurdu@gmail.com`

In this paper, a new approach is used to study the characteristics of circular version of systematic sampling schemes. This approach enables us to derive the variance expressions of Circular Systematic Sampling (CSS) and Diagonal Circular Systematic Sampling (DCSS) for the populations having linear trend, which were not available up till now. Using this new approach, corrected sample mean and average variance of the corrected sample mean for CSS and DCSS are also obtain. Further, we extend the whole work for the case where some label of units of certain samples exceeds , which were also not considered earlier. A detailed efficiency comparison is also carried out.

Key words: CSS, DCSS, linear trend, corrected mean, average variance.

A Direct Utility Model for Economies of Scope

Jaehwan Kim
Korea University Business School
`jbayes@korea.ac.kr`
Coauthors: Dong Soo Kim (Ohio State University) Sanghak Lee(Universiyt of Iowa) Greg Allenby (Ohio State University)

Economies of scope exist when consumers find it beneficial to consume multiple offerings because it reduces the cost of consumption. Examples include shopping at malls and other venues where products are in close proximity, and ordering season tickets that do not require seat requests and billing information for every home game. Economies of scope lead to the consumption of multiple goods because of cost reductions, in contrast to models of satiation that explain multiple purchases in terms of diminishing marginal returns in utility. We develop a model of scope economies for a latent, unobserved resource that constrains consumer purchases. The model is applied to a conjoint dataset of theme park demand, where we find that economies of scope lead to a better model fit and smaller price elasticities relative to existing models of demand.

Generalized Functional Concurrent Model

Janet S. Kim
North Carolina State University
`jskim3@ncsu.edu`
Coauthors: Arnab Maity, Ana-Maria Staicu

We consider the generalized functional model, where both the response and the covariate are functional data and are observed on the same domain. In contrast to typical functional linear concurrent models, we allow the relationship between the response and covariate to be nonlinear, depending on both the value of the covariate at a specific time point as well as the time point itself. In this framework we develop methodology for estimation of the unknown relationship, allowing for correlated error structure as well as sparse and/or irregular design. In this study we mainly focus on prediction of an unobserved functional response. We construct point-wise prediction bands for the predicted trajectories under the flexible settings of observations. We test the existence of unknown dependence between the response and the covariate through bootstrapping. We investigate this approach in finite sample size through simulations and a real data application.

Database Storage Engines for Big Data Management

Rajesh Babu Kodali
Marshall University
`kodali@marshall.edu`
Coauthors: Dr. Venkat N Gudivada

Special needs of Web 2.0, Big Data, and mobile applications have ushered in several new classes of systems for data management. This renaissance in database management systems (DBMS) is ushering a new trend where a database system can be built using modular components to closely match the application data needs. Database storage engine is a central component of this approach.

In this talk, we present our investigation of the following storage engines: InnoDB, MyISAM, TokuDB, MERGE MEMORY (HEAP), BDB (BerkeleyDB), EXAMPLE, ARCHIVE, CSV, BLACKHOLE, ISAM, MariaDB, Tokutek, ARCHIVE, ndbcluster, FEDERATED, MRG_MYISAM, and PostgreSQL. We discuss their functional and performance characteristics and identify Big Data application scenarios that are ideal for each engine.

Clustering Financial Time Series: A Polyspectral SLEX Approach

Priya Kohli
Connecticut College
`pkohli@conncoll.edu`
Coauthors: Nalini Ravishanker and Jane L Harvill

Clustering of time series has applications in diverse fields like business, finance and geophysical processes. Time series data in many scientific or engineering disciplines exhibit nonstationarity and/or nonlinearity. The most widely used clustering methods are based on the second order properties which are generally insufficient for the nonlinear processes. There are several open problems in the estimation of time-varying, higher-order spectra for nonlinear time series, and the development of optimal clustering algorithms based on these spectral estimates. We address these issues using an approach based on the smooth localized complex exponential library of orthogonal transforms for polyspectra (higher order spectra) that is statistically and computationally efficient. The performance of the proposed approach is illustrated through extensive simulations and real-data analysis.

On Lotka-Volterra Population Games

Vlastimil Křivan
Biology centre of AS CR and Faculty of Science USB, Czech Republic
`vlastimil.krivan@gmail.com`

Lotka-Volterra predator-prey and competition models played an instrumental role in developing fundamental principles of contemporary population ecology. These dynamics assume that the interaction strength between different species is fixed (e.g., predator's food preferences are fixed and independent of prey densities). In other words, they do not consider animal behavior (e.g., food preferences, habitat preferences, predator avoidance etc.) that can change as a response to changing environment. Such adaptive animal behaviors have been documented empirically, and are often postulated in game theoretical evolutionary models. To integrate population dynamics with evolutionary models requires to specify time-scales on which these processes do operate. In my talk I will adopt the classical dogma of ecology where behavior is much faster when compared to population dynamics which are still much faster than evolutionary changes in traits. With this perspective I will discuss effect of adaptive animal behaviors on classical Lotka-Volterra predator-prey and competition models. I will show that adaptive animal behaviors when applied to the Lotka-Volterra models can strongly change their predictions and provide new hypothesis for species coexistence.

Quantitative Analysis of Effectiveness of Teaching an Online Versus In-Class Course

Ashok Kumar
Kentcuky State University
`ashok.kumar@kysu.edu`

More and more public and private educational institutions and training institutes are moving towards online education. There is a tremendous growth of online courses for either making profit by the private organization or in order to meet large funding cuts to the public institutions over the last 10 years. During this period, there has been an explosion of these online courses all over the world, especially in USA. Our research is being conducted for quantitatively analyze the data obtained from two sections of a course to measure and compare the effectiveness of an online course vs and traditional on campus course.

Generalized Signed-Rank Estimation for Nonlinear Models with Multidimensional Indices

Eddy Kwessi
Trinity University
`ekwessi@trinity.edu`
Coauthors: B. Nguelifack, A. Abebe

We consider a nonlinear regression model when the index variable is multidimensional. Such models are useful in signal processing, texture modeling, and spatio-temporal data analysis. A generalized form of the signed rank estimator of the nonlinear regression coefficients is proposed. This general form of the signed-rank estimator includes Lp estimators and hybrid variants. Sufficient conditions for strong consistency and asymptotically normality of the estimator are given. It is shown that the rate of convergence to normality can be different from root-n. The sufficient conditions are weak in the sense that they are satisfied by harmonic type functions for which results in the current literature are not applicable. A simulation study shows that the certain generalized signed-rank estimators (eg. signed-rank) perform better in recovering signals than others (eg. least squares) when the error distribution is contaminated or is heavy-tailed

Functional Feature Construction for Personalized Medicine

Eric Laber
North Carolina State University
`laber@stat.ncsu.edu`
Coauthors: Ana-Maria Staicu, Leslie Sall

Evidence-based personalized medicine formalizes treatment selection as a decision rule that maps up-to-date patient information into the space of possible treatments. Available patient information may include static features such race, gender, family history, genetic and genomic information, as well as longitudinal information including the emergence of comorbidities, waxing and waning of symptoms, side-effect burden, and adherence. Dynamic information measured at multiple time points before treatment assignment should be included as input to the decision rule. However, subject longitudinal measurements are typically sparse, irregularly spaced, noisy, and vary in number across subjects. Existing estimators for decision rules require equal information be measured on each subject and thus standard practice is to summarize longitudinal subject information into a scalar ad hoc summary during data pre-processing. This reduction of the longitudinal information to a scalar feature precedes estimation of a decision rule and is therefore not informed by subject outcomes, treatments, or covariates. We propose a data-driven method for constructing maximally prescriptive yet interpretable features that can be used with standard methods for estimating optimal decision rules. In our proposed framework, we treat the subject longitudinal information as a realization of a stochastic process observed with error at discrete time points. The estimated feature is expressible as the integral of the subject longitudinal process against a smooth coefficient function; the estimated coefficient function therefore describes the optimal weighting of subject-specific longitudinal information which is potentially informative for clinical practice.

Two Years in Silicon Valley – Some Experiences, Stories, and Thoughts about Big Data

Lexin Li
University of California, Berkeley
`lexinli@berkeley.edu`

There has been an enormous and continuously growing interest in Big Data in recent years – Just see how often we hear this term from the media and from everyone around us nowadays. In 2011-2013, I was on leave from North Carolina State University, and had been visiting Yahoo Research Labs and Stanford University. Then in fall 2013, I offered a topic course on Big Data at NC State, Statistics Department. In this talk, I would like to share some of my own experiences and personal thoughts on issues related to Big Data, and will talk about the Big Data course that I taught. This talk has been based upon many conversations and interactions with people from leading Internet Technology companies in Silicon Valley, as well as my students who took the course.

Modeling Nonstationary Covariance Function with Convolution on Sphere

Yang Li
University of Minnesota Duluth
`yangli@d.umn.edu`
Coauthors: Zhengyuan Zhu

The wide use of satellite-based instruments provides measurements in climatology on a global scale, which often have nonstationary covariance structure. In this paper we address the issue of modeling axially symmetric spatial random fields on sphere with a kernel convolution approach. The observed random field is generated by convolving a latent uncorrelated random field with a class of Matern type kernel functions. By allowing the parameters in the kernel functions to vary with locations, we are able to generate a flexible class of covariance functions and capture the nonstationary properties. Since the corresponding covariance functions generally do not have a closed form, numerical evaluations are necessary and a pre-computation

table is used to speed up the computation. For regular grid data on sphere, the circulant block property of the covariance matrix enables us to use Fast Fourier Transform (FFT) to get its determinant and inverse matrix efficiently. We apply this approach to the Total Ozone Mapping Spectrometer (TOMS) ozone data.

A Nonparametric Approach to Combining Information from Multiple Sources

Dungang Liu
University of Cincinnati
`liudg@ucmail.uc.edu`
Coauthors: Regina Liu and Minge Xie

For the purpose of combining inferences from several nonparametric studies for a common hypothesis, we develop a new methodology using the concepts of data depth and confidence distribution. A confidence distribution (CD) is a sample-dependent distribution function that can be used to estimate parameters of interest. It is a purely frequentist concept yet can be viewed as a distribution estimator of the parameter of interest. Examples of CDs include Efrons bootstrap distribution and Frasers significance function (also referred to as p-value function). In recent years, the concept of CD has attracted renewed interest and has shown high potential to be an effective tool in statistical inference. In this project, we use the concept of CD, coupled with data depth, to develop a new approach for combining the test results from several independent studies for a common multivariate nonparametric hypothesis. Specifically, in each study, we apply data depth and bootstraps to obtain a p-value function for the common hypothesis. The p-value functions are then combined under the framework of combining confidence distributions. This approach has several advantages. First, it allows us to resample directly from the empirical distribution, rather than from the estimated population distribution satisfying the null constraints. Second, it enables us to obtain test results directly without having to construct an explicit test statistic and then establish or approximate its sampling distribution. The proposed method provides a valid inference approach for a broad class of testing problems involving multiple studies where the parameters of interest can be either finite or infinite dimensional. The method will be illustrated using simulations and flight data from the Federal Aviation Administration (FAA).

Modeling Novel Strategies for Controlling Mosquito-Borne Diseases

Alun Lloyd
North Carolina State University
`alun_lloyd@ncsu.edu`

Dengue is the most significant mosquito-borne viral infection of humans, causing hundreds of millions of infections annually. The main line of attack against dengue has been traditional mosquito control measures, such as insecticides. The coming years will see the broadening of our anti-dengue arsenal to include genetically-modified mosquitoes, biocontrol methods—such as Wolbachia—and vaccines. Mathematical models are being, and have been, used to assess the feasibility and effectiveness of a number of these approaches at various stages of their development. In this talk, I shall discuss the biological bases of some of these novel strategies and the accompanying modeling work, illustrating the use of a number of different models as the projects move along the path from theoretical proposals, through lab-based studies to field deployment.

Clinical Utility Estimation for Assay Cut-offs in Early Phase Oncology Enrichment Trials

Jared Lunceford
Merck & Co., Inc.
`jared_lunceford@merck.com`
Coauthors: Jeff Cheng, Peggy Wong, Devan V. Mehrotra

Predictive enrichment strategies use biomarkers to selectively enroll oncology patients into clinical trials to more efficiently demonstrate therapeutic benefit. Because the enriched population differs from the patient population eligible for screening with the biomarker assay, there is potential for bias when estimating clinical utility for the screening eligible population if the selection process is ignored. We write estimators of clinical utility as integrals averaging regression model predictions over the conditional distribution of the biomarker scores defined by the assay cut-off and discuss the conditions under which consistent estimation can be achieved while accounting for some nuances that may arise as the biomarker assay progresses toward a companion diagnostic. We outline and implement a Bayesian approach to estimating these clinical utility measures and use simulations to illustrate performance and the potential biases when estimation naively ignores enrichment. Results suggest that the proposed integral representation of clinical utility in combination with Bayesian methods provide a practical strategy to facilitate cut-off decision making in this setting.

Higher Order Asymptotics for Generalized Fiducial Distribution

Abhishek Pal Majumder
University of North Carolina at Chapel Hill
`palmabhishek@gmail.com`
Coauthors: Jan Hannig

When no prior information is available about the parameter space it's common to use non-informative prior despite its undesirable performances in some pathological cases. The study of Generalized Fiducial Inference has its own merit since it evolved from an extension of Fisher's Fiducial argument where so called "prior" depends on the data generating structural equation (i.e $X = G(U, \theta)$). Now if we look at the transformation A on the data generating equation $A(X) = A \circ G(U, \theta)$ then one can associate a weight function $w() = A'()$ in the Jacobian part of the Fiducial distribution. Under few strong assumptions, it can be shown that if a smooth transformation A is a minimal sufficient statistics then the fiducial quantile is exact in classical coverage sense. But if those assumptions aren't met, we used the idea of probability matching prior to get conditions under which the Fiducial quantile will be asymptotically first/second order exact. Conditions are similar to Bayesian paradigm but difference occurs in the asymptotic expansion of fiducial quantile since due to the empirical structure, one has some noise terms inside of both first and second order terms. This work bridges and analyze two paradigms on a common ground and raise some important questions on the validity of the first/second order terms.

Dynamic Spatial Evaluation of Deterministic Air Quality Models Using Monitoring Station Observations

Elizabeth Mannshardt
North Carolina State University
`mannshardt@stat.ncsu.edu`
Coauthors: Montserrat Fuentes, Soumendra Lahiri, and Kristen Foley

The Air Quality System (AQS) network of monitoring stations provides a rich source of information about air pollution. However, station observations do not allow consideration of different emission and meteorological scenarios. Air quality models can be used to predict changes in air quality due to changes in emissions or meteorology. The Community Multiscale Air Quality System (CMAQ) allows for estimations under different climate scenarios, but is not as accurate as network sensors. Of interest is the comparison of modeled ozone output to observed ozone, as well as how ozone changes in response to emissions and

meteorology. We consider ozone observations from AQS monitoring stations to evaluate CMAQ ozone estimates. We construct spatial quantile functions for observed ozone versus modeled ozone, considering the quantile of each metric that corresponds to the EPA standard for ozone emissions. A dynamic evaluation approach explicitly focuses on predicted pollutant responses due to changes in emissions or meteorology. Findings from such dynamic evaluation studies can be directly relevant in regulatory decisions.

Nonnegative Nested Cone Analysis

J. S. Marron
University of North Carolina
`marron@unc.edu`
Coauthors: Lingsong Zhang, Dept. of Statistics, Purdue University

The usefulness of Principal Component Analysis as a statistical visualization device is illustrated. In a number of situations, maximal interpretability comes from nonnegative loadings and scores. The standard Nonnegative Matrix Factorization solution suffers in multi-scale applications (such as visualization) because the solutions are not nested over rank of approximation. The recently developed notion of Backwards Principal Component Analysis is seen to lead to a useful solution to this problem, giving a new method called Nonnegative Nested Cone Analysis.

Efficient Computation of the Distribution of Spaced Seed Coverage

Donald E.K. Martin
North Carolina State University
`martin@stat.ncsu.edu`
Coauthors: Laurent Noe, LIFL (UMR 8022 Lille 1 University/CNRS) - INRIA Lille Nord-Europe, Villeneuve d'Ascq, France

When using an auxiliary Markov chain to compute probabilities for pattern statistics, the computational complexity is directly related to the number of Markov chain states. Thus in recent years, minimal deterministic finite automata have been used as data structures that facilitate computation while keeping the number of states at a minimum. For statistics where overlapping and non-overlapping pattern occurrences are treated differently, one could form an extended automaton that includes prefixes of initial and overlapping word occurrences, and then minimize the extended automaton. However, there are situations where forming a full extended automaton and then minimizing it is computationally expensive. We give a method to bypass the formation of a full extended automaton before minimization that facilitates efficient computation. The method is applied to the distribution of the number of sequence positions covered by spaced seed hits, a pattern matching paradigm that has proven fruitful in similarity searches for DNA sequences.

On Optimal Inferential Models

Ryan Martin
University of Illinois at Chicago
`rgmartin@uic.edu`

The inferential model (IM) framework provides provably valid probabilistic inference in cases where no genuine prior information is available. This is accomplished through the use of random sets for predicting unobservable auxiliary variables associated with the observable data and unknown parameters. The efficiency of the corresponding IM is related to the choice of this random set. In this talk, I will present a theory of optimal predictive random sets. Besides efficiency gains, this optimality theory also removes any apparent subjectivity in the choice of association or random set, since the resulting IM depends only on the sampling model. Examples, extensions, and open questions will also be discussed.

Variation between Self- and Mutual Assessment in Animal Contests

Mike Mesterton-Gibbons
Florida State University
`mesterto@math.fsu.edu`
Coauthors: Stephen M. Heap (Department of Biological and Environmental Science, University of Jyväskylä, Finland)

Limited resources lead animals into conflicts of interest, which are resolved when an individual withdraws from a direct contest. Current theory suggests that the decision to withdraw can be based on a threshold derived from an individual's own state (self-assessment) or on a comparison between their own state and their opponent's (mutual assessment). The observed variation between these assessment strategies in nature does not conform to theory. Thus, we require theoretical developments that explain the functional significance of different assessment strategies. We consider a hawk-dove game with two discrete classes that differ in fighting ability, in which the players strategically decide on their investment toward mutual assessment. Analysis of the model indicates that there are simultaneous trade-offs relating to assessment strategies. First, weak individuals in a population must decide on whether to acquire information about their opponents at the cost of providing opponents with information about themselves. Secondly, all individuals must decide between investing in mutual assessment and being persistent in contests. Our analysis suggests that the potential for individuals to make errors during contests and differences in the consequences of sharing information within a population may serve as fundamental concepts for explaining variation in assessment strategy.

An Adaptive Decision Making Algorithm to Manage the Spread of a Disease Over a Network

Nicholas J. Meyer
North Carolina State University
`nick.j.meyer@gmail.com`
Coauthors: John Drake, Eric B. Laber, Krishna Pacifici, Brian J. Reich

Managing the spread of a disease over a network is a complex decision making problem with both spatial and temporal components. Optimally choosing the times and locations to apply treatment requires a high quality model for the disease dynamics and the ability to use that model to simulate and optimize potential treatment strategies. In some settings the disease dynamics are non-stationary and the model must be re-estimated as new data become available. We construct an adaptive sequential treatment strategy that minimizes the disease spread over a network by a pre-specified finite time horizon. Using observed data, the method estimates a dynamics model to develop an initial treatment strategy. In real time, as data accrue, the initial strategy is updated using a stochastic approximation algorithm. The method is illustrated using a series of simulated experiments.

Optimality of Variance Balanced Designs and Their Constructions

Nutan Mishra
University of South Alabama
nmishra@southalabama.edu

Optimality properties of non-binary , variance balanced designs with unequal replications are discussed. The criteria for optimality considered are based on the canonical efficiency factors instead of the characteristic roots of the C-matrix. Further the necessary conditions for the existence of such designs and their constructions are given.

Measuring Health Outcomes in America Using a Multivariate Approach

Prashant Mittal
University of Southern Maine
pmittal@usm.maine.edu

It is challenging to accurately measure and compare health outcomes nationally. Several survey and surveillance systems collect health data on a yearly basis to make such comparisons. Researchers tend to investigate state level data for national comparisons with the underlying assumption that within each state the health outcomes are similar. The reality is unquestionably far from this notion. Each state behaves as its own ecosystem with a huge diversity of outcomes at smaller geographic boundaries like counties, cities and towns. Imagine a jig-saw puzzle made up of approximately 3,200 pieces of counties in the nation, rearranged not based on their geography but based on how sick or how healthy they are. The results will be anything but the geographic map of America. This paper makes a first of its kind attempt to study the phenomenon of diversity in self-reported health outcomes using a sophisticated multivariate approach. The study compares comorbidity and protective factors to create profiles of similar counties. Such analysis may inform policy makers, health organizations and service providers, so they better understand the diversity of health nationally and locally, and can allocate resources accordingly. This study uses Behavioral Risk Factors Surveillance System (BRFSS) national data. BRFSS is the worlds largest telephonic health surveillance system, conducted by the Centers for Disease Control and Prevention (CDC) in all 50 states and territories in the US. BRFSS uses a complex survey methodology and weighting scheme to measure behavioral factors and related health outcomes.

Multi-factor Youden Rectangles

John P. Morgan
Virginia Tech
jpmorgan@vt.edu
Coauthors: Donald A. Preece (Queen Mary University of London and University of Kent, Canterbury)

A *Youden rectangle* is a $k \times v$ row-column design on v treatments, where row blocks are complete and column blocks comprise a symmetric BIBD. This design also travels by the name *Youden square*, owing to the fact that if one interchanges the roles of rows and treatments, the result is a $v \times v$ square on k symbols with $v(v-k)$ empty cells. This square, termed a (v, k)-*Youden array*, has the properties that (i) each of the k symbols appears once in each row and once in each column, and (ii) if each of the k symbols is replaced by a one, and each empty cell is filled with a zero, the result is an incidence matrix for a BIBD.

Consider now two (v, k)-Youden arrays based on the same BIBD incidence matrix. For simplicity, take the k symbols in each of the individual arrays to be the integers $0, 1, \ldots, k-1$. The pair of arrays can then be considered as a single $v \times v$ array with an ordered pair of integers in each non-empty cell. This is a *double Youden array* if, throughout the array, the ordered pair p, q ($p = 0, 1, \ldots, k-1$; $q = 0, 1, \ldots, k-1$) occurs n_1 times if $p = q$, and n_2 times if $p \neq q$, where n_1 and n_2 are constants with $n_1 = n_2 \pm 1$. Extending this notion, if any two of m Youden arrays form a double Youden array, one has a (v, k, m)-*Youden array*, corresponding to a Youden rectangle for m factors.

For $m = 3$ two examples of triple Youden arrays are known, having $(v, k) = (13, 4)$ and $(15, 7)$. To our knowledge, no examples are known for $m > 3$. Here (v, k, m)-Youden arrays are constructed for $m = k - 4$ with $v = 2k + 1$, where k is an odd prime or prime power, and their properties as statistical designs are explored. This is the last paper in the long and distinguished career of Donald Preece, who died January 6, 2014.

Evaluation of Low-Dimensional Structures in High-Dimensional Twitter Data

Chitra Reddy Musku
The University of North Carolina at Greensboro
`c_musku@uncg.edu`
Coauthors: Shan Suthaharan

The complicated linguistic structure combined with terabytes of data makes the analysis of twitter data a challenging research problem. The high dimensionality of the data adds more complexity and it is difficult to apply Machine Learning techniques to model the data for prediction and decision making with these problems. In this case, preprocessing techniques (called representation learning) can help reduce the complexity.

Recently, a preprocessing technique using twitter data, related to Dutch keywords, has been proposed and reported in the Big Data literature. This technique extracts word count, phrase count and conversation count from terabytes of twitter data using Hadoop Distributed File System (HDFS). In our research, we collect our own twitter data and transform phrases from keywords (e.g. iPhone 6 and Samsung Galaxy) and its neighboring words within a predetermined window size, and then hash the phrases to construct low dimensional structures. This transformation helps reduce the dimensionality of data, volume and linguistic complexity. We then execute the preprocessing techniques in the hashed domain, hence leads to faster processing and reduced complexity in HDFS. We present our twitter data, results and findings at the conference.

Big Data Storage And Management

Christopher Nyamful
NC A&T State University
`cnyamful@aggies.ncat.edu`
Coauthors: Rajeev Agrawal

The amount of data generated daily by industries, large organizations and research institute is increasing at a very fast rate. These huge volumes of data need to be kept not just for analytic purposes, but also in compliance with laws and service level agreements to protect and preserve data. Storage and management are major concern in this era of big data. The ability for storage devices to scale to meet the rate of data growth, enhance access time and data transfer rate is equally challenging. These factors, to a considerable extent, determine the overall performance of data storage and management. Big data storage requirements are complex and thus needs a holistic approach to mitigate its challenges. This survey paper analyses current technologies and techniques available for big data storage and management. In addition, we also provide useful suggestions to meet the demands posed by big data storage and management.

Co-Directing from Two Institutions for the National Research Experience for Undergraduates Program

Hyunju Oh
Bennett College
`hoh@bennett.edu`
Coauthors: Jan Rychtar

In this talk, we will present our two years of experiences in the National Research Experience for Undergraduates Program funded by Mathematical Association of America. Jan Rychtar at the University of North Carolina at Greensboro and Hyunju Oh at the Bennett College co-directed students projects in Game Theory and Applications during the six weeks each summer 2013 and 2014. We will report how we support African-American female undergraduate students from Bennett College in research experience and the students achievements, and how we collaborate between two project directors in two different institutions.

Big Data Scale Content Based Image Retrieval

Jordan Paris
Marshall University
`paris2@live.marshall.edu`
Coauthors: Dr. Gudivada

Digital images are being generated at unprecedented levels by both humans and machines. Cameras embedded in mobile devices and smart camera networks are the primary contributors of these images. In 2013 Facebook reported that over 350 million images per day are uploaded to its site. However, image search is still based on textual descriptors and other metadata associated with the images. Ability to search large image collections based on their contents is an active research problem.

In this talk, we describe our approach to content based image retrieval that can scale to Big Data image collections. We divide the image into non-overlapping blocks of equal size and compute various pixel-based metrics to characterize the image blocks. Using these block metrics, we create a structure similar to a merkle tree. In essence, each image is represented by a merkle tree and similarity between two images is cast as a problem of finding similarity between merkle trees. We demonstrate the effectiveness of our approach using the images in the Geograph collection.

Case-Control Sampling for Brain Imaging and Enhancement Prediction

Gina-Maria Pomann
North Carolina State University
`gina.pomann@gmail.com`
Coauthors: A. Staicu, R.T. Shinohara, and E. Sweeney, and D.S. Reich

Multiple sclerosis (MS) is a disease associated with inflammation of the brain, morbidity, and disability. Irregular blood flow in white matter of the brain is associated with clinical relapses in patients with MS. This abnormality in active MS lesions is commonly evaluated through identification of enhancing voxels on Magnetic Resonance (MR) images. Current clinical practice includes intravenous injection of contrast agents that are occasionally toxic to the patient and can increase the cost of imaging by over 40%. Local image regression methodology has recently been developed to predict enhancing MS lesions without using such contrast agents. We extend this model to account for the rarity of enhancement as well as to incorporate historical information about lesions. Incorporation of a historical covariate which characterizes newly enhancing lesion behavior is found to substantially improve prediction. We consider 77 brain MR imaging studies on 15 patients which include historical imaging covariates along with T1-w, T2-w images. Additionally, we present a novel scan-stratified case-control sampling technique that accounts for the rarity of enhancement and reduces the computational cost.

Autoregressive-Stochastic Volatility with Threshold Nonlinear Time-Series Model

Prajneshu
Indian Agricultural Statistics Research Institute, India
`prajneshu@yahoo.co.in`

A parametric nonlinear time-series model, viz. Autoregressive-Stochastic volatility with threshold (AR-SVT) model with mean equation for forecasting level and volatility is proposed. Methodology for estimation of parameters of this model is developed by first obtaining recursive Kalman filter time-update equation and then employing Unrestricted quasi-maximum likelihood method. Further, optimal multi-step ahead out-of-sample forecasts formulae along with forecast error variances are derived analytically by recursive use of conditional expectation and variance. As an illustration, volatile all-India monthly spices export is considered. Entire data analysis is carried out using EViews and MATLAB software packages. The AR-SVT model is fitted and interval forecasts for ten hold-out data points are obtained. Superiority of this model for describing and forecasting over other competing models for volatility is shown for the data under consideration

Undergraduate Research at Delhi University, India: An Overview

Meenakshi Prajneshu
University of Delhi, India
`prajneshum@gmail.com`

Aim of undergraduate research is the supportive, encouraging, and intellectual partnership among students as well as between students, their faculty and project mentor through which they apply knowledge gained in the classrooms to real-life problems. Delhi University constituent colleges have been challenged to involve students in hands-on research experiences to enhance their learning. Interdisciplinary research is conducted through teams of ten students from two or three disciplines, three faculty members and one mentor. Every year, a Symposium is organized by each team in which imminent scholars give lectures and interact with scholars to make them interested in result-oriented research. Teams also present their work to the whole University and outside World by participating in an annual programme of the Universitys Antardhwani. Essentially, students become junior partners in faculty research. This promotes students to become scholars. Undergraduate research at Delhi University colleges started in the year 2012 in a modest way but very next year, more than 250 projects were sanctioned. Salient aspects of some completed as well as some ongoing projects will be described in the talk. Rigorous monitoring mechanism being followed will also be discussed. My own experiences as a Coordinator since inception of this programme will be highlighted.

Measuring the Lack of Monotonicity in Functions

Danang Teguh Qoyyimi
Gadjah Mada University, Indonesia
`dqoyyimi@uwo.ca`
Coauthors: Ričardas Zitikis (University of Western Ontario, Canada)

Problems in econometrics, insurance, reliability engineering, and statistics quite often rely on the assumption that certain functions are non-decreasing. To satisfy this requirement, researchers frequently model the underlying phenomena using parametric and semi-parametric families of functions, thus effectively specifying the required shapes of the functions. To tackle these problems in a non-parametric way, in this presentation we suggest indices for measuring the lack of monotonicity in functions. We investigate properties of the indices and also offer a convenient computational technique for practical use.

Canonical Discriminate Analysis of Functional Metagenomes

Dhana Rao
Marshall University
`raod@marshall.edu`
Coauthors: Venkat N. Gudivada

Canonical discriminate analysis of functional metagenomes Metagenomics is a relatively new area that applies genomics techniques to study microbial communities in their natural environments. It has shown tremendous potential to drive the discovery of the many microorganisms that cannot be cultured. Analyses of a metagenomic dataset provide insights into the functionality and diversity of the corresponding microbial community. However, analyses of datasets can be challenging because of sequencing errors, chimeric reads and complex underlying biology with high inter-subject variability. Furthermore, the sheer volume of data makes it difficult to identify key differences in the function and taxonomy between communities. Canonical discriminate analysis (CDA) can be used to parse out nuances that separate different environments. CDA combines the dimensionality of principal component analysis (PCA) with the classification of K-means approaches. In our project we present results of applying this approach to re-analyze sequence data from multiple environmental metagenomic datasets with varying complexity and sampling levels.

Critique of Trends in Phase II Trials

Russell Reeve
Quintiles
`russell.reeve@quintiles.com`

Phase II is a critical time in drug development. Success in answering key questions in this phase has been shown to improve success in Phase III. In this talk, we talk about recent trends in the design of Phase II trials. Adaptive trials are becoming more frequently observed, with many options available, in both oncology and non-oncology settings. We shall discuss some of these designs, and offer suggestions for improvement. In many settings, fixed designs are still preferable, and here we shall discuss the current state of Phase II designs. The learn-and-confirm paradigm of Sheiner, harkening to the TQC philosophy of Deming, is starting to play a larger role in Phase II development, but still has substantial room for improvement. We shall discuss how its use could improve both Phase II development, but also improve the odds for success in Phase III.

Securitization and Equilibrium Pricing Under Relative Performance Concerns

Goncalo dos Reis
University of Edinburgh, UK
`G.dosReis@ed.ac.uk`
Coauthors: Jana Bielagk (Humboldt University of Berlin, Germany)

We investigate the effects of a finite set of agents interacting socially in an equilibrium pricing mechanism.

A derivative written on non-tradable underlyings is introduced to the market and priced in an equilibrium framework by agents who assess risk using convex dynamic risk measures expressed by Backward Stochastic Differential Equations (BSDE). An agent is not only exposed to financial and non-financial risk factors, but he also faces performance concerns with respect to the other agents. The equilibrium analysis leads to systems of fully coupled multi-dimensional quadratic BSDEs.

Within our proposed models we prove the existence and uniqueness of an equilibrium. We show that aggregation of risk measures is possible and that a representative agent exists. We analyze the impact of the problem's parameters in the pricing mechanism, in particular how the agent's concern rates affect prices and risk perception.

Protecting Confidentiality in an Era with No Privacy

Jerome Reiter
Duke University
`jerry@stat.duke.edu`

Data stewards seeking to share large-scale social science data face a difficult challenge: they have to release data that (i) protect privacy and confidentiality, (ii) are informative for a wide range of analyses, and (iii) are relatively straightforward for a large number of data analysts to use. As the size, richness, and quality of social science data have increased, so too have the threats to confidentiality and the scope of the dissemination challenge. In this talk, I discuss some of the approaches taken by statistical organizations when sharing confidential data. I focus particularly on releasing synthetic data, i.e., data simulated from models estimated with the confidential data, which is emerging as a potential solution to data sharing.

Imputation of Missing Data via Flexible, Mixture Modeling

Jerome Reiter
Duke University
`jerry@stat.duke.edu`

Multiple imputation is a common approach to handling missing values. It is usually implemented using either chained equation approaches (as in the software MICE and IVEWARE) or joint models. For the latter, analysts typically use a multivariate normal or log-linear model as imputation engines. While useful, these models impose sets of assumptions that may not be satisfied in many data sets. In this talk, I describe several imputation engines based on nonparametric Bayesian models. I present comparisons against default (main effects only) implementations of chained equations in simulated and actual data.

Wilks Λ and the Growth Curve Model

Dietrich von Rosen
Swedish university of agricultural sciences
`Dietrich.von.rosen@slu.se`

Let $\mathbf{W}_{H_0}$ and $\mathbf{W}_{H_1}$ be two independently distributed Wishart matrices which build up Wilks Λ, i.e.

$$\Lambda = \frac{|\mathbf{W}_{H_0} + \mathbf{W}_{H_1}|}{|\mathbf{W}_{H_1}|}.$$

The matrices appear when testing H_0: $\mathbf{BG} = \mathbf{0}$ versus H_1: $\mathbf{B}$ unrestricted in a MANOVA model, i.e.

$$\mathbf{X} = \mathbf{BC} + \mathbf{E},$$

where $\mathbf{X}$ is a random matrix which represents the observations, $\mathbf{C}$ and $\mathbf{G}$ are a known matrixes, and $\mathbf{E} \sim N_{p,n}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I})$, where $\mathbf{B}$ and $\mathbf{S}igma$ are unknown matrices. The distribution for Λ equals a product of independently beta-distributed variables. When approximating the distribution several approaches are available, where the most commonly applied uses approximations of the gamma-function.

Somewhat remarkable is an old classical result which states that the likelihood ratio test for testing H_0: $\mathbf{FBG} = \mathbf{0}e$ versus H_1: $\mathbf{B}$ unrestricted in a GMANOVA model, i.e.

$$\mathbf{X} = \mathbf{ABC} + \mathbf{E},$$

where the difference with the MANOVA model is the known matrix $\mathbf{A}$ also follows a Wilks Λ distribution. It is remarkable since the maximum likelihood estimators in the MANOVA and GMANOVA are very different. The talk will derive the distribution in a somewhat different way than what usually is applied which also sheds some light on some conditional arguments.

Autoregressive Methodology In Detection of Irregular Spiking Activity in Neurons

Sudip Roy
University of Texas at San Antonio
`sudipstat@gmail.com`
Coauthors: Dr. Daijin Ko

Neuron emits different firing patterns such as bursts or pauses, which are important for information processing in the brain. In dopaminergic neurons, prominent patterns include repetitive bursts, regular pacemakers and pauses. Bursts are more important than single spike for evoking responses in post synaptic cells. We have introduced a method, robust way of detecting additive and innovative patches outliers considering them as bursts and pauses from contaminated autoregressive time series model of the single neuron spike train, based on inter spike interval (ISI) length. Additive outlier in time series model is the type that affects a single observation. After this disturbance, the series returns to its normal path as if nothing happened. In neuroscience spike train analysis we call it as doublet. Innovation outliers are a kind that affects subsequent observations starting from its position or an initial shock that propagates in the subsequent observations, also characterized as additive patches of outliers. Similar things we look for in burst clusters and pause clusters in neuron spike train. We have standardized the log transformed ISIs (SLISI) with robust location and scale parameters and we will use this transformed data to analyze the behavior of neural spike train. The robust serial correlation and neurophysiology literature show autoregressive nature of spike trains. Robust AR regression model with high breakdown point (Gervini and Yohai, 2002) is fitted as the model for regular firing pattern. The residuals from the fitted model are used in calculating significance of outlyingness. In particular we developed an automated data adaptive method to identify outliers based on the distribution of the standardized residuals and detect extended clusters of bursts and pauses. My method is a modification to Ko et al., 2012, Detection of bursts and pauses method in spike trains where I have taken into account the autocorrelation of the inter event time (ISI) with its past few ISIs. In Ko et. al., Gaussian Surprise Method the central part of the spike train are determined by the function of E-center. In our method we have used robust autoregressive model to recognize the central part of the spike train. We have not go further with other time series modeling like Moving Average process or ARIMA to make our analysis simple, function of the current ISI based on its near past ISIs. The distribution assumption is made for the residuals of the fitted time series model for the ISIs of the spike train and based on it the outliers are being detected. The question lies why the outliers are of concern to get the initial burst or pause seeds. The reason behind for doing so is presumably the bursts or pauses are outlier subsequences whose frequency differs from the expected frequency got from the usual distribution of the regular firing spikes in the neuron. So if we can robustly model the regular spiking activity then significantly large positive and negative residuals of the model can be due to those pause and burst respectively. My method furthermore attenuate the fact that detection of outliers based on the residuals is data adaptive and not only lies on the threshold based on the distribution of the residuals itself. The similarity with the Random Gaussian Method by Ko et. al, is again the technique to evaluate the extension of burst or pause patches from its seed we have used False Discovery Rate (FDR) for multiple testing whereas in the former method more conservative Bonferroni multiple testing method is been used . Our motivation of the study is to find a proper distribution for the background regular firing part of the data which is more robust in nature and based on which we can detect the irregular firing considering them as significant outlier from our distribution, considering the negative autocorrelation of the ISIs.

Mentoring Interdisciplinary Research Projects

Jan Rychtář
The University of North Carolina at Greensboro
`rychtar@uncg.edu`

In 2006 faculty members within the Departments of Biology, Mathematics and Statistics and the Office of Undergraduate Research at the University of North Carolina at Greensboro came together to develop an undergraduate MathBio program supported by the National Science Foundation. Our objectives are to generate new knowledge at the interface of mathematics and biology, showcase the importance of mathematics outside the discipline, increase the use of mathematics and statistics in the field of biology, and guide students to an increased proficiency of research skills while preparing them for interdisciplinary graduate education. During this talk, we will address how we began and how we evolved the program over the years; talk about the successes and pitfalls of the program and discuss challenges we face as we move forward.

Ideal Cost-Free Distributions in Structured Populations for General Payoff Functions

Jan Rychtář
The University of North Carolina at Greensboro
`rychtar@uncg.edu`
Coauthors: Mark Broom, City University London, UK

The important biological problem of how groups of animals should allocate themselves between different habitats, has been modelled extensively. Such habitat selection models have usually involved infinite well-mixed populations. In particular the problem of allocation over a number of food patches when movement is not costly, the ideal free distribution (IFD) model is well-developed. Here we generalize (and solve) a habitat selection game for a finite structured population. We show that habitat selection in such a structured population can have multiple stable distributions (in contrast to the equivalent IFD model that is practically unique). We also define and study a "predator dilution game" where unlike in the habitat selection game, individuals prefer to aggregate (to avoid being caught by predators due to the dilution effect) and show that this model has a unique solution when movement is unrestricted.

Information Security in NoSQL Systems for Big Data

Jothilakshmi S
Annamalai University
`jothi.sekar@gmail.com`
Coauthors: Venkat N Gudivada (Marshall University)

Special needs of Big Data recently ushered in several new classes of systems for data management. They are referred to in the literature by various names including NoSQL, NewSQL, and non-RDBMS. These systems differ greatly in data models employed, architectures used for scalability and performance, approaches used for high availability, query languages and client interfaces provided. More importantly, some systems provide rudimentary security features and other provide no security and assume operation in a trusted environment. Given the availability of numerous NoSQL systems, selecting a right system for a given application is not a trivial task. To address this problem, Database as a Service (DBaaS) is emerging as a solution. DBaaS enables organizations to effectively leverage NoSQL choices by matching the right data model and query language with the type of data without requiring in-house resources such as specialized and dedicated hardware and personnel. However, information security is an unsolved problem in DBaaS. In this talk, we describe information security issues of NoSQL systems under the DBaaS model. A comparative study of security features of existing NoSQL systems will be discussed including their strengths and limitations. Emergent security challenges that come into play when multiple NoSQL systems operate under DBaaS model will also be presented.

Using Meta-Analysis to Obtain Desired Random Effect Structures in Cluster-Randomized Studies

Roy T. Sabo
Virginia Commonwealth University
`rsabo@vcu.edu`
Coauthors: Ghalib Bello (Oklahoma Medical Research Center), Aobo Wang (Virginia Commonwealth University), Alexander Krist (Virginia Commonwealth University)

Analysis of data from cluster-randomized trial (CRT) designs can be complicated by (i) multi-layered or nested clusters, (ii) large data-bases or cohorts, (iii) non-continuous outcomes, and (iv) repeated measure outcomes, each of which requires sophisticated optimization techniques. The presence of any of these features or their combination can reduce the likelihood of estimator convergence and prohibit conceptual inference all-together. Fortunately, meta-analytic methods can be adjusted to allow for parameter estimation in CRT designs featuring the aforementioned characteristics, while continuing to capture and account for the complicated cluster design. This methodology effectively partitions the existing sample into several manageable samples, based on either independent sampling or through use of the existing clusters. We show that this method allows for parameter estimation and inference in complicated situations, as shown in a contextual example from a cancer screening study. The results suggest an iterative process of estimation works best.

Sudokus as Experimental Designs

Jyotirmoy Sarkar
Indiana University - Purdue University Indianapolis
`jsarkar@iupui.edu`
Coauthors: Bikas Kumar Sinha, ISI (retired)

Viewing Sudokus as experimental designs, one can go a step beyond traditional Latin Square Designs. Subramani and Ponnuswamy (2009) introduced such an additional [third] component of variation in a Sudoku. But they did not properly account for the loss of orthogonality. A correct analysis was given by Saba and Sinha (2014). We recover some lost orthogonality by constructing cylindrical-shift Sudokus by suitable composition of Latin Squares. We extend the idea to Sudokus of arbitrary order. Such Sudoku designs provide extra utility and flexibility to the experimenters.

On a Generalized Class of Mean Estimators in Survey Sampling

Javid Shabbir
Quaid-I-Azam University, Pakistan
`javidshabbir@gmail.com`
Coauthors: Sat Gupta

We propose an improved generalized class of estimators of finite population mean that leads to many more classes of estimators including Singh and Solanki (2013) estimator. Expressions for bias and mean squared error for the proposed class of estimators are obtained up to first order of approximation. The proposed class of estimators is more efficient than many other competing estimators.

Perceptions of Statistical Consulting in a University Setting

Julia Sharp
Clemson University
`jsharp@clemson.edu`
Coauthors: Pat Gerard, James Wrenn

The role that statistical consulting should play for an academic statistician has long been debated. The aim of this study is to determine the perceived value of statistical consulting in a university setting. To gauge this value, a survey was designed to obtain responses from statistical consultants in academia. Statistical consultants were asked questions about their personal value of statistical consulting and the perceived value their universities place on statistical consulting. Our results show that statistical consulting is beneficial to the university, department, and consulting statistician. A majority of statistical consultants strongly agreed or agreed that their consulting duties were and should be taken into consideration for tenure and partial fulfillment of research/publication requirements. A majority of statistical consultants not in a center were either neutral or did not agree that their statistical consulting duties are taken into consideration when considered for a pay raise.

A Modified Binary Optional RRT Model

Jeong Sep Sihm
The University of North Carolina at Greensboro
`j_sihm@uncg.edu`

We propose a modified two stage binary optional randomized response technique (RRT) model in which respondents are given the option of answering a sensitive question directly without using randomization device in the second stage if they think they don't need extra measures to protect their privacy. This particular approach was proposed by Sihm and Gupta (2013) previously. However, instead of using the split sample technique and estimating the sensitivity level of the research question and the prevalence of the sensitive characteristic simultaneously, we propose to use the Warner (1965) method twice. First to estimate the sensitivity level of the research question, and then to estimate the prevalence of the sensitive characteristic in the same sample by asking two separate questions using two separate randomization devices. Simulation results of this new model are compared with those of the split sample based two stage binary optional RRT model. Computer simulation shows that the proposed model has smaller variance than the previous two stage optional binary RRT model for two parameters.

Improved Exponential Methods Of Imputation In Estimation Of Current Population Mean In Two-Occasion Successive Sampling

G. N. Singh
Indian School of Mines, India
`gnsingh_ism@yahoo.com`
Coauthors: M. Khetan and S. Maurya

This work suggests some improved exponential methods of imputation for missing data due to non-response and subsequently effective estimation procedures have been proposed to estimate the current population mean in two-occasion successive sampling. Properties of the proposed estimation procedures have been examined and optimum replacement strategies are suggested which are helpful in reducing the cost of the survey. Empirical studies are carried out to assess the performance of suggested estimation procedures and suitable recommendations have been made to the survey practitioners.

Test for the Mean in a Growth Curve Model for High Dimensions

Martin Singull
Linköping University
`martin.singull@liu.se`
Coauthors: Muni S. Srivastava (University of Toronto)

In this paper, we consider the problem of estimating and testing a general linear hypothesis in a general multivariate linear model, the so called Growth Curve Model, when the $p \times N$ observation matrix is normally distributed with an unknown covariance matrix.

The maximum likelihood estimator (MLE) for the mean is a weighted estimator with the inverse of the sample covariance matrix which is unstable for large p close to N and singular for p larger than N. We modify the MLE to an unweighted estimator and propose a new test which we compare with the previous likelihood ratio test (LRT) based on the weighted estimator, i.e., the MLE. We show that the performance of the LRT and the new test based on the unweighted estimator are similar.

For the high-dimensional case, when p is larger than N, we construct two new tests based on the trace of the variation matrices due to the hypothesis (between sum of squares) and the error (within sum of squares).

To compare the performance of these four tests we compute the attained significance level and the empirical power.

Optimal Designs for Generalized Linear Mixed Models

Sanjoy Sinha
Carleton University, Canada
`sinha@math.carleton.ca`
Coauthors: Xiaojian Xu, Brock University

Generalized linear mixed models (GLMMs) are commonly used for analyzing clustered correlated discrete binary and count data including longitudinal data and repeated measurements. We explore techniques for the design of experiments, where the design issue is formulated as a decision of choosing the values of the predictor(s) for GLMMs. The usual measures of performance of a design in GLMMs depend on the parameters being estimated. We investigate a sequential approach to choosing a design point that maximizes a measure of performance evaluated at the estimates obtained from observations made at previous design points. We assess the performance of the proposed design using simulations.

Influence of Climate Change on Extreme Weather Events

Richard Smith
University of North Carolina and SAMSI
`rls@email.unc.edu`
Coauthors: Michael Wehner, Lawrence Berkeley National Laboratory

The increasing frequency of extreme weather events raises the question to what extent such events are attributable to human causes. The human influence may be characterized through the fraction of attributable risk (FAR) or equivalently the risk ratio (RR), which is the ratio of the probability of a given extreme event under an anthropogenic forcing scenario to the same probability when only natural forcings are taking into account. However, there is no generally accepted method of calculating these quantities. We propose a method based on extreme value theory, incorporated into a Bayesian hierarchical model for combining climate models runs and the observational record. The same method solves the closely related question of projecting changes in the probability of extreme events over some future period. These methods are applied to three extreme events: (a) the extreme heatwave that occurred in parts of Europe in 2003, (b) the heatwave of Russia in 2010, and (c) the central USA heatwave of 2011. In each case we find posterior median risk ratios (from anthropogenic versus control-run models) of between 2 and 3, with wide posterior intervals, but still at least a two-thirds probability that the true risk ratio is greater than 1.5, implying substantial anthropogenic influence. Projections of future probabilities for the same three extreme events show substantial differences: an event equivalent to Europe in 2003 has nearly a 90% probability of reoccurrence in any given year by 2040; corresponding results for the other two events also show an increase, but much more slowly.

Assess the Uncertainty of Probability of Success in Phase III Trial after Phase II Data

Guochen Song
Quintiles
`guochen.song@quintiles.com`
Coauthors: Ilya Lipkovich; Alex Dmitrienko

The Probability of Technical Success (PTS) assesses the probability that a clinical development program at a particular phase will meet its predefined efficacy and safety criteria and advance to the next stage of development. However, the PTS depends heavily on assumptions. As PTS is calculated under the Bayesian framework, Bayesian robustness assessment can be used to evaluate the uncertainties of such calculation.

Bonds Historical Simulation Value at Risk

J. Beleza Sousa
M2A/ISEL/IPL CMA/FCT/UNL
`jsousa@deetc.isel.ipl.pt`
Coauthors: M. L. Esquvel, R. M. Gaspar, P. C. Real

Bonds historical returns can not be used directly to compute VaR by historical simulation because the maturities of the interest rates implied by the historical prices are not the relevant maturities at time VaR is computed. In this paper we adjust bonds historical returns so that the adjusted returns can be used directly to compute VaR by historical simulation. The adjustment is based on the prices, implied by the historical prices, at the times to maturity relevant for the VaR computation. Besides other features, we show that the obtained VaR values agree with the usual market trend of smaller times to maturity being traded with smaller interest rates, hence, carrying smaller risk and consequently having a smaller VaR.

Improved Mean Estimation Using Auxiliary Information in Stratified Sampling

Rita Sousa
New University of Lisbon
`ritasousa03@hotmail.com`
Coauthors: Sat Gupta, Javid Shabbir and Pedro Corte-Real

In survey research, there are many situations when the primary variable of interest is sensitive. The sensitivity of some queries can give rise to a refusal to answer or to false answers given intentionally. Survey can be conducted in a variety of settings, in part dictated by the mode of data collection, and these settings can differ in how much privacy they offer the respondent. A variety of techniques have been used to improve reporting by increasing the privacy of the respondents. The Randomized Response Technique (RRT), introduced by Warner in 1965, develops a random relation between the individual's response and the question. This technique provides confidentiality to respondents and still allows the interviewers to estimate the characteristic of interest at an aggregate level. Sousa et al. (2010) introduced a ratio estimator for the mean of a sensitive variable which performs better than the ordinary mean based on a RRT. Gupta et al. (2012) proposed a regression estimator that performs better than the ratio estimator based on the same technique. The present study considers the RRT focusing on estimating the mean of a sensitive variable in stratified random sampling. In this paper, we present the combined ratio and regression estimators using auxiliary information in order to estimate the population mean from a population divided into strata. The performance of the proposed estimators is evaluated based on a simulation study. Similarly to simple random

sampling, it is shown that the proposed estimators perform better than the ordinary mean estimator which does not use auxiliary information unless in the strata definition. Also, we show that the proposed estimators perform better than the corresponding estimators based on simple random sampling.

Tests For Covariance Matrices in High Dimension with Less Sample Size

Muni Srivastava
University of Toronto, Canada
`srivastava@sympatico.ca`
Coauthors: Hirokazu Yanagihara, Hiroshima University Tatsuya Kubokawa, University of Tokyo

In this article, we propose tests for covariance matrices of high dimension with fewer observations than the dimension for a general class of distributions with positive definite covariance matrices. In one-sample case, tests are proposed for sphericity and for testing the hypothesis that the covariance matrix Σ is an identity matrix, by providing an unbiased estimator of $tr[\Sigma^2]$ under the general model which requires no more computing time than the one available in the literature for normal model. In the two-sample case, tests for the equality of two covariance matrices are given. The asymptotic distributions of proposed tests in one-sample case are derived under the assumption that the sample size $N = O(p^\delta)$, $1/2 < \delta < 1$, where p is the dimension of the random vector.

Economic and Mathematical Models of Credit Risk Assessment in Commercial Bank

O.P. Stefanets
Kyiv National University of Trade and Economics
`o.stefanets@mail.ru`
Coauthors: O.A.Kharchenko

The purpose of the study is to develop a model of risk assessment of the commercial bank loan portfolio. As a result of the research basic phases of the commercial bank credit risk management have been substantiated. Strengths and weaknesses of the most famous models of credit risk assessment - Z-Score, Basel 2, Credit Metrics, Moody's KMV Portfolio Manager, CreditRisk +, CreditPortfolioView have been analyzed. Requirements for their use in Ukrainian practice have also been formulated. The analysis of practice of using and development methodology of the statistical model of credit risk assessment proved that there are no unique universally recognized methods for building credit risk assessment models. The problem of developing a model is complicated by the fact that local banks have not accumulated enough data to build such models. Periodic revision is required for a number of parameters. Borrowers risk rating system based on the default probability model which takes into consideration both quantitative and qualitative factors has been developed. Two algorithms for credit risk modeling using a Monte Carlo based method for obtaining empirical distribution function of the portfolio loss by taking correlation of the borrowers defaults probability into account have been proposed. For this purpose covariance loss matrix is applied in the first algorithm, and covariance default probability matrix is used in the second one. A cumulative default probability prediction method based on auto-regressive model has been recommended. Cumulative probabilities calculations from 1998 to 2013 based on data from Moody's have been made. Calculations confirm that the general form of the auto-regressive model of the cumulative default probability time dependence is maintained for long periods of time. Economic and mathematical model for determining the cumulative default probability based on autoregressive models with supplement of additional variables has been created.

Modeling Savanna Water Resource Dynamics with Stochastic Daily Rainfall

Christopher Strickland
SAMSI and the University of North Carolina, Chapel Hill
`strickla@math.colostate.edu`
Coauthors: Adam Liedloff

Modeling has become an essential part of understanding ecosystem dynamics, and within the savanna ecology community, models are used as a key tool to advance theories about the determinants of savanna as an ecological state between forest and grassland. Many of these models adopt mean annual precipitation (MAP) as the primary variable for water resources, based on the observation of Sankaran *et al.* (2005) that MAP is highly correlated with an upper bound on maximum woody cover in African savannas and their subsequent hypothesis that MAP can be used to determine climatic savanna stability. In this talk, I will introduce a new water-resource model based on FLAMES, a process-based, demographic model developed by Liedloff and Cook (2007) for Australian savannas. Analysis of this model using stochastic, daily rainfall distributions suggests that length and severity of the dry season is the true driver behind climatic savanna equilibria, and also provides a method to predict the total basel area and structure of a stand given the local rainfall record. Finally, I will describe the effect of various fire regimes on the model, and the implications of these results for management.

On Optimal Designs for Nonlinear Mixed Effects Models

John Stufken
Arizona State University
`jstufken@gmail.com`
Coauthors: Linwei Hu, University of Georgia

Identifying optimal designs for correlated data is a problem for which few tools have been developed. Many classical results established for independent data have no obvious generalization to correlated data. Our proposed method identifies locally optimal designs for a class of linear and nonlinear mixed effects models under commonly used optimality criteria by extending results for independent data. We demonstrate our method by an example, and investigate robustness of design efficiency to mis-specification of the covariance matrix for the random effects.

Greater Mobility and Larger Neighborhood Size Inhibit the Evolution of Cooperation

David A. Suarez
The University of North Carolina at Greensboro
`d_suarez@uncg.edu`
Coauthors: Elizabeth Bergen (Cornell University) Jonathan Rowell (UNCG) Jan Rychtar (UNCG) Praveen Suthaharan (NCSU)

Researchers have long sought to identify which mechanisms can overcome basic selection pressures and support the establishment of cooperation within a population. In a slight variation of their traditionally conception, cooperators are individuals who pay a cost for another individual to receive a benefit, whereas wildtype individuals are free-riders or defectors who receive but do not provide benefits, thus paying no cost. In models of well- mixed populations, free-riders are favored; however, cooperation can emerge and evolve in spatially structured populations. In this paper we extend previous results on the evolution of cooperation to a system of a finite, fixed population of mobile competitors. We conduct a stochastic simulation study examining how parameters controlling an individual's mobility and neighborhood size influences the likelihood that cooperation can evolve within the population. We find that both greater mobility and larger neighborhood size inhibit the evolution of cooperation because it allows the free-riders to find the cooperators faster and exploit them more.

Analysis of Multi-Sequence Time Series Data from MS Lesions on Structural MRI

Elizabeth M. Sweeney
Johns Hopkins Bloomberg School of Public Health
`emsweene@jhsph.edu`
Coauthors: Russell T. Shinohara, Daniel S. Reich, Ciprian M. Crainiceanu, Ani Eloyan

Structural magnetic resonance imaging (MRI) can be used to detect lesions in the brains of multiple sclerosis (MS) patients. The formation of these lesions is a complex sequence of inflammation, degeneration, and repair that MRI has been shown to be sensitive to. We characterize the lesion formation process with multi-sequence structural MRI. We have longitudinal MRI from 10 MS patients, each with between 10 and 40 studies. Each study consists of multiple sequences: a T1-weighted, T2-weighted, fluid attenuated inversion recovery (FLAIR) and proton density (PD) volume. To identify voxels in the brain that are part of new and enlarging lesions, we use the algorithm Subtraction-Based Logistic Inference for Modeling and Estimation (SuBLIME). In the voxels that are identified by SuBLIME, we extract the multi-sequence time series of the voxel intensities from the four volumes. We also collect the same information from a control area, the voxels from the contralateral of the brain. From these times series we are able to develop and classify categories of MS lesion formation.

Toxoplasmosis Vaccination Strategies

David Sykes
The University of North Carolina at Greensboro
`dgsykes@uncg.edu`
Coauthors: Jan Rychtar

The protozoan Toxoplasma Gondii is a parasite often found in wild and domestic cats, and it is the cause of the disease Toxoplasmosis. More than 60 million people in the United States carry the parasite, and the Centers for Disease Control have placed toxoplasmosis in their disease classification group Neglected Parasitic Infections in which there are five parasitic diseases targeted as priorities for public health action. In recent years, there has been significant progress toward the development of a practical vaccine, so vaccination programs may soon be a viable approach to controlling the disease. Anticipating the availability of a toxoplasmosis vaccine, we are interested in determining when cat owners should vaccinate their own pets. To investigate this, we have created a mathematical model describing the conditions under which vaccination is advantageous for a person living with cats. Analysis of our model shows that under any fixed set of parameters the population vaccination level will stabilize around a value that can be directly computed. We find that populations may achieve herd immunity if the cost of vaccine is zero, there is a critical cost threshold above which no one will use the vaccine, and a vaccine cost slightly below this critical threshold results in high usage of the vaccine, conferring a significant reduction in population seroprevalence.

Chronological Bias Caused by Unobserved Time Trends in Randomized Clinical Trials

Miriam Tamm
RWTH Aachen University
`mtamm@ukaachen.de`
Coauthors: Ralf-Dieter Hilgers (RWTH Aachen University)

In clinical trials randomization is used to achieve comparable treatment groups in order to minimize bias in treatment comparison. Since in many clinical trials patients are recruited sequentially over time, one source of bias are time trends. Time trends occur e.g. because of (unobserved) changes in the patients characteristics, improvements in the concomitant therapy or for various other possible reasons. Even in randomized clinical trials the time trend can bias the treatment comparison if, by chance, a large imbalance between the treatment groups occurs during the recruitment. To minimize this so called chronological bias, a suitable randomization procedure has to be chosen. The ICH E9 guideline recommends the randomization

of subjects in blocks. Nevertheless, to avoid predictability very short block lengths should be avoided. The aim of the presentation is to evaluate and compare the amount of chronological bias in randomized clinical trials - with focus on the permuted block randomization and the choice of different block sizes. We use the following setting: We consider a parallel group design where the response to treatment is measured by a continuous normally distributed random variable. To model different shapes of time trend, we consider linear, logarithmic and stepwise trends. The impact of time trends on the treatment effect estimation and on the test decision is investigated by considering the expected value and mean squared error of the treatment effect estimator and the empirical type I error rate when using the two-tailed t-test. Results regarding the overall evaluation of the randomization procedure as well as results with respect to the maximum extent of bias for a realized sequence in the analysis stage of a clinical trial are presented. This research is part of the IDeAl EU-FP7 project, Grant-Agreement No. 602 552.

Intrinsic-Objective Bayes Variable Selection for Site-Occupancy Models

Daniel Taylor-Rodriguez
SAMSI/Duke University
`dantaylor60@hotmail.com`
Coauthors: Nikolay Bliznyuk, Andrew Womack

Colombia plays a central role in the success of the Jaguar Corridor Initiative (JCI). The corridors, that are to be established across the country, aim to ensure an active pathway that preserves the link between the jaguar populations in South and Central America. A preliminary but crucial step in defining the corridor in Colombian territory is to produce distribution maps for jaguars and their prey. Occupancy models are well suited for this purpose. They use repeated survey detection / non-detection data to estimate the percentage of area occupied after adjusting for imperfect detection. However, Bayesian variable selection methods using suitable testing priors, both on the parameter and model spaces, are yet to be put forth. By setting up the occupancy model as an extension of Albert and Chib's algorithm for probit regression, we build an Objective Bayes variable selection procedure based on intrinsic priors, using prior distributions on the model space that automatically correct for the multiplicity of the testing problem. We conduct several simulation experiments to gauge the performance of our methods and, us the procedure to identify the drivers of jaguar population occupancy in Colombia.

Game-Theoretical Model of Oviposition Site-Selection of Gravid Female Mosquitoes

Ana Tognasoli
University of North Carolina at Greensboro
`a_tognas@uncg.edu`
Coauthors: Jan Rychtar and Gideon Wasserberg, University of North Carolina at Greensboro.

We aim to produce a game-theoretical model for the oviposition site-selection behavior of gravid female mosquitoes, which would allow us to further infer on the validity of the Hump-shaped model. Hematophagous mosquitoes can serve as vectors for many diseases, including but not limited to Chickengunya, Malaria and Dengue. Fitness of the individual is dependent upon the optimal oviposition site-selection by the female. Thus the understanding of reproduction and behavioral patterns related to their oviposition site-selection is essential to the creation of effective methods of population control of those species. A Hump shaped model for oviposition site-selection has been recently suggested, in which different densities of conspecifics, enlist different site-selection responses from gravid females. A trade-off occurs between the cost and benefits related to the presence and density of conspecifics. It is assumed that every female will inherently behave optimally due to forces of natural selection, and the presence of conspecifics in immature stages is indicative of the suitability of the site, since other offspring has persisted on that site. While enough nutrients, proper a-biotic conditions and absence of predators are viewed as benefits, the larval competition will emerge as a cost, hence the trade-off. An optimal density of conspecifics, will provide the best site for oviposition, while sites with lower density of conspecifics are less attractive because of the lack of reassurance, and sites with

larger density are less attractive due to an increasing competition for resources. Within the same game-theoretical scheme, we may evaluate the importance of other aspects of site selection, such as predation and resource enrichment.

Implementing Personalized Medicine: Estimation of Optimal Dynamic Treatment Regimes. Part II: Value Search Estimators and Classification Perspective

Anastasios A. Tsiatis
North Carolina State University
`tsiatis@ncsu.edu`

A dynamic treatment regime is a list of sequential decision rules that formalizes clinical practice. Each rule corresponds to a key treatment decision point in the course of patient's disease or disorder at which clinicians would select a treatment for the patient from among the available options based on information on the patient to that point. Given data from a clinical trial or observational study, the statistical problem is to estimate the optimal dynamic treatment regime, that leading to the most favorable outcome if followed by all patients in the population. Regression-based methods have the drawback that they can yield estimators for an optimal regime that are far from a true optimal regime if the regression models involved are misspecified. Even if the models are correctly specified, complex models can lead to regimes that are infeasible for practical implementation. An alternative approach to estimating an optimal dynamic treatment regime that addresses these issues is based on restricting attention to a practically feasible class of regimes and then maximizing a suitable estimator for the "value" of a regime, the expected outcome if all patients in the population were to follow it. We focus first on describing these so-called value search estimators in the case of a single treatment decision. We then discuss how this problem can be recast from a classification perspective. Extension to multiple decision points is then discussed. This talk is based on joint work with Baqun Zhang, Phillip Schulte, Eric B. Laber, and Marie Davidian.

A Revisit to Testing the Equality of Several Poisson Parameters

Suntaree Unhapipat
University of Louisiana at Lafayette
`u.suntaree@gmail.com`
Coauthors: Montip Tiensuwan, Mahidol University (Thailand); Nabendu Pal, University of Louisiana at Lafayette (USA).

Recently Chang et al (2010, Communication in Statistics-Simulation and Computation) considered testing the equality of several Poisson parameters, and proposed a new parametric bootstrap (PB) method, called CAT. The CAT was compared against fourteen other tests including the asymptotic likelihood ratio test (ALRT) as well as the parametric bootstrap version of the likelihood ratio test (henceforth, PBLRT), and all have been found to be quite conservative unless the common parameter under the null hypothesis is not too small. In this paper we propose five new test procedures based on two broad adjustments, namely (i) using different metrics which measure deviation from the null hypothesis; and (ii) using shrinkage estimators in the aforementioned metrics. All these five new tests, are PB in nature, and their critical values are obtained internally through computational steps (i.e., one does not need to provide the critical values explicitly for these five tests). The resultant tests are then studied through simulation and compared against ALRT as well as PBLRT in terms of size and power. In this study we also investigate the role of sample sizes in determining the size and/or power of the tests. Finally, we apply our test methods to U.S. homicide data (FBI data) for comparing different states.

Training Statistical Consultants to Become Collaborative Statisticians

Eric A. Vance
Virginia Tech
`ervance@vt.edu`

At LISA, Virginia Tech's Laboratory for Interdisciplinary Statistical Analysis, we train our students to become effective statistical collaborators who focus on helping researchers answer their research questions rather than merely answering their statistics questions. This change in mindset can be difficult for some, but usually results in a successful collaboration that leaves the researcher happy and satisfied with LISA and more appreciative of the contribution that statistics and statisticians can make to his or her research. Over 95% of clients reporting about their collaborations with LISA statisticians report that the statisticians were helpful and that they are satisfied with LISA. This talk will highlight the key features of LISAs training program, including our use of video to record data on our statistical practice and how we analyze that data to improve our statistical collaboration skills. These training methods form the basis for the LISA 2020 program to train statisticians from developing countries to become collaborative statisticians and help them create a network of statistical collaboration laboratories to enable and accelerate research; help government officials, NGOs, and local businesses make better decisions using data; and train the next generation of collaborative statisticians.

Bayesian Reliability Estimation of Inverted Exponential Distribution under Progressive Type-II Censored Data

Pradeep Kumar Vishwakarma
Banaras Hindu University, India
`vpradeep4u@gmail.com`
Coauthors: Sanjay Kumar Singh, Umesh Singh and Abhimanyu Singh Yadav

The present study deals with the estimation procedure for the parameter, reliability function and hazard function of the inverted exponential distribution under progressive Type-II censored data. For estimation purpose, we have considered both classical and Bayesian method of estimation. The Bayes estimates of the parameter, reliability function and hazard function are calculated under symmetric and asymmetric loss function and also computed the 95% highest posterior density (HPD) intervals of the parameter. Further, Monte Carlo simulation technique has been used to compare the performances of the Bayes estimators with corresponding MLEs under both the loss functions. Finally, we analysed one data set for illustrative purposes of the study.

SMART Estimators for SMART Studies

Abdus S Wahed
University of Pittsburgh
`wahed@pitt.edu`
Coauthors: Semhar Ogbagaber

Hypothesis testing to compare adaptive treatment strategies (ATS) from a sequential multiple assignment randomization trial (SMART) are usually based on inverse probability weighting or g-estimation. However, regression methods that allow for comparison of treatment strategies that flexibly adjusts for baseline covariates using these methods are not as straight-forward due to the fact that one patient can belong to multiple strategies. This poses a challenge for data analysts as it violates basic assumptions of regression modeling of unique group membership. In this talk, we will propose an artificial randomization" technique to make the data appear that each subject belongs to a single ATS. This enables treatment strategy indicators to be inserted as covariates in a regression model and apply standard tests such as t- or F-tests. The properties of this method are investigated analytically and through simulation. We demonstrate the application of this method by applying to a SMART study of cancer regimens.

Optimal Algorithms for Logistic Regression with Big Data

HaiYing Wang
University of New Hampshire
`haiying.wang@unh.edu`
Coauthors: Rong zhu, Ping Ma

For non-linear models with Big Data, it is computationally infeasible to obtain the maximum likelihood estimates for unknown parameters. This paper proposes algorithms to approximate the maximum likelihood estimates for logistic regression based on random sub-sampling. We prove the consistency of the sub-sampling algorithms and the asymptotic normality of resultant estimators. Then we develop two optimal strategies by minimizing functions of the variance-covariance matrix of the asymptotic normal distribution. The proposed methods are evaluated using synthetic and real data sets.

Releasing Synthetic Magnitude Microdata with Fixed Marginal

Lan Wei
Duke University
`lw178@stat.duke.edu`
Coauthors: Jerome P. Reiter

Establishment microdata is subjected to strict confidential protection. Releasing it while providing meaningful statistical inference is a challenge. This is especially true for micro discrete data with fixed marginal totals. We develop a new statistical disclosure limitation (SDL) technique based on data synthesis. We propose a nonparametric Bayesian mixture of Poisson distributions for multivariate count data and synthesis methods subject to fixed marginal totals. We apply the procedure to data from the Colombian Annual Manufacturing Survey. We also present methods for disclosure risk assessment.

Sparsity in Multivariate Dynamic Models

Mike West
Duke University
`mw@stat.duke.edu`

Bayesian model developments bringing a range of approaches to sparsity modelling in multivariate time series have recently been demonstrably valuable in a number of topical application areas. Practical deployment of models and coupled, relevant decision analytic strategies, have defined advances in evaluating time-varying inter-relationships among series, dynamic dependency networks (with time-evolving predictor-response interconnections and lagged structures), and applications in prediction and decisions across a range of financial, economic and scientific areas.

My talk will link to some of the history of Bayesian sparsity modelling in multivariate time series analysis and forecasting, and then discuss recent modelling innovations for dynamic sparsity modelling, and scaling of Bayesian dynamic model-based analysis, inference and decisions. Applications in dynamic network modelling, finance and econometrics will feature centrally for context and motivation. These applications will then also help to convey experiences we have had in using these modelling approaches to aid in isolating and estimating sparse, dynamic structures in complex multivariate, time-evolving systems, and the impact they can have on improving predictions and decision making as a result.

Multiple-objective Optimal Designs for the 4-parameter Logistic Model

Weng Kee Wong
UCLA
`wkwong@ucla.edu`
Coauthors: Guanghao Qi, Fudan University Seung Wong Hyun, North Dakota State University

Multiple-objective optimal designs have several advantages over traditional single-objective optimal designs. They can properly incorporate several study objectives at the design stage and deliver user-specified efficiencies according to their relative importance.

I demonstrate different ways of finding multiple-objective optimal designs using both meta-heuristic and non-metaheuristic methods. To fix ideas, I use the flexible 4-parameter logistic model as a dose response model and use different algorithms to generate 3-objective optimal designs. The objectives of interest are (i) estimate all parameters in the model,(2) estimate the IC50 and (3) estimate a user-specified minimal effective dose. The constructed designs allow the possibility that each objective has unequal interest to the researcher and they can all be verified to be globally optimum.

High Dimensional Sparse MANOVA

Yin Xia
The University of North Carolina at Chapel Hill
`xiayin@email.unc.edu`
Coauthors: Tony Cai

This talk considers testing the equality of multiple high-dimensional mean vectors under dependency. We propose a test that is based on a linear transformation of the data by the precision matrix which incorporates the dependence structure of the variables. The limiting null distribution of the test statistic is derived and is shown to be the extreme value distribution of type I. The convergence to the limiting distribution is, however, slow when the number of groups is relatively large. An intermediate correction factor is introduced which significantly improves the accuracy of the test. It is shown that the test is particularly powerful against sparse alternatives and enjoys certain optimality. A simulation study is carried out to examine the numerical performance of the test and compare with other tests given in the literature. The numerical results show that the proposed test significantly outperforms those tests against sparse alternatives.

High Dimensional Tests for Functional Brain Networks

Jichun Xie
Duke University
`jichun.xie@duke.edu`
Coauthors: Jian Kang, Emory University

Large-scale resting-state functional magnetic resonance imaging (fMRI) studies have been conducted for patients with autism, and the existence of abnormalities in the functional connec- tivity between brain regions (containing more than one voxel) have been clearly demonstrated. Due to the ultra-high dimensionality of the data, current methods focusing on studying the connectivity pattern between voxels are often lack of power and computation-efficiency. In this paper, we introduce a new framework to identify the connection pattern of gigantic networks with desired resolution. We propose two pairwise testing procedures to test region dependence, and one multiple testing procedure to test global structures of the network. The limiting null distributions of the test statistics are derived, together with their power. It is also shown that the tests are rate optimal when the alternative networks are sparse. The numerical studies show that the tests are valid and powerful. We apply our method to a resting-state fMRI study on autism. The analysis yields interesting insights about the mechanism of autism.

Joint Estimation of Multiple Dependent Gaussian Graphical Models with Applications to Mouse Genetics

Yuying Xie
UNC at Chapel Hill
`xyy@email.unc.edu`
Coauthors: Yufeng Liu and William Valdar

Gaussian graphical models are widely used to represent conditional dependence among random variables. In this paper we propose a novel estimator for data arising from a group of GGMs that are themselves dependent. A motivating example is that of modelling gene expression collected on multiple tissues from the same individual: a multivariate outcome that is affected by dependencies defined at the level of both the tissue and the whole body. Existing methods that assume independence among graphs are not applicable in this setting. To estimate multiple dependent graphs, we decompose the problem into two graphical layers: the systemic layer, which is the network affecting all outcomes and thereby inducing cross-graph dependency, and the category-specific layer, which represents the graph-specific variation. We propose a new graphical EM technique that estimates the two layers jointly, establish the consistency and selection sparsistency of the proposed estimator, and confirm by simulation that the EM method is superior to a simple one-step method. Lastly, we apply our graphical EM technique to mouse genetic data and obtain biologically plausible results.

Evolution of Mobility in Predator-Prey Systems

Fei Xu
Wilfrid Laurier University, Canada
`fxu.feixu@gmail.com`
Coauthors: Ross Cressman and Vlastimil Krivan

In this talk, we investigate the dynamics of a predator-prey system with the assumption that both prey and predators use game theory-based strategies to maximize their per capita population growth rates. The predators adjust their strategies in order to catch more prey per unit time, while the prey, on the other hand, adjust their reactions to minimize the chances of being caught. We assume each individual is either mobile or sessile and investigate the evolution of mobility for each species in the predator-prey system. When the underlying population dynamics is of the Lotka-Volterra type, we show that strategies evolve to the equilibrium predicted by evolutionary game theory and that population sizes approach their corresponding stable equilibrium (i.e. strategy and population effects can be analyzed separately). This is no longer the case when population dynamics is based on the Holling II functional response, although the strategic analysis still provides a valuable intuition into the long term outcome. Numerical simulation results indicate that, for some parameter values, the system has chaotic behavior. Our investigation reveals the relationship between the game theory-based reactions of prey and predators, and their population changes.

Optimal Design of Clinical Trials with Randomly Censored “Time-To-Event End Points

Xiao Qiang Xue
Quintiles
`valerii.fedorov@quintiles.com`
Coauthors: Valerii Fedorov and Xiao Qiang Xue Predictive Analytics, Quintiles, USA

In the typical clinical trials the subject arrival times can be modeled as the outcomes of a point process. Subsequently the follow-up times can be viewed as random variables and they become known only after trial completion. Thus the information that is gained depends on the sample size (number of subjects) and random follow up times (specific for each subject). We present results for a few popular survival functions that follow from optimal design theory and illuminate them with numerical examples.

Two-Block Non-Negative Matrix Factorization for Unknown Chemical Identification

S. Stanley Young
National Institute of Statistical Sciences
`young@niss.org`
Coauthors: Chris Beecher, IORA Technologies, nextgenmetabolomics.com Paul Fogel, Consultant, Paris, France

Often two blocks of omic data are available that are linked as the same sample is matched in both blocks. It would be useful to links sets of variables in one block with sets of variables in the other block. An example would be genes in one block and proteins in the other. In our case we have known compounds in one block and unknown compounds in the second block. If the known compounds in block one can be linked to the unknowns in block two, then the ability to identify the unknowns is greatly improved. Our idea is to use non-negative matrix factorization on each block, grouping together sets of compounds that are statistically highly related. The score vectors from each factorization are then correlated to link sets from each block. The benefit of this linking is to put the unknowns into a context of the knowns thereby facilitating identification. A compound from the block of unknowns could be in the same biochemical pathway as a known. We demonstrate the process with a metabolomics data set.

Sparse Learning Incorporating Graphical Structure among Predictors

Guan Yu
University of North Carolina at Chapel Hill
`guanyu@live.unc.edu`
Coauthors: Yufeng Liu, University of North Carolina at Chapel Hill.

With the abundance of high dimensional data in various disciplines, sparse regularized techniques are very popular these days. In this paper, we make use of the structure information among predictors to improve sparse regression models. Typically, such structure information can be modeled by the connectivity of an undirected graph using all predictors as nodes of the graph. Most existing methods use this undirected graph information edge-by-edge to encourage the regression coefficients of corresponding connected predictors to be similar. However, such methods may require expensive computation when the predictor graph has many more edges than nodes. Furthermore, they do not directly utilize the neighborhood information of the graph. In this paper, we incorporate the graph information node-by-node, instead of edge-by-edge as used in most existing methods. To that end, we decompose the true p-dimensional regression coefficient vector as the sum of p latent parts and incorporate convex group penalty functions to utilize predictor graph neighborhood structure information. Our proposed method is very general and it includes adaptive Lasso, group Lasso and ridge regression as special cases. Both theoretical and numerical studies demonstrate the effectiveness of the proposed method for simultaneous estimation, prediction and model selection.

Detection of Very Short Signal Segments

Ming Yuan
University of Wisconsin-Madison
`myuan@stat.wisc.edu`

Motivated by a range of applications in engineering and genomics, we consider in this paper detection of very short signal segments in three settings: signals with known shape, arbitrary signals, and smooth signals. Optimal rates of detection are established for the three cases and rate-optimal detectors are constructed. The detectors are easily implementable and are based on scanning with linear and quadratic statistics. Our analysis reveals both similarities and differences in the strategy and fundamental difficulty of detection among these three settings.

Sorting Out Genuine and Spurious Discoveries in Large Scale Genetic Studies

Dmitri Zaykin
National Institute of Environmental Health Sciences
`dmitri.zaykin@nih.gov`

Imagine an utterly uninformed epidemiologist who is studying effects of various predictors on susceptibility to a disease. The epidemiologist is oblivious to any external knowledge regarding possible effects of predictors on the outcome and simply tests every predictor in sight. In this scenario, predictors that do in fact influence the outcome, i.e., ”true signals” occur at a constant rate. In other words, the rate of true signal occurrence does not diminish as additional predictors are tested. At the end of the day, the epidemiologist reports the predictor yielding the smallest P-value as a potentially true signal. This strategy is often perceived with disdain as ”data torturing”. However, a predictor with the smallest P-value in such a study becomes increasingly less likely to be a spurious association as more tests are performed. More generally, a set of ”top hits” in a multiple-testing experiment becomes increasingly enriched with true signals as more tests are carried out. To understand and quantify this phenomenon, we develop theory for evaluating rates of spurious findings among predictors with the smallest P-values. Our results can be used for evaluating the proportion of real discoveries in modern genetic association studies, where millions of variants are tested in an agnostic, hypothesis-generating manner.

A Two-Stage, Phase II Clinical Trial Design with Nested Criteria for Early Stopping and Efficacy

Daniel Zelterman
Yale University
`Daniel.Zelterman@yale.edu`

A Phase II clinical trial is a hypothesis test on the probability of success in a binomial distribution, with a historical control value as the null hypothesis. We propose a two-stage design for a clinical trial with an early stopping rule for safety. This design employs different criteria to assess early stopping and efficacy. The early stopping rule is based on criteria that can be determined more quickly than efficacy. These separate criteria are also nested in the sense that efficacy is a special case of, but not identical to, the early stopping criteria. The design readily allows for planning in terms of statistical significance, power, and expected sample size necessary to assess an early stopping rule. This method is illustrated with a Phase II design comparing patients treated for lung cancer with a novel drug combination to a historical control. In this example, the early stopping rule is based on the number of patients who exhibit progression-free survival (PFS) at 2 months post treatment follow-up. Efficacy is judged by the number of patients who have PFS at 6 months.

Estimating Biomarker-Adjusted Treatment Effects in Randomized Clinical Trials Using Auxiliary Covariates: Application to a Phase 3 HIV-1 Trial

Bo Zhang
FDA/CDRH
`Bo.Zhang@fda.hhs.gov`
Coauthors: Zhiwei Zhang,Yanping Qu, Lei Nie, Guoxing Soon

A biomarker-adjusted treatment effect (BATE) model describes the effect of one treatment versus another on a subpopulation of patients defined by a biomarker. Such a model can be estimated from clinical trial data without relying on additional modeling assumptions, and the estimator can be made more efficient by incorporating information on the main effect of the biomarker on the outcome of interest. Motivated by an HIV trial known as THRIVE, we consider the use of auxiliary covariates, which are usually available in clinical trials and have been used in overall treatment comparisons, in estimating a BATE model. Such covariates can be incorporated using an existing augmentation technique. For a specific type of estimating

functions for difference-based BATE models, the optimal augmentation depends only on the joint main effects of marker and covariates. For a ratio-based BATE model, this result holds in special cases but not in general; however, simulation results suggest that the augmentation based on the joint main effects of marker and covariates is virtually equivalent to the theoretically optimal augmentation, especially when the augmentation terms are estimated from data. Application of these methods and results to the THRIVE data yields new insights on the utility of baseline CD4 cell count and viral load as predictive or treatment selection markers. The research results were published by Statistical Methods in Medical Research (doi: 10.1177/0962280213515572).

Smoothing Splines Using Compactly Supported, Positive Definite, Radial Basis Functions

Guoyi Zhang
University of New Mexico
`gzhang123@gmail.com`

In this research, we develop a fast algorithm for a smoothing spline estimator in multivariate regression. To accomplish this, we employ general concepts associated with roughness penalty methods in conjunction with the theory of radial basis functions and reproducing kernel Hilbert spaces. It is shown that through the use of compactly supported radial basis functions it becomes possible to recover the band structured matrix feature of univariate spline smoothing and thereby obtain a fast computational algorithm. Given n data points in R^2, the new algorithm has complexity $O(n^2)$ compared to $O(n^3)$, the order for the thin plate multivariate smoothing splines.

Covariance Structures of Axially Symmetric Spatial Processes on the Sphere

Haimeng Zhang
UNCG
`h_zhang5@uncg.edu`
Coauthors: Chunfeng Huang and Scott Robeson

In this presentation, I will discuss a class of axially symmetric spatial processes in the analysis of global-scale data, whereby their covariance function depends on differences in longitude alone. In addition, a simplified representation of a valid axially symmetric process and its covariance function will be presented, and further construction of parametric covariance models for axially symmetric processes will be explored.

Accelerated Intensity Frailty Model for Recurrent Events Data

Jiajia Zhang
University of South Carolina
`jzhang@mailbox.sc.edu`
Coauthors: Bo Liu; Wenbin Lu

We propose an accelerated intensity frailty (AIF) model for recurrent events data and derive a test for the variance of frailty. In addition, we develop a kernel-smoothing-based EM algorithm for estimating regression coefficients and the baseline intensity function. The variance of the resulting estimator for regression parameters is obtained by a numerical differentiation method. Simulation studies are conducted to evaluate the finite sample performance of the proposed estimator under practical settings and demonstrate the efficiency gain over the Gehan rank estimator based on the AFT model for counting process (Lin et al., 1998, *Biometrika* **85,** 605-618). Our method is further illustrated with an application to a bladder tumor recurrence data.

Valid Post-Selection Inference

Kai Zhang
University of North Carolina, Chapel Hill
`zhangk@email.unc.edu`
Coauthors: Richard Berk, Lawrence D. Brown, Andreas Buja, Linda Zhao

It is common practice in statistical data analysis to perform data-driven model selection and derive statistical inference from the selected model. Such inference is generally invalid. We propose to produce valid post-selection inference by reducing the problem to one of simultaneous inference. Simultaneity is required for all linear functions that arise as coefficient estimates in all submodels. By purchasing simultaneity insurance for all possible submodels, the resulting post-selection inference is rendered universally valid under all possible model selection procedures. This inference is therefore generally conservative for particular selection procedures, but it is always less conservative than full Scheffe protection. Importantly it does not depend on the truth of the selected submodel, and hence it produces valid inference even in wrong models. We describe the structure of the simultaneous inference problem and give some asymptotic results.

Interpretable and Parsimonious Optimal Treatment Regime Using Decision List

Yichi Zhang
North Carolina State University
`yzhang52@ncsu.edu`
Coauthors: Eric Laber (North Carolina State University) Anastasios Tsiatis (North Carolina State University) Marie Davidian (North Carolina State University)

A treatment regime is a rule that recommends a treatment for each patient based on his/her characteristics. Due to patient heterogeneity, using a treatment regime could lead to better clinical outcome than giving the same treatment to every patient. We employ the concept of decision list, a finite sequence of if-then statements, to obtain a treatment regime that is readily interpretable and can potentially provide insights into the therapy. Moreover, by requiring the condition in each if-clause be some simple inequality involving at most two variables, the most relevant variables are identified. We propose novel algorithms to learn the optimal treatment regime from a randomized or observational study, by maximizing the augmented inverse probability weighted estimator of the mean clinical outcome. The proposed method is robust against model misspecification, compared to regression-based methods like Q-learning. Simulation studies show that the proposed method performs favorably to competing methods. In addition, the proposed method is applied to two clinical trials, one on breast cancer and the other on chronic depression.

New Analytical Methods for Non-Inferiority Trials: Covariate Adjustment and Sensitivity Analysis

Zhiwei Zhang
FDA
`zhiwei.zhang@fda.hhs.gov`
Coauthors: Lei Nie, Guoxing Soon, Bo Zhang

Even though an active-controlled trial provides no information about placebo, investigators and regulators often wonder how the experimental treatment would compare to placebo should a placebo arm be included in the study. Such an indirect comparison often requires a constancy assumption, namely that the control effect relative to placebo is constant across studies. When the constancy assumption is in doubt, there are ad hoc methods that discount" the historical data in conservative ways. Recently, a covariate adjustment approach was proposed that does not require constancy or involve discounting, but rather attempts to adjust for any imbalances in covariates between the current and historical studies. This covariate-adjusted approach is valid under a conditional constancy assumption which requires only that the control effect be constant within each subpopulation characterized by the observed covariates. Furthermore, a sensitivity analysis

approach has been developed to address possible departures from the conditional constancy assumption due to imbalances in unmeasured covariates. This presentation describes these new approaches and illustrates them with examples.

Universally Optimal Designs for Two Interference Models

Wei Zheng
Indiana University-Purdue University Indianapolis
`weizheng@iupui.edu`

A systematic study is carried out regarding universally optimal designs under the interference model, previously investigated by Kunert and Martin (2000) and Kunert and Mersmann (2011). Parallel results are also provided for the undirectional interference model, where the left and right neighbor effects are equal. It is further shown that the efficiency of any design under the latter model is at least its efficiency under the former model. Designs universally optimal for both models are also identified. Most importantly, Kushner's type linear equations system is developed as a necessary and sufficient condition for a design to be universally optimal. This result is novel for models with at least two sets of treatment-related nuisance parameters, which are left and right neighbor effects here. It sheds light on other models in deriving asymmetric optimal or efficient designs.

Extremal Risks: Paths and Indices of Maximal Tail Dependence

Ricardas Zitikis
University of Western Ontario
`zitikis@stats.uwo.ca`
Coauthors: Edward Furman and Jianxi Su (York University, Toronto)

Estimating extremal risks has been a particularly important problem for actuaries and risk analysts. When portfolios contain multiple business lines, extreme risks become even more troublesome and their co-movements play a key role in risk management. When dependent risks are modeled using copulas, the co-movements are usually modeled by their upper tails in the case of insurance risks and by the lower tails in the case of financial risks. In this talk, we shall argue that the classical indices of tail dependence quite often underestimate the tail dependence in copulas and thus may not always convey useful information. We shall illustrate this phenomenon using a number of bivariate copulas and suggest an alternative way for assessing tail dependence.

Trends in Disguise

Ricardas Zitikis
University of Western Ontario
`zitikis@stats.uwo.ca`
Coauthors: Vytaras Brazauskas (University of Wisconsin-Milwaukee) and Bruce L. Jones (University of Western Ontario)

Human longevity is changing; but at what rate? Insurance claims are increasing; but at what rate? Are the trends that we glean from data true or illusionary? The shocking fact is that true trends might be quite different from those that we actually see from visualized data. Indeed, in some situations the upward trends (e.g., inflation) may even look decreasing (e.g., deflation). In this talk, we shall elucidate this trends in disguise phenomenon, whose further mathematical and statistical details will soon be published by the Annals of Actuarial Science in a joint paper with V. Brazauskas and B.L. Jones.

Contents