IDWSDS 2024

October 8, 2024

International Day of Women in Statistics and Data Science

Empowering the next generation of statisticians and data scientists

IDWSDS 2024

2024 INTERNATIONAL DAY OF WOMEN IN STATISTICS AND DATA SCIENCE

CONFERENCE PROGRAM AND ABSTRACTS

OCTOBER 8TH, 2024
A VIRTUAL 24 HOUR CONFERENCE
HTTPS://IDWSDS.ORG

ORGANIZED BY

CAUCUS FOR WOMEN IN STATISTICS AND DATA SCIENCE

HOSTED BY

CAUCUS FOR WOMEN IN STATISTICS AND DATA SCIENCE VIA ZOOM

IDWSDS 2024

TABLE OF CONTENTS

IDWSDS 2024

IDWSDS 2024

IDWSDS 2024

IDWSDS 2024

Welcome

Welcome to the third annual International Day of Women in Statistics and Data Science conference! This year's theme is Empowering the Next Generation of Statisticians and Data Scientists. We are thrilled to celebrate the achievements and experiences of women around the world in these vital fields. By sharing our career journeys, research interests, challenges, lessons learned, and passions, we are strengthening our community and paving the way for the next generation of leaders and innovators in statistics and data science.

As you explore the program, please pay close attention to the timing of each session and ensure you convert UTC time to your local time. Each session includes an abstract, along with photos and bios of the presenters. We are excited to offer over 60 sessions covering technical topics in statistics and data science, career insights, and the history of our field, featuring speakers from across the globe. The majority of the conference will take place in three different Zoom rooms, so be sure to check the link for each session to access the correct room.

We are delighted to see our community grow each year and look forward to continuing this momentum in the future. As you review the program and engage in the conference, consider how you or other statisticians in your area might collaborate for a future presentation.

Jessica Kohlschmidt and Dong-Yun Kim
Co-chairs, 2024 International Day of Women in Statistics and Data Science Organizing Committee

IDWSDS 2024

Organizing Committee

CYTHIA BLAND
RTI International

Cynthia Bland is the President of the Caucus for Women in Statistics and Data Science. Her strategic focus for the Caucus is to increase membership and deepen the connection members have to the organization. Professionally she is a survey statistician and directs the Center for Official Statistics at RTI International, a non-profit research firm specializing in government contracting. In her two decades at RTI, she has designed and analyzed national surveys, with an emphasis on patient experience of care surveys. Cynthia leads a center of more than 80 statisticians as they innovate for clients and grow in their own careers. Cynthia serves on the Executive Board of RTI Press, an open-access, peer-reviewed journal, has been active with the NC Chapter of the American Statistical Association, and is a lecturer in statistics for the Kenan-Flagler School of Business at the University of North Carolina at Chapel Hill. Outside of the statistical world, she serves on the board of a local credit union and enjoys ballet.

PROF. LÍGIA HENRIQUES-RODRIGUES
University of Évora and CIMA

https://www.researchgate.net/profile/Ligia-Rodrigues-2

Lígia Henriques-Rodrigues (LHR) is an Assistant Professor in the Department of Mathematics at the School of Science and Technology, University of Évora. LHR completed a postdoctoral fellowship in Statistics of Extremes at the Faculty of Sciences, University of Lisbon, and earned her Ph.D. in Probability and Statistics from the same university. Her primary research focuses on the statistics of extremes and computational statistics, particularly in developing new classes of semi-parametric reduced-bias estimators for extreme event parameters. The applications of her research span environmental and financial sciences. Beyond her academic pursuits, LHR has been a board member of the Portuguese Statistical Society (SPE) since January 2024 and is a member of the executive committee for the Master's Degree in Statistical Modeling and Data Analysis at the University of Évora.

CO-CHAIR

DR. DONG-YUN KIM
NHLBI/NIH

https://www.kimdongyun.com/

Dr. Kim is a mathematical statistician at the Office of Biostatistics Research within National Heart, Lung, and Blood Institute (NHLBI), Bethesda, Maryland, US and adjunct professor at the department of statistics, George Mason University. She received a PhD in Statistics from the University of Michigan, Ann Arbor in 2003. Before joining NIH in 2013, she held a faculty position at Virginia Tech. Her research interests include fully sequential monitoring in clinical trials, change-point inference, and statistical genetics. Currently she is involved in large NHLBI-sponsored clinical trials and intramural projects in MRI imaging, pulmonary disease and cancer research. Dr. Kim has years of experience in collaborative research in other areas including mobile health, bioengineering, health services, and environmental science. Dr. Kim was the President of the Caucus for Women in Statistics and Data Science (CWS) in 2023. Currently, she is serving as a chair of several committees in ASA and CWS including Michael Woodroofe Award which she established last year to honor the memory of Prof. Woodroofe. She is also a board member for the Korean International Statistics Society (KISS), and the co-Chair of International Day of Women in Statistics and Data Science.

CO-CHAIR

DR. JESSICA KOHLSCHMIDT
The Ohio State University

https://fisher.osu.edu/people/kohlschmidt.1

Dr. Jessica K. Kohlschmidt, a first-generation college student from Houston, Texas, developed a passion for math and teaching early on. She pursued a degree in mathematics with a minor in statistics, eventually earning her Ph.D. from The Ohio State University. Since 1999, Dr. Kohlschmidt has taught in colleges and served as a biostatistician at the Clara D. Bloomfield Center for Leukemia Outcomes Research. She currently teaching in the Fisher College of Business at The Ohio State University.
An active leader in the Caucus for Women in Statistics (CWS), she was its first executive director and now oversees operations, focusing on international outreach and youth engagement. Passionate about encouraging young people, particularly girls, to explore careers in statistics, she contributes to initiatives like the Florence Nightingale Day event. Dr. Kohlschmidt also holds leadership roles in the American Statistical Association (ASA), where she serves as Vice Chair for District 3, membership chair of the Columbus Chapter, and co-chair of the History in Statistics Interest Group. She finds fulfillment in helping students and collaborators understand statistics.

DR. SHILI LIN
The Ohio State University

Shili Lin is a Professor of Statistics at the Ohio State University. Her research interests lie in the development and application of statistical methods to multi-omics data from cell lines, populations, and family samples. Shili is an active member of and serves the statistical profession in various capacities, including editorial service for various journals (e.g. current Associate Editor (AE) of Biometrics and former AE of the Journal of the American Statistical Association), severing as panel members in NIH Study Sections (e.g. former standing member of the Biostatistical Methods and Research Design), and serving professional organizations such as the American Statistical Association (ASA), the Institute of Mathematical Statistics (IMS), the International Biometrics Society (IBS), and the International Statistical Institute (ISI). She also served as the 2018 Caucus for Women in Statistics (CWS) President and is currently serving in the Board of Directors of the Canadian Statistical Sciences Institute (CANSSI). Shili in is a co-founder and the president of the Florence Nightingale Day for Statistics and Data Science. She is a Fellow of the ASA, the IMS, the American Association for the Advancement of Science (AAAS), and an elected member of the ISI.

IDWSDS 2024

Organizing Committee

continued...

DR. ALTEA LORENZO-ARRIBAS

Biomathematics and Statistics Scotland (BioSS)

https://www.bioss.ac.uk/people/altea

Altea is a socio-economic statistician at Biomathematics and Statistics Scotland (BioSS) working in collaboration with researchers at the Scottish Environment, Food and Agriculture Research Institutions. She is an elected council member of the Royal Statistical Society, member of the Society's AI Task Force, chair of the Celebrating Diversity Special Interest Group and the secretary of the History of Statistics Section, as well as a member of the Women Committee of the Spanish Society of Statistics and Operations Research (SEIO), and the Spanish Biostatistics Network (Biostatnet).

SARAH LOTSPEICH

Wake Forest University

Sarah Lotspeich is an Assistant Professor in Statistical Sciences at Wake Forest University, with a secondary appointment in Biostatistics and Data Science at the Wake Forest University School of Medicine. She is enthusiastic about mentoring student research and co-leads the Spatial and Environmental Statistics in Health (SESH) Lab at Wake Forest and the Missing and INcomplete Data (MIND) Lab at the University of North Carolina (UNC) at Chapel Hill. She co-organizes Florence Nightingale Day at Wake Forest annually, engaging local students in statistics and data science, and holds elected positions in the Caucus for Women in Statistics and Data Science and other organizations. Sarah completed a postdoctoral fellowship in Biostatistics at UNC Chapel Hill and earned her Ph.D. in Biostatistics from Vanderbilt University. Her research tackles challenges in analyzing error-prone observational data, focusing on international HIV cohorts, electronic health records, and health disparities. She also develops methods for statistical modeling with censored covariates, applicable to Huntington's disease. When she's not writing code, you can find Sarah cross-stitching, adventuring in new places, or rewatching her favorite TV shows.

DR. VANDA LOURENCO

NOVA University of Lisbon

Vanda M. Lourenço (VML) is an Assistant Professor in the Department of Mathematics at the NOVA School of Science and Technology, part of the NOVA University of Lisbon. She earned her PhD in Statistics and Stochastic Processes from Instituto Superior Técnico at the University of Lisbon. Her primary research focuses on robust, non-parametric, and computational statistics, particularly in genetic and genomic association studies and the prediction of quantitative traits. While her expertise has been predominantly applied to challenges in plant breeding, she is also keen on extending these methodologies to animal and human studies. In addition to her academic work, VML actively contributes to professional and statistical communities. She served as President of the Supervisory Board of the Portuguese Statistical Society (SPE) from 2021 to 2024 and currently represents Portugal in the Caucus for Women in Statistics (CWS), where she is also a member of the Nominations Committee. Additionally, she is an associate editor for the Journal of Data Science, Statistics, and Visualization (JDSSV). Beyond her formal roles, she is involved in various other initiatives within and outside statistical societies.

DR. UMUT ÖZBEK

Eli Lilly and Company

Umut Özbek is a director of oncology statistics at Eli Lilly and Company and a part-time associate professor at the Icahn School of Medicine at Mount Sinai. Over the past sixteen years, Dr. Özbek has applied her expertise as a biostatistician to conduct rigorous and collaborative research in many different areas. She aims to advance the field of cancer research using statistical techniques that are appropriate and current, with an eye toward innovation on the methodological, the applied, and the translational fronts. Importantly, she strives to maximize the impact of her work through communication that is precise yet accessible to non-statisticians. She takes great satisfaction in working with clinical and scientific investigators to improve outcomes for patients. Umut Özbek is president-elect of the Caucus for Women in Statistics and Data Science.

DR. HUNNA WATSON

The Ohio State University

https://www.linkedin.com/in/hunna-watson-52030352/

Dr. Hunna Watson is a Research Associate Professor in the School of Medicine at the University of North Carolina at Chapel Hill, where she serves as a biostatistician in the Center of Excellence for Eating Disorders. She also holds Adjunct Research Fellow appointments in the School of Medicine at The University of Western Australia and in the Discipline of Psychology at Curtin University in Australia. Specializing in the field of eating disorders, Dr. Watson has authored over 120 peer-reviewed scientific articles and book chapters, focusing on critical areas such as diagnostic nosology, prevention and treatment, molecular genetics, and randomized clinical trials. Her research has gained global recognition through lectures, papers, and workshops presented worldwide. Dr Watson serves on the editorial board of the International Journal of Eating Disorders and is a member of the Alumni Group of the National Eating Disorders Collaboration Steering Committee in Australia. Dr Watson completed her PhD in Clinical Psychology at Curtin University and holds Master's degrees in Biostatistics from the University of Melbourne and Clinical Psychology from Curtin University.

IDWSDS 2024

Program Schedule

All sessions will take place within three Zoom Rooms. You can find the designated Zoom Room for sessions listed here and on each session page.

CLICK HERE FOR ZOOM ROOM 1 | CLICK HERE FOR ZOOM ROOM 2 | CLICK HERE FOR ZOOM ROOM 3

Session	Time (UTC)	Title	Zoom Room #
		October 8, 2024	
S0	00:00-00:30	Opening	1
S1	00:30-01:00	Graduate and Undergraduate Students' Research in Statistics and Data Science from United States and South Korea	1
S2	00:30-01:30	Advances in Statistical Inference	2
S3	00:30-01:00	Innovative Software Tools for Data Insight Generation	3
S4	01:00-01:30	Statistics' Hidden History: Places Where Social Justice Issues Changed Our Narrative	3
S5	01:30-02:00	High School Research: Accuracy of Confidence Intervals and AI Ethics in Medicine	1
S6	01:30-02:30	Advancing Healthcare with Data Science and Machine Learning: The Future of Data-Driven Care	2
S7	01:30-02:00	Diversity in Data Science: Empowering Women Through Outreach	3
S8	02:00-03:00	Advances in Statistical Learning	1
S9	03:00-04:00	Advanced Statistical Methods for Complex Data Challenges	1
S10	04:00-05:00	The Significance of Mentoring for NZ Women	1
S11	05:00-06:00	Monash EBS PhD Contest	1
K1	**06:00-07:00**	**Modern Bayesian Statistical Science - Challenges and Opportunities**	**1**
S12	07:00-08:00	Research and Other Stories of Korean Women Leaders in Data Sciences	1
S13	07:00-08:00	Research from Women Statisticians in China	2
S14	07:00-08:00	Serbian Contribution to Advanced Statistical Methods for Data Analysis: Privacy, Geometry, and Testing	3
S15	08:00-08:30	The International Biometric Society: A Journey Through History and its Impact on Statistical Science	1
S16	08:00-09:00	New developments in dependent censoring with unknown association	2
S17	08:00-09:00	Causal Inference	3
S18	08:30-09:00	Next Generation: Showcasing Young Portuguese Talent in Biometry and Data Science	1
S19	09:00-10:00	Analyzing complex data in biostatistics and public health	1
S20	09:00-10:00	Measuring stylized facts to bridge gaps, inspire innovation and shape the future	2
S21	09:00-09:30	Round Table: Type 2 Diabetes Mellitus (T2DM) Through the Gender Lens: A Case Study of Ghanaian Community	3
S22A	09:30-09:45	Test for Symmetry and Confidence Interval of the Parameter μ of Skew-Symmetric Laplace Uniform Distribution	3
S22B	09:45-10:00	Evaluating Randomness Assumption: A Novel Graph Theoretic Approach for Linear and Circular Data	3
S23	10:00-11:00	Women in Statistics and Data Science: Overcoming Career Hurdles and Leveraging context	1
S24	10:00-11:00	Methodologies in Time Series and Spatial Statistics With Applications	2
S25	10:00-10:30	Data Science for Health Equity	3
S26	10:30-11:00	Modeling the Impact: Non-Pharmaceutical Interventions (NPIs) and COVID-19 Transmission	3
S27	11:00-11:30	Predicting Recurrent Events in a Survival Framework: Development of a Machine Learning Approach and an Application in Oncology	1
S28	11:00-12:00	SEIO Women in Statistics and Data Science: A Research Sample From Different Perspectives and Career Stages	2
S29	11:00-11:30	Working as a Statistician in the Healthcare Industry	3
S30	11:30-12:00	Using Functional Data Analysis for Surrogate Model Development	1

IDWSDS 2024

Program Schedule

continued...

CLICK HERE FOR ZOOM ROOM 1 | CLICK HERE FOR ZOOM ROOM 2 | CLICK HERE FOR ZOOM ROOM 3

S31	11:30-12:00	Use of Artificial Intelligence and Statistics in the World of Mental Health	3
S32	12:00-13:00	Next Generation: Showcasing Young Portuguese Talent in Statistics and Data Science	1
S33	12:00-13:00	Statistics and data science with machine learning and AI	2
S34	12:00-13:00	Showcasing Research by PhD Students from MRC Biostatistics Unit (Efficient Study Design)	3
K2	**13:00-14:00**	**Leading for a Long, Long Time**	**1**
S35	14:00-15:30	Pioneering Women in Statistics and Data Science: Bridging Global Research and Innovation	1
S36	14:00-15:00	Integration of Mobile and Wearable Data to Study the Interrelationships Between Biological Processes and Mood in Real Time	2
S37	14:00-15:00	The Role of Women in Pharma Analytics: End-to-End Analytics for Dynamic Targeting	3
S38	15:00-16:00	The Art of the Invite: Crafting Successful Invited Session Proposals	2
S39	15:00-16:00	Navigating the Noise: Statistical Methods for Measurement Error in Data	3
S40	15:30-16:30	Perspectives on Local and Global Health: Showcasing Work by Women and Non-Binary People at the London School of Hygiene & Tropical Medicine's New Data Science and Statistics Centre	1
S41	16:00-17:00	Showcasing Research by Postodoctoral researchers from MRC Biostatiscts Unit (Efficient Study Design)	2
S42	16:00-17:00	From Animal Health & Welfare to Plant & Crop Science: Contributions from Statisticians and Modellers at BioSS	3
S43	16:30-17:00	Chart-Toppers and Cliffhangers: Stats in Pop Culture	1
S44	17:00-17:30	Causal Inference Methods for Evaluating Surrogate Markers	1
S45	17:00-18:00	Recent advances in Bayesian methods for covariance estimation and network data	2
S46	17:00-18:00	Innovations in Precision Medicine and Optimal Treatment Strategies	3
S47	17:30-18:00	Statistical Analysis Plans: An Overview and Practical Guidance for Enhancing Research Rigor	1
K3	**18:00-19:00**	**When Prediction Meets Causal Inference**	**1**
S48	19:00-20:00	Infection Insights: Women's Contributions to Infectious Disease Modeling	1
S49	19:00-19:30	Bridging the Gender Gap in Statistics Education: Strategies to Encourage More Women to Pursue Studies in Statistics and Data Science	2
S50	19:00-20:00	Empowered Voices: Brazilian Award Winning Women Share Their Research	3
S51	19:30-20:00	The Stories Behind the History of Women in Statistics	2
S52	20:00-21:00	Resilience, Innovation, and Impact: Women's Journeys in Statistics and Data Science	1
S53	20:00-21:00	Bayesian vs. Frequentist Approaches in Group Sequential Clinical Trial Design	2
S54	20:00-21:00	Statistics Helping the Growth of Business	3
S55	21:00-22:00	Navigating Scientific Challenges and Personal Milestones: Stories from Woodroofe Awardees, leading statisticians and data scientists	1
S56	21:00-22:00	Innovative Health Data Science Applications	2
S57	21:00-22:00	Methods for Diverse Types of Outcomes, Mediators, and Confounders in Causal Mediation Analysis	3
S58	22:00-23:00	Legacies of Women in Data Science and Statistics	1
S59	22:00-23:00	HER Impact on EHR: Women Innovators in Electronic Health Records Research	2
S60	22:00-23:00	The Thriving Neurodivergent Statistician and Data Scientist: Success and Failures (Learning Opportunities) Within the Field	3
K4	**23:00-24:00**	**Shaping the Future: Empowering Women in Statistics and Data Science**	**1**
October 9, 2024			
S61	00:00-00:30	Closing	1

IDWSDS 2024

Time Conversions

IDWSDS 2024 is a 24 hour conference being held Tuesday, October 8th. All session times are given in Universal Time Coordinated (UTC) format. Please use the conversion table below or visit a site like https://www.utctime.net/ to get the current UTC time and time in your area.

Country	Local Time Zone	Conversion
Argentina	ART	UTC-3
Australia	AEST	UTC+10
Bangladesh	BDT	UTC+6
Belgium	CEST	UTC+2
Brazil	Multiple*	UTC-2
Canada	Multiple*	UTC-4
China	CST	UTC+8
Costa Rica	CST	UTC-6
Croatia	CEST	UTC+2
Denmark	CEST	UTC+2
France	CEST	UTC+2
India	IST	UTC+5:30
Indonesia	Multiple*	UTC+7
Italy	CEST	UTC+2
Japan	JST	UTC+9
Korea	KST	UTC+9
Nigeria	WAT	UTC+1
Pakistan	PKT	UTC+5
Poland	CEST	UTC+2
Portugal	WET	UTC+1
Serbia	CEST	UTC+2
South Africa	CAT	UTC+2
Spain	CEST	UTC+2
Taiwan	CST	UTC+8
UK	BST	UTC+1
USA	Multiple*	UTC-4
Zimbabwe	CAT	UTC+2

IDWSDS 2024

Keynote Sessions

IDWSDS 2024

Session Info

K1

KEYNOTE SESSION

October 8th, 06:00 - 07:00 UTC

Modern Bayesian Statistical Science - Challenges and Opportunities

CLICK HERE TO ACCESS THE ZOOM ROOM

Bayesian methods are now pervasive in statistical modelling and analysis. Interestingly, although this approach has its origins in the late 18th century, some of the key theoretical, methodological and practical challenges faced then are still being addressed today. These include formulation of the prior distribution, characterisation of the data through the likelihood and computational algorithms that enable implementation of the approach.

In this presentation, I will review the state of Bayesian statistical science in our time. I will touch on the above issues, discuss their role in data-informed decision making, learning and AI, and describe some of the ways in which we have been implementing Bayesian approaches to address substantive environmental, health and societal challenges.

This research is joint with a range of collaborators who will be acknowledged in the presentation.

Dr. Kerrie Mengersen

QUEENSLAND UNIVERSITY OF TECHNOLOGY

Kerrie Mengersen is a Distinguished Professor of Statistics and Director of the Centre for Data Science at the Queensland University of Technology in Brisbane, Queensland, Australia. Her primary research interests are in Bayesian statistics and its applications in environment, health and society. Kerrie is an elected member of the Australian Academy of Science and the Academy of the Social Sciences of Australia, and a current Vice-President of the International Society of Australia.

ORGANIZER: Altea Lorenzo-Arribas, Biomathematics and Statistics Scotland

CHAIR: Altea Lorenzo-Arribas, Biomathematics and Statistics Scotland

IDWSDS 2024

Session Info

K2

KEYNOTE SESSION

October 8th, 13:00 - 14:00 UTC

Leading for a Long, Long Time

CLICK HERE TO ACCESS THE ZOOM ROOM

Nancy Geller has been the Director of the Office of Biostatistics Research at the National Heart, Lung, and Blood Institute of the National Institutes of Health since 1990. She simply loves her job and doesn't want to stop! She will describe her philosophy of leadership. This includes relatively few rigid rules, respect for colleagues, never forcing a colleague to do a project that needs a statistician, encouraging open expression of differences of opinion, good communication skills, and having a sense of humor. The goal-oriented but relaxed atmosphere of OBR has successfully retained staff with little turnover.

Dr. Nancy Geller

NATIONAL HEART, LUNG AND BLOOD INSTITUTE (NHLBI)

Nancy L. Geller has been the Director of the Office of Biostatistics Research (OBR) at the National Heart, Lung and Blood Institute (NHLBI) of the National Institutes of Health for over 30 years. She leads a group of 13 statisticians who collaborate in the design, implementation, monitoring and analysis of clinical studies in heart, lung and blood diseases and sleep disorders and administers the statistical activities of NHLBI. Despite technically being a government administrator, she has published over 250 papers in the statistical and medical literature collaborating with both statistical and medical colleagues. As such, she is a role model for OBR.

She has been active in the statistics profession, most notably as 2011 President of the American Statistical Association (ASA). She is the 2009 winner of the Janet L. Norwood Award For Outstanding Achievement By A Woman In The Statistical Sciences and the 2024 winner of the ASA Jeanne E. Griffith Mentoring Award.

Throughout her career she has taken on the incidental responsibility of mentoring colleagues and other statisticians who have passed through her life.

ORGANIZER: Dong-Yun Kim, NHLBI/NIH

CHAIR: Nairanjana "Jan" Dasgupta, Washington State University

IDWSDS 2024

Session Info

K3

KEYNOTE SESSION

October 8th, 18:00 - 19:00 UTC

When Prediction Meets Causal Inference

CLICK HERE TO ACCESS THE ZOOM ROOM

This talk will focus on statistical methods used in health research, where common aims are to predict the risk of an adverse outcome or estimate the causal effect of a medical intervention. The tasks of making predictions and investigating causal effects are often viewed as separate. However, tools traditionally used in prediction modelling are increasingly used to help to solve certain challenges in causal inference, and it has also been shown how causal concepts are important in the context of clinical prediction modelling. For example, prediction methods for estimating outcomes conditional on a set of covariates, including machine learning methods, are used to fit nuisance models which are ingredients to procedures for estimating causal effects. In the other direction, it is often of interest to use risk predictions to inform whether a person should initiate a particular treatment, but it has been underappreciated that this requires causal considerations and analysis tools. This talk will discuss what prediction can do for causal inference and vice versa.

Dr. Ruth Keogh

CENTRE FOR DATA AND STATISTICS SCIENCE FOR HEALTH (DASH) AT LONDON SCHOOL OF HYGIENE AND TROPICAL MEDICINE

Ruth Keogh is Professor of Biostatistics & Epidemiology at the London School of Hygiene & Tropical Medicine (LSHTM), where she is co-director of the LSHTM Centre for Data and Statistical Science for Health (DASH). Ruth's research focuses on statistical methodology for the analysis of observational health data with a particular emphasis on methods for analysis of time-to-event data and causal inference methods. She has also been involved in a number of areas of applied health research using data arising from patient registries and electronic health records, including in cystic fibrosis, cancer, organ transplantation, Covid-19, and kidney disease.

ORGANIZER: Sarah Lotspeich, Wake Forest University

CHAIR: Sarah Lotspeich, Wake Forest University

IDWSDS 2024

Session Info

K4

KEYNOTE SESSION

October 8th, 23:00 - October 9th, 00:00 UTC

Shaping the Future: Empowering Women in Statistics and Data Science

CLICK HERE TO ACCESS THE ZOOM ROOM

Despite a growing proportion of college degrees earned by women, their representation in the mathematical sciences remains disproportionately low. With the proliferation of data in our society, the need for people trained in statistics and data science is increasing. Addressing the gender gap in the math sciences requires a concerted effort to empower and support women entering these fields. In this talk, I will share my journey into statistics, highlighting how mentorship and support have been pivotal to my success. I will explore actionable strategies to foster the growth of young women in statistics and data science, drawing from my personal experiences and leadership insights. My hope is to contribute to closing the gender gap and to inspire and elevate the next generation of leaders in statistics and data science.

Dr. Leslie McClure

SAINT LOUIS UNIVERSITY

Leslie McClure is Dean of the College for Public Health and Social Justice at Saint Louis University. Prior to joining the SLU community, she was Professor & Chair of the Department of Epidemiology and Biostatistics and Associate Dean for Faculty Affairs at the Dornsife School of Public Health at Drexel University. She earned a Bachelor of Science in Mathematics from the University of Kansas, an MS in Preventive Medicine and Environmental Health from the University of Iowa, and her PhD in Biostatistics from the University of Michigan. Dr. McClure does work to try to understand health inequities, particularly racial and geographic, and the role that the environment plays in them. Her methodological expertise is in the design and analysis of multicenter trials, as well as issues of multiplicity in clinical trials. Dr. McClure is a Fellow of the American Statistical Association, the Society for Clinical Trials, and of the American Heart Association and completed the Executive Leadership in Academic Medicine (ELAM) Fellowship. In addition to her research, Dr. McClure is passionate about increasing diversity in the math sciences, advocates for women and minorities in science, and devotes considerable time to mentoring younger scientists. When she is not working, Dr. McClure enjoys reading, spending time outdoors and with her family, and watching NCIS reruns.

ORGANIZER: Sarah Lotspeich, Wake Forest University

CHAIR: Sarah Lotspeich, Wake Forest University

IDWSDS 2024

Invited Sessions

IDWSDS 2024

Session Info

S0

OPENING SESSION

October 8th, 00:00 - 00:30 UTC

Opening

CLICK HERE TO ACCESS THE ZOOM ROOM

This session will feature Caucus for Women in Statistics and Data Science (CWS) President, Cynthia Bland, American Statistical Association (ASA) President-elect, Ji-Hyun Lee, and the Vice President of the Portuguese Statistical Society (SPE), Lisete Sousa, as we kick off the third annual International Day for Women in Statistics and Data Science.

Dong-Yun Kim will host a brief presentation covering important conference information and tips for navigating this exciting event.

This session is open to all conference participants and provides a great opportunity to ask questions.

CYNTHIA BLAND
CWS President

JI-HYUN LEE
ASA President-elect

LISETE SOUSA
SPE Vice President

DONG-YUN KIM
Program Co-Chair

IDWSDS 2024

Session Info

S1

TECHNICAL SESSION

October 8th, 00:30 - 01:30 UTC

Graduate and Undergraduate Students' Research in Statistics and Data Science from United States and South Korea

CLICK HERE TO ACCESS THE ZOOM ROOM

In this session, two students from the United States (one undergraduate and one graduate) and two students from South Korea (both graduate) will present their current research topics, discussing the analysis of gaze-based data in attention dynamics, Bayesian models for spatial datasets and clustering, continuous monitoring of the mean for different distributions, and a new R package that uses win time to measure treatment effects. Our first speaker is Amia Graye, and she will discuss how the margin of error and the minimum sample size are used to address the uncertainty of parameter monitoring and compute a sequential confidence interval for the mean. Eugene Hwang is the second speaker, and she will present on the study of visual attention within dynamic media environments using gaze-based data and attentional fluctuations. Our third speaker is Mongju Jeong, and she will discuss her research that addresses cluster-wise variations and effect changes across boundaries using Bayesian partitioning with Voronoi tessellations. Sam Lawrence is our last speaker, and he will talk about how to use win time to analyze clinical trials with composite endpoints and introduce a new R package he created for this method.

ORGANIZER: Amia Graye, Graduate Student

CHAIR: Ruba Shalhoub, Statistician

IDWSDS 2024

Speaker Bios

S1

MS. AMIA GRAYE

Graduate Student

Amia Graye is a current master's student at Johns Hopkins University Bloomberg School of Public Health in Biostatistics program. Prior to this, she worked at the US National Institutes of Health at the National Heart, Lung, and Blood Institute in the Office of Biostatistics Research. She worked closely under the advisement of Dong-Yun Kim, PhD, veteran mathematical statistician and former president of the Caucus for Women in Statistics and Data Science. Here she studied sequential monitoring and worked under Marcus Carlsson, MD, PhD, in his cardiovascular magnetic resonance imaging lab as a statistical analyst. She earned a Bachelor of Arts degree in Mathematics with a pre-med concentration from Georgetown University.

MS. EUGENE HWANG

Graduate Student

Eugene Hwang holds a dual Bachelor of Arts degree in Media and Communications and Business from Korea University and a Master of Science degree from the Graduate School of Culture and Technology at KAIST. Currently, Eugene is a Ph.D. candidate in the same graduate school and a researcher at the Visual Cognition Lab. Her areas of research interest include Human-Computer Interaction, Virtual Reality and Augmented Reality, attention, and the application of cognitive psychology. Eugene's research focuses on the use of eye-tracking and electroencephalography (EEG) methods.

MS. MONGJU JEONG

Graduate Student

Mongju Jeong is currently a PhD candidate in Statistics at Seoul National University, where she also earned her MD in Statistics. She is part of Professor Chae Young Lim's Spatial Statistics Lab. She holds a Bachelor's degree in Mathematics with a double major in Applied Statistics from Yonsei University. Her research interest lies in spatiotemporal modeling and clustering, and she is deeply engaged in advancing these areas through her studies and research.

MR. SAM LAWRENCE

Undergraduate Student

Samuel Lawrence is a current undergraduate student at Hood College pursuing a Bachelor of Arts degree in mathematics and a Bachelor of Science degree in computer science. Prior to this, he worked at the US National Institutes of Health, National Heart Lung and Blood Institute in the Office of Biostatistics Research. He worked under James Troendle, PhD, a statistician and senior investigator. Here he was exposed to win statistics and introduced to the idea of "win time." He spent the summer developing an R package that implements the win time methods.

Abstracts

Key Parameter Sequential Monitoring: Computing a Sequential Confidence Interval for the Mean and Discussing its Variability Amongst Different Distributions

Amia Graye, Graduate Student

In this talk, we introduce a method for continuous monitoring for the mean under different distributions. Using the margin of error and the minimum sample size needed for the study, we address the uncertainty of parameter monitoring during a trial or clinical study and compute a sequential confidence interval for the mean if the data meets the stopping criteria. During the trial, the continuous monitoring can detect if the empirical mean is substantially different from the target mean. The accrued data can be used to answer questions that may be useful for adaptive purposes. We illustrate the method using data from a large Phase III clinical trial.

What is your focus?: Gaze Data Based Research in Attention Dynamics

Eugene Hwang, Graduate Student

This presentation delves into the study of visual attention within dynamic media environments, with a particular focus on XR technologies. By analyzing gaze-based data and attentional fluctuations, this research uncovers key insights into cognitive processes, particularly in how attention is managed and sustained in complex virtual spaces. The findings have implications for various real-world applications, including automated media production, personalized learning, and accessibility solutions. This talk will also explore ongoing work aimed at defining gaze-based signatures of attention, with the ultimate goal of bridging digital gaps and enhancing human experiences through technology.

Exploring Spatial Dynamics in Regression Coefficients: A Bayesian Regularization Method with Clustering

Mongju Jeong, Graduate Student

Many fields are seeing expansive spatial datasets with increasing observations and covariates on a large spatial domain. As study domains expand, simple spatial assumptions of homogeneity in relationships might not capture complex dynamics. Using Bayesian partitioning with Voronoi tessellations, we address cluster-wise variations and effect changes across boundaries. We propose a Bayesian regularized spatially clustered coefficient (BRSCC) regression model, which can identify key covariates affecting the response variable and their spatial cluster patterns. The joint execution is facilitated through a latent indicator variable that shapes the prior model for each regression coefficient and cluster arrangement. We've developed a reversible jump MCMC-based algorithm and tested the model's effectiveness through simulation studies.

wintime R Package for Analysis of Time to Event Data in Clinical Trials

Sam Lawrence, Undergraduate Student

In this presentation, we introduce new methods of analyzing clinical trials and showcase the way they are implemented in a versatile, user-friendly software package. We will explore the foundational idea of these new methods, called "win time." The win time methods measure treatment effects by using the excess time spent in a better clinical state by the treatment arm. This methodology is appealing for ordered composite endpoints because it captures the entire clinical experience of the patient during a trial. Each win time method offers its own unique advantages and uses this idea in a different way. The provided software package will make these calculations quick and easy due to the numerous tools it provides to users.

IDWSDS 2024

Session Info

S2

TECHNICAL SESSION

October 8th, 00:30 - 1:30 UTC

Advances in Statistical Inference

CLICK HERE TO ACCESS THE ZOOM ROOM

The session is dedicated to the latest advancements in mathematical statistical inference. We aim to present and explore recent developments innovative in either methodological or theoretical aspects, or both. The studies address the challenges imposed by the complex data structure, such as high-dimensionality, and temporal and cross-sectional dependence. The talks cover a range of topics, including change point tests in spatial time series, tests of independence between random vectors, and the inference in non-parametric least squares.

ORGANIZER: Dr. Mengyu Xu, University of Central Florida

CHAIR: Jessica Kohlschmidt, The Ohio State University

IDWSDS 2024

Speaker Bios

S2

DR. LIKAI CHEN

University of Washington in St. Louis

Dr. Likai Chen is an Assistant Professor of Mathematics and Statistics at Washington University in St. Louis. She received her Ph.D. degree in statistics at the University of Chicago in 2018. Her Ph.D. adviser is Prof. Wei Biao Wu. Prior to graduate school, she obtained her B.S. in mathematics at Tsinghua University in 2013.
She is intersted in high dimensional data analysis, time series, statistical learning theory.

DR. MEIMEI LIU

Virginia Tech

Meimei liu is an Assistant Professor in the Department of Statistics at Virginia Tech. Prior to VT, she worked with Prof. David B. Dunson as a post-doc in Department of Statistical Science at Duke University, where she developed the methodology for nonparametric Bayesian, machine learning, and neural imaging. She was a Ph.D. student in Purdue University, where she was advised by Prof. Guang Cheng and Prof. Zuofeng Shang. She received a M.S. in statistics at University of Science and Technology of China, where she was supervised by Prof. Weiping Zhang.
Her research interests include deep learning, Bayesian data analysis, learning theory, big data analysis, and semi/non-parametric inference.

DR. DANNA ZHANG

University of California, San Diego

Danna Zhang received her Ph.D. in statistics at the University of Chicago under the supervision of Wei Biao Wu and Peter McCullagh. Her research areas include high-dimensional significance testing, change-point detection, high-order statistics, nonlinear/non-stationary time series, time-varying network and random graph. She is currently interested in high-dimensional time series.

Abstracts

Adaptive Two-Way MOSUM: Testing for Multiscale Change Points in Spatial Time Series

Likai Chen, University of Washington in St. Louis

Moving sum (MOSUM) test statistic is popular for multiple change-point detection due to its simplicity of implementation and effective control of the significance level for multiple testing. However, its performance heavily relies on the selection of the bandwidth parameter for the window size, which is extremely difficult to determine in advance. To address this issue, we propose an adaptive MOSUM method, applicable in both multiple and high-dimensional time series models. Specifically, we adopt an ℓ^2-norm to aggregate MOSUM statistics cross-sectionally, and take the maximum over time and bandwidth candidates. We provide the asymptotic distribution of the test statistics, accommodating general weak temporal and cross-sectional dependence. By employing a screening procedure, we can consistently estimate the number of change points, and the convergence rates for the estimated timestamps and sizes of the breaks are presented. The asymptotic properties and the estimation precision are demonstrated by extensive simulation studies. Furthermore, we present an application using real-world COVID-19 data from Brazil, wherein we observe distinct outbreak stages among subjects of different age groups and geographic locations. These findings may facilitate analysis of epidemics, pandemics, and data from various fields of knowledge exhibiting similar patterns.

Scalable Statistical Inference in Non-parametric Least Squares

Meimei Liu, Virginia Tech

Stochastic approximation (SA) is a powerful and scalable computational method for iteratively estimating the solution of optimization problems in the presence of randomness, particularly well-suited for large-scale and streaming data settings. In this work, we propose a theoretical framework for stochastic approximation (SA) applied to non-parametric least squares in reproducing kernel Hilbert spaces (RKHS), enabling online statistical inference in non-parametric regression models. We achieve this by constructing asymptotically valid pointwise (and simultaneous) confidence intervals (bands) for local (and global) inference of the nonlinear regression function, via employing an online multiplier bootstrap approach to a functional stochastic gradient descent (SGD) algorithm in the RKHS. Our main theoretical contributions consist of a unified framework for characterizing the non-asymptotic behavior of the functional SGD estimator and demonstrating the consistency of the multiplier bootstrap method. And the theory specifically reveals an interesting relationship between the tuning of step sizes in SGD for estimation and the accuracy of uncertainty quantification.

Test of Independence Based on Generalized Distance Correlation

Danna Zhang, University of California, San Diego

We study the fundamental statistical inference concerning the testing of independence between two random vectors. Existing asymptotic theories for test statistics based on distance covariance can only apply to either low dimensional or high dimensional settings. In this work we develop a novel unified distributional theory of the sample generalized distance covariance that works for random vectors of arbitrary dimensions. In particular, a non-asymptotic error bound on its Gaussian approximation is derived. Under fairly mild moment conditions, the asymptotic null distribution of the sample generalized distance covariance is shown to be distributed as a linear combination of independent and identically distributed chi-squared random variables. We also show high dimensionality is necessary for the null distribution to be asymptotically normal. To estimate the asymptotic null distribution practically, we propose an innovative Half-Permutation procedure and provide the theoretical justification for its validity. The exact asymptotic distribution of the resampling distribution is derived under general marginal moment conditions and the proposed procedure is shown to be asymptotically equivalent to the oracle procedure with known marginal distributions.

IDWSDS 2024

Session Info

S3

TECHNICAL SESSION

October 8th, 00:30 - 01:00 UTC

Innovative Software Tools for Data Insight Generation

CLICK HERE TO ACCESS THE ZOOM ROOM

Analyzing and reporting the data collected in a clinical trial follows highly formalized and standardized procedures. The results of the analyses are routinely summarized in the form of listings, tables, and figures (TLFs). However, this traditional approach has its limitations, given the increasing complexity of the study designs and the resulting ever-growing amounts of data collected especially in Phase 3 trials.

In this talk, we introduce our newly established concept for how all parties involved in a pharmaceutical company in the evaluation of study data can develop together a solid understanding of the data in a very inspiring and highly efficient way. Further, we present the software tools we currently use for this approach, which we would like to share with the scientific community.

ORGANIZER: Dr. Erya Huang, Bayer U.S. LLC.

CHAIR: Dr. Erya Huang, Bayer U.S. LLC.

Speaker Bios

DR. ERYA HUANG

Bayer U.S. LLC.

Dr. Huang is the Associate Director at Bayer U.S. LLC. She has many years of experiences working on global drug developments and registrations. She is also the main contact person of the data visualization app center in Bayer North America since 2016, being passionate about introducing efficient data visualization apps to the communication among statisticians and non-statisticians. Dr. Huang received her Ph.D. in statistics from Stony Brook University.

IDWSDS 2024

Session Info

S4

STATISTICS EDUCATION, HISTORY, AND DATA ETHICS

October 8th, 01:00 - 01:30 UTC

Statistics' Hidden History: Places Where Social Justice Issues Changed Our Narrative

CLICK HERE TO ACCESS THE ZOOM ROOM

Academic statistics/data science programs spend little (or no) time discussing our history; however, like other scientific fields, the social context in which our theory and methods developed directly impacts our practice today. Similarly, few academic programs teach students about the myriad of ethical issues that can arise for a statistician or data scientist, and we rarely review case studies related to ethics violations involving data. This needs to change, if we wish to avoid repeating the mistakes of the past.

This session will elucidate specific issues from our history as a field. We will start by discussing how the British Eugenics movement was deeply entwined with the development of inferential statistics, and focus on Ronald Aylmer Fisher's life as a case study. We will then discuss how social marginalization led us to "forget" the contributions of William DuBois to data visualization; specifically, he arguably created interactive data visualization, an idea that subsequently lay dormant for almost 100 years. Finally, we will discuss our current practice and reasons why every statistician and data scientist should have a keen understanding of our past mistakes to guide their decision-making during every step of data analysis, from data collection to dissemination of results. Attention will be paid to current controversies regarding the collection of gender data and the marginalization of women and nonbinary people in data.

ORGANIZER: Dr. Jana Asher, Goucher College

CHAIR: Sarah Lotspeich, Wake Forest University

Speaker Bios

DR. JANA ASHER

Goucher College

https://www.goucher.edu/faculty/jana-asher

Jana Asher is an Associate Professor of Data Science at Goucher College in Baltimore, MD, USA. Asher earned an MS (1999) & PhD (2016) in statistics from Carnegie Mellon University. Asher is serving a 2023-2025 term on the American Statistical Association's (ASA) Board of Directors & is Chair of the Committee on the History of Statistics of the International Statistics Institute (ISI). Asher is an internationally recognized statistician for her work on human rights & sexual violence. As well as authoring/coauthoring over 65 peer-reviewed articles, book chapters, and manuscripts, she was the lead editor for the book Statistical Methods for Human Rights. Asher was elected a Fellow of the ASA in 2009 and a member of the ISI in 2010. In 2022 she received the CWS Societal Impact Award for "her work combating societal injustice through accurate and ethical quantitative measurement, and for her commitment toward teaching civic responsibility and JEDI principles through statistical practice.

IDWSDS 2024

Session Info

S5

TECHNICAL SESSION

October 8th, 01:30 - 02:00 UTC

High School Research: Confidence Interval Accuracy, AI Ethics in Healthcare Law, and AI Applications in Space Science

CLICK HERE TO ACCESS THE ZOOM ROOM

Three high school students from the Midwest present their research. One student developed a program to provide accurate confidence intervals, thereby addressing challenges in clinical studies with small proportions. One student explores the ethical issues of AI in healthcare and current legislation through a review of recent literature and case studies. The last student investigates the role of machine learning in space science and, more specifically, forecasting astronomical phenomena. All three papers encompass the applications of data science.

ORGANIZER: Mulan Wu, University of Chicago Laboratory Schools

CHAIR: Menyu Xu, University of Central Florida

IDWSDS 2024

Speaker Bios

S5

MULAN WU

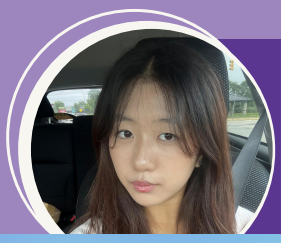
University of Chicago Laboratory Schools

I am a junior at the University of Chicago Laboratory Schools with a keen interest in biology, medicine, and statistics.
I have earned state-level medals in Science Olympiad, including a 4th place medal in Forensics and a 5th place medal in Green Generation, along with regional and invitational level medals in high school.
In my free time, I play the violin, and I have earned a 2nd place trophy in the Senior Division in the Illinois Music Association Competition in June 2024.

DANIEL WU

University of Chicago Laboratory Schools

I am a high school junior at the University of Chicago Laboratory Schools with a passion and determination to intersect STEM and legal applications.
I have earned numerous medals for Science Olympiad, received the Creativity in Engineering award for FIRST Robotics, interned for an intellectual property law firm in New York, and contributed to a start-up family law clinic for pro se litigants in high school.
I've also played piano for 9 years and won a number of regional and state awards.



JIAQI HUANG

Huron High School

Jiaqi Huang is a high school student in Ann Arbor Michigan who has a passion for applied maths, statistics and computer science. She has worked with researchers at the University of Michigan Transportation Research Institute to conduct surveys and co-publish papers as well as collaborating with the University of Michigan Statistics Department to develop a set of machine learning models that monitor and predict solar flare movement. Twice awarded the honor of being in the Mathematical Association of America's AMC Young Women in Mathematics Award and Certificate for her performance on the AMC 10, Jiaqi is active in competition math as both a competitor and a coach at her local Chinese school.

Confidence Intervals Based on the Modified Chi-Squared Distribution and its Applications in Medicine

Mulan Wu, University of Chicago Laboratory Schools

Small sample sizes in clinical studies offer advantages such as reduced costs, limited subject availability, and the rarity of studied conditions. However, this creates challenges for accurately calculating confidence intervals using the normal distribution approximation. In this talk, we employ a modified equation from Xu et al. to provide more accurate confidence intervals, particularly for data with small sample sizes or proportions. Based on the simulations, we suggest reasonable values of sample sizes and proportions for the application of the quadratic method. Consequently, this method enhances the reliability of statistical inferences. We illustrate this method with real medical data.

The End of the World: AI Ethics in Medicine and Healthcare Law

Daniel Wu, University of Chicago Laboratory Schools

Artificial intelligence (AI) is an powerful technological development which simulates the problem-solving capabilities of the human mind. AI softwares are dynamic; they consistently improve their abilities to learn, recognize patterns, and predict responses based on growing data sets and user input. This allows for a wide range of practical AI applications in countless fields and professions. However, this paramount achievement in technology induces controversial and ethical questions in medicine and healthcare. Issues of explainability, algorithmic bias, patient confidentiality, and sustainability challenge a ubiquitous reliance on AI in the medical field. This paper provides a response to the development of AI, its ethical implications, and its applications to the future of medicine and healthcare through a review of recent and relevant academic literature. Additionally, two relevant cases decided between the years 2023 and 2024 (or currently in decision) and in American jurisdiction were included for individual review.

Machine Learning Applications in Optimizing Solar Flare Predictions

Jiaqi Huang, Huron High School

This talk seeks to explore and introduce the ways and roles of machine learning in the field of space science. Much of the research done in space science relies on the use of forecasting various astronomical phenomena and the idea of using machine learning to aid with such a nuanced task is a new and promising development. Machine learning can play a versatile role in optimizing forecasting accuracy for a multitude of different purposes. In this session, common uses of machine learning in space science will be covered as well as specific techniques and methods used to develop machine learning models geared towards forecasting accuracies and the various pitfalls and issues that arise with using machine learning to predict astronomical events.

IDWSDS 2024

Session Info

S6

TECHNICAL SESSION

October 8th, 01:30 - 02:30 UTC

Advancing Healthcare with Data Science and Machine Learning: The Future of Data-Driven Care

CLICK HERE TO ACCESS THE ZOOM ROOM

This session delves into the intersection of data science and machine learning in healthcare, focusing on how these technologies are revolutionizing personalized medicine. We will explore how advanced algorithms and multimodal data integration are being applied to enhance the diagnosis, prognosis, and treatment of complex conditions. The session also addresses the ethical and practical challenges of implementing AI in clinical settings, emphasizing the importance of responsible and equitable approaches. Through these discussions, the session aims to highlight the potential of data science to transform healthcare by enabling more accurate, individualized care strategies.

ORGANIZER: Divya Sharma, York University

CHAIR: Jessica Kohlschmidt, The Ohio State University

IDWSDS 2024

Speaker Bios

S6

DR. DIVYA SHARMA

York University

Dr. Divya Sharma is an Assistant Professor in the Department of Mathematics and Statistics at York University, with cross-appointments as an Assistant Professor (status-only) at the Dalla Lana School of Public Health, University of Toronto, and as a Clinician Investigator at the University Health Network (UHN). Dr. Sharma holds a PhD in Computer Science with a specialization in Machine Learning from the Indian Institute of Technology Jodhpur. Following her doctoral studies, she completed a postdoctoral fellowship at the Biostatistics Department of the Princess Margaret Cancer Centre, UHN. Her research program focuses on developing bespoke deep learning models for integrative, high-dimensional modeling of multimodal big healthcare data, with a strong emphasis on clinical interpretability. Her innovative modeling approaches, published in journals like Lancet Digital Health and Bioinformatics, emphasize developing robust ML models for personalized medicine and clinical deployability.

DR. SOUMITA GHOSH

University Health Network

Soumita earned her doctoral degree in biostatistics from the Saw Swee Hock School of Public Health, National University of Singapore, where she developed bioinformatics tools for large-scale visualization of multi-omics data and mining high-throughput datasets. As a Research Fellow in the Department of Medicine, Yong Loo Lin School of Medicine, Singapore, she developed tools for predicting cancer-related outcomes, integrating genotype and phenotype information. Subsequently, at the Cancer Science Institute of Singapore, she worked on developing machine learning-based drug recommendation algorithms for refractory cancer patients at the National University Hospital, Singapore. Currently, as a Schmidt AI in Science postdoctoral fellow in Dr. Bhat's lab she is applying machine learning based methods integrating omics and clinical data for improving outcomes in liver diseases.

DR. GHAZAL AZARFAR

University Health Network

Ghazal earned her PhD from the University of Wisconsin, where she specialized in analytical chemistry and bioanalysis, utilizing machine learning to analyze chemical images and explore how single cells respond to environmental stress. Driven by a passion for innovative multimodal AI solutions to improve clinical practice, Ghazal is now a member of Dr. Bhat's lab within the Transplant AI Initiative. In this role, she integrates clinical expertise, computer science, and data analytics to advance transplant medicine, with a particular focus on managing and understanding post-transplant complications, including the recurrence of hepatocellular carcinoma, sepsis, and lymphoproliferative disorders.

Abstracts

Harnessing Data Science and Machine Learning for Personalized Osteoarthritis Treatment: A Multimodal Deep Learning Approach

Divya Sharma, York University

Osteoarthritis (OA) is a complex condition that varies widely among patients, making it challenging to predict outcomes and tailor treatments. In my talk, I'll present a novel deep learning framework that integrates multiple types of biological data to uncover hidden patterns in OA. Using Variational Autoencoders (VAEs), we tackled the challenge of analyzing high-dimensional data—a common hurdle known as the "curse of dimensionality." Our approach brings together diverse datasets, including genetic, biochemical, and molecular information, to identify distinct patient subgroups, or endotypes. These endotypes provide new insights into the underlying mechanisms of OA. We further advanced this work by creating a machine learning model that uses these insights to predict how patients might respond to knee replacement surgery, particularly in terms of pain relief and functional improvement. This study exemplifies how cutting-edge data science, combining deep learning with machine learning, can integrate complex data from multiple domains, paving the way for more personalized and effective treatments in health research.

Advancing Personalized Diagnosis, Prognosis, and Treatment of Liver Diseases through Machine Learning and Data Science applied to Omics and Clinical Data

Soumita Ghosh, University Health Network

Advancements in omics technologies and artificial intelligence (AI) are revolutionizing personalized treatment approaches for liver diseases. AI, including both traditional machine learning (ML) and emerging deep learning (DL) algorithms, leverages data-driven methods to efficiently analyze high-throughput data, uncovering hidden patterns and generating insights that can inform clinical decisions. Despite the potential for richer insights from omics data, high costs have limited their widespread adoption in clinical settings. Applying advanced AI techniques to multi-omics datasets requires substantial data for training and validation, and challenges remain in achieving unbiased results with clinical relevance. This review provides an overview of the different omics levels investigated in various liver diseases and categorizes the AI methodologies employed in these studies. We discuss strategies to address data limitations and capitalize on emerging opportunities. Given the significant global burden of chronic liver diseases, establishing multi-center collaborations to generate large-scale omics datasets is crucial for early disease detection and intervention. Additionally, exploring advanced AI methods is essential to fully utilize these datasets and enhance early detection and personalized treatment strategies.

Leveraging Data Science for Responsible Multimodal AI in Healthcare: Promises and Challenges

Ghazal Azarfar, University Health Network

Clinicians routinely use multiple data modalities—such as physical assessments, patient history, clinical signs, vital signs, imaging, laboratory tests, and microbiology results—to enhance decision-making for individual patients. Multimodal artificial intelligence (AI) systems are a recent advancement in AI, showing potential to process and analyze these diverse data sources. However, implementing such models in clinical practice presents significant challenges due to data heterogeneity and integration complexities. Adopting multimodal AI in healthcare requires both technological innovation and careful consideration of ethical, legal, and practical implications to ensure safe and effective patient outcomes. This talk will explore the current landscape of multimodal AI, highlighting the opportunities and challenges in clinical settings. I will provide practical examples of how multimodal AI can enhance clinical practice, such as in early sepsis diagnosis, cardiology, and neonatal intensive care units. Additionally, I will address deployment issues and offer actionable insights for overcoming barriers. Key challenges include managing multimodal data, model selection, validation, generalization, explainability, interpretability, ease of use, safety, liability, fairness, and international considerations. By addressing these challenges, we can harness the potential of multimodal AI to transform healthcare delivery, provide individualized care, and improve patient outcomes.

IDWSDS 2024

Session Info

S7

STATISTICAL/DATA SCIENCE EDUCATION AND OUTREACH

October 8th, 01:30 - 02:00 UTC

Diversity in Data Science: Empowering Women Through Outreach

CLICK HERE TO ACCESS THE ZOOM ROOM

In 2024, the Data Science Council of America reported that although 57% of women make up the overall workforce, only 27% of employees in the technology industry are women. We are currently still facing the challenge of insufficient representation of women and other gender minorities in data science and statistics. However, it is has been observed that perhaps this gender imbalance may start from a much earlier point than when individuals are entering the workforce. The "leaky pipeline" is a term that describes the phenomenon of the loss of women in STEM fields , despite them having strong abilities and interest in these domains. To help overcome systemic barriers that women face in these industries, I believe it is important that we impact students early on to expose them to new topics and perhaps inspire them to pursue careers in data science. In this talk, I will discuss my experiences with data science outreach and how engaging our communities can promote future women in the field.

ORGANIZER: Katie Burak, University of British Columbia

CHAIR: Katie Burak, University of British Columbia

SPONSOR: University of British Columbia

Speaker Bios

DR. KATIE BURAK

University of British Columbia

Dr. Katie Burak is an Assistant Professor of Teaching in the Department of Statistics at the University of British Columbia. She is also a member of the academic team for UBC's Masters of Data Science program. Prior to this, she completed her PhD in Statistics at the University of Alberta. Katie has a passion for statistical outreach and education and has organized a variety of outreach events in her community. Additionally, she is interested in investigating modern pedagogical methods and best teaching practices in the field of data science.

IDWSDS 2024

Session Info

S8

TECHNICAL SESSION

October 8th, 02:00 - 03:00 UTC

Advances in Statistical Learning

CLICK HERE TO ACCESS THE ZOOM ROOM

Advances in statistical learning involves multidimensional problems and density estimation for probability distributions.

New Multidimensional clustering and nonparametric density estimation with penalization methods are provided.

These statistical learning methods can be used in practice including clinical research.

ORGANIZER: Jaehee Kim, Duksung Women's University

CHAIR: Hye-Young, Hanyang University

SPONSOR: WISK (Women in Statistics in Korea)

IDWSDS 2024

Speaker Bios

S8

PROF. SOON-SUN KWON

Ajou University

https://scholar.google.com/citations?user=JSOv5boAAAAJ&hl=ko

She got the bachelor' degree from Ajou University, master and Ph.D. from Seoul National University.
Her research interests are Medical statistics, Longitudinal data analysis, and Clinical Trials.

PROF. KYOUNG HEE KIM

Korea University

https://sites.google.com/site/kyoungheearlene

She got the bachelor' degree and master's from Korea University and Ph.D. from Yale University.
She was postdoctoral associate at the Statistical Laboratory in the University of Cambridge.
Her research interests broadly lie in statistical theory and methodology.

PROF. JEONGYOUN AHN

KAIST

She got the bachelor' degree and master's from Seoul National University and Ph.D. from UNC-Chapel Hill.
Her research interest is high-dimensional statistical learning methodology.

Clustering for multi-dimensional functional data in Clinical research

Soon-Sun Kwon, Ajou University

Multi-dimensional functional data analysis has become a contemporary research topic in medical research as patients' various records are measured over time. Functional data analysis, introduced by Ramsay [1] and Ramsay and Dalzell [2] has been increasingly used for modeling, predicting, and analyzing data with functional structures. Traditional statistical approaches such as analysis of variance, cluster analysis, discriminant analysis, outlier detection, principal component analysis, and regression analysis have been extended to functional data analysis. In this study, there're proposed two clustering methods using the Fréchet distance for multi-dimensional functional data. The first method extends an existing K-means type approach from one-dimensional to multi-dimensional longitudinal data. The second method enforces sparsity on functional variables while grouping observed trajectories and enables us to assess the contribution from each variable. Both methods utilize the generalized Fréchet distance to measure the distance between trajectories with irregularly spaced and asynchronous measurements. It produces interpretable clusters and weighs the importance of functional variables through a comparative study using various simulation examples and real data analysis.

REFERENCES
1. Ramsay, J.O. When the data are functions. Psychometrika 47, 379–396 , 1982
2. Ramsay, J.O., Dalzell, C.J. Some tools for functional data analysis. J. Roy. Stat. S. B. 53, 539–572, 1991

Density estimation using Total variation regularization

Kyoung Hee Kim, Korea University

We study the problem of nonparametric estimation of an unknown density on the real line using penalized maximum likelihood, where the penalty is based on the total variation of an appropriate derivative of the log-density. This estimator has been around for a while, but many theoretical properties including rates of convergence are unavailable. We prove such rates of convergence, and explain connections to shape-constrained density estimation.

Finding hidden ordinal labels via monotone clustering

Jeongyoun Ahn, KAIST

There has been a growing demand in various sectors, such as healthcare, finance, and social sciences, for clustering methods that not only find quality clusters in data but also provide inherent orders within the clusters for better decision-making and risk assessment. Traditional clustering methods, though effective at grouping data, often fall short in delivering interpretable and structured clusters. This paper introduces Monotone Clustering (MOCL), a novel unsupervised algorithm specifically designed to address these challenges by focusing on ordinal interpretability. MOCL integrates generalized additive models with clustering techniques in an iterative manner. A remarkable benefit of MOCL is that it can select variables that are monotonically related to the found clusters, which greatly enhances MOCL's applicability to high-dimensional datasets. The algorithm's effectiveness is demonstrated through both synthetic and real-world datasets.

IDWSDS 2024

Session Info

S9

TECHNICAL SESSION

October 8th, 03:00 - 04:00 UTC

Advanced Statistical Methods for Complex Data Challenges

CLICK HERE TO ACCESS THE ZOOM ROOM

This session brings together innovative statistical approaches to address challenges in analyzing complex and high-dimensional data. Topics include a Bayesian quantile regression framework for handling incomplete longitudinal medication data, a unified procedure for small area estimation under informative sampling, and the application of Empirical Bayes methods to high-dimensional data in various fields. These talks collectively showcase cutting-edge methodologies that enhance the accuracy and reliability of statistical inference in diverse and challenging settings.

ORGANIZER: HoYoung Park, Department of Statistics, Sookmyung Women's University

CHAIR: YangJin Kim, Department of Statistics, Sookmyung Women's University

SPONSOR: Department of Statistics, Sookmyung Women's University

IDWSDS 2024

Speaker Bios

S9

PROF. MINJAE LEE

University of Texas Southwestern

Dr. MinJae Lee is an Associate Professor of Biostatistics in O'Donnell School of Public Health at University Texas Southwestern. She developed new statistical methods that can deal with various types of measurement issues in longitudinal/multi-level data on biomarkers, cancer-preventive behaviors, environmental measurements.

DR. YANGHYEON CHO

Columbia University

Yanghyeon Cho is a Postdoctoral Research Scientist at the Department of Biostatistics and the Center for Statistical Genetics, The Gertrude H. Sergievsky Center, Columbia University, New York, NY, USA.

PROF. HOYOUNG PARK

Department of Statistics, Sookmyung Women's University

Hoyoung Park is an Assistant Professor in the Department of Statistics at Sookmyung Women's University. His research focuses on high-dimensional non-parametric estimation techniques, multiple-testing procedures, and medical data analysis, particularly in the field of epidemiology. Park's experience includes a postdoctoral fellowship at the NIH, where he collaborated on epidemiologic and community health research. His work is characterized by interdisciplinary collaboration and a commitment to advancing statistical methodologies.

Abstracts

Disentangling unobserved heterogeneity: Statistical approaches to characterizing longitudinal trajectory of medication usage among various at-risk patients

MinJae Lee, Peter O'Donnell Jr. School of Public Health, University of Texas Southwestern

In the era of precision medicine, researchers are increasingly sensitive to the heterogeneity among at-risk individuals. Evaluating the association between disease progressions and the longitudinal pattern of pharmacological therapy has become more important. However, in many longitudinal studies, self-reported medication usage data collected at patients' follow up visits could be missing and/or inaccurate/untenable information. These patterns may also dramatically differ between individuals at varying risks, and thus complicate determining the trajectory of medication use and its complete effects for patients. Although traditional existing methods can deal with specific types of missing/incomplete data, inappropriate handling of this complex issue can lead to misleading findings especially when it depends upon multiple sources of variation over time. We propose a new statistical approach under Bayesian quantile regression framework that incorporates cluster of unobserved heterogeneity for handling medication usage data with various measurement issues. Findings from our simulation study indicate that the proposed method performs better than traditional methods under certain scenarios of data distribution. We also illustrate applications of the proposed method to real data obtained from the longitudinal study.

Optimal Predictors of General Small Area Parameters Under an Informative Sample Design Using Parametric Sample Distribution Models

YangHyeon Cho, Columbia University

Two challenges in small area estimation occur when (1) the sample selection mechanism depends on the outcome variable and (2) the parameter of interest is a nonlinear function of the response variable in the assumed model. If, given the values of the model covariates, the sample selection mechanism depends on the model response variable, the design is said to be informative for the model. Pfeffermann and Sverchkov (2007) develop a small area estimation procedure for informative sampling, focusing on prediction of small area means. Molina and Rao (2010) develop a small area estimation procedure for general parameters that are nonlinear functions of the model response variable. The method of Molina and Rao (2010) assumes noninformative sampling. We combine the approaches of Molina and Rao (2010) and Pfeffermann and Sverchkov (2007) to develop a procedure for the estimation of general parameters in small areas under informative sampling. We evaluate the validity of the proposed procedures through extensive simulation studies and illustrate the procedures utilizing agricultural survey data.

Leveraging Empirical Bayes for High-Dimensional Data: Techniques and Applications Across Diverse Fields

HoYoung Park, Department of Statistics, Sookmyung Women's University

In this talk, we explore the unique features of Empirical Bayes methodology, particularly its application to high-dimensional data analysis. Empirical Bayes approaches offer a powerful framework for combining information across different data sources, enabling more accurate and stable parameter estimates in complex settings. We will discuss how these methods are particularly well-suited for high-dimensional contexts, where traditional parametric techniques often fall short. The talk will also cover various applications of Empirical Bayes in fields such as genomics, finance, and medical research, illustrating the versatility and effectiveness of this approach in tackling real-world challenges. By the end of the presentation, attendees will gain a deeper understanding of how Empirical Bayes methods can be leveraged to enhance statistical inference and decision-making in diverse domains.

IDWSDS 2024

Session Info

S10

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

October 8th, 04:00 - 05:00 UTC

The Significance of Mentoring for NZ Women

CLICK HERE TO ACCESS THE ZOOM ROOM

While many workplaces and institutions support their early career statisticians by assigning mentors and having a strong focus on professional development, there are many others where individuals have little to no support as they navigate the early stages of their careers. While these under supported individuals can still be very effective statisticians, they may experience significant challenges and be forced to learn things the hard way. This scenario can be even more challenging for women who are often working in male dominated teams, navigating parental leave, and balancing childcare responsibilities.

Three years ago, the New Zealand Statistical Association (NZSA) launched a mentoring program to support our student and early career members. The program exists to provide opportunities for members to connect and share their expertise and to provide opportunities for networking and collaboration. This session will highlight the successes and learnings from the first three years of the NZSA mentoring program, including a panel discussion from four women who have been involved in the program as mentors and / or mentees who will share their personal experiences of the program.

ORGANIZER: Lisa Thomasen, Fonterra Co-Operative Group Ltd

CHAIR: Lisa Thomasen, Fonterra Co-Operative Group Ltd

Speaker Bios

DR. CLAIRE CAMERON

University of Biostatistics Centre, University of Otago

https://www.otago.ac.nz/healthsciences/expertise/profile?id=1088

"Claire is the Director of the Biostatistics Centre (University of Otago, New Zealand) providing biostatistical collaboration and advice to people involved with health research. Her involvement in research is varied but, usually, she contributes methodological leadership and expertise to a project. Her position is Research Associate Professor.
She has worked as a statistician in various guises since 1990 and has been involved in a wide range of application areas. She has been working as a biostatistician since late August 2011. She has been a member of the New Zealand Statistical Association since 1990 and a member of the Caucus of Women in Statistics and Data Science since 2020.
She has contributed to several networks for different academic groups including biostatisticians. She has also been involved in mentoring (formally and informally) for many years. She sees supporting colleagues at all levels and being part of these networks as an essential part of a healthy working life."

DR. RINA HANNAFORD

AgResearch

Rina Hannaford is a senior statistician based at the Grasslands campus of AgResearch, a New Zealand Crown Research Institute. Her work involves thinking about other peoples' (data analysis) problems. Rina joined the mentoring program in the first year and returned with even greater enthusiasm this year. She is looking forward to singing its praises!

DR. ALICE HYUN MIN KIM

University of Otago Wellington

https://www.otago.ac.nz/wellington/departments/central-services/biostatisticalservices/staff/alice-kim

Dr. Alice Hyun Min Kim is a Biostatistician/Senior Research Fellow based in Wellington, New Zealand. Her research interests focus on the applications of Epidemiology and Statistics in Medicine, Psychology and Public Health. She studied Economics at the University of Auckland and Harvard University, and has an MSc in Statistics from the University of Auckland and a PhD in Health Sciences from the University of Canterbury. Her teaching experiences include Epidemiology (first year) and Health Issues in the Community (postgrad) courses. In her current role at the University of Otago Wellington, she provides biostatistical and methodological input to various health research projects working with a diverse group of researchers across disciplines and institutions. She is a member of the WHO Thematic Platform for Health Emergency and Disaster Risk Management Research Network and has authored research methods chapters on natural experiment design and health disaster and emergency research data.

DR. OLIVIA ANGELIN-BONNET

The New Zealand Institute for Plant and Food Research Ltd

https://olivia-angelin-bonnet.netlify.app/

Olivia completed her PhD in Statistics at Massey University, where she worked on unravelling genotype-to-phenotype relationships from multi-omics data, with a focus on polyploid organisms. After a year as a lecturer in Statistics at Massey University, she is now a Statistical Scientist at Plant & Food Research. Her research interests include Systems Biology, multi-omics data integration, reproducible research, and the development of R packages for visualising and integrating omics data.

IDWSDS 2024

Session Info

S11

TECHNICAL SESSION

October 8th, 05:00 - 06:00 UTC

Monash EBS PhD Contest

CLICK HERE TO ACCESS THE ZOOM ROOM

This session features presentations from three PhD Students in Statistics and Data Science.

Felicia Bongiovanni (Walter and Eliza Hall Institute): Inferring time of infection by understanding antibody dynamics through a hierarchical Bayesian framework
Malvika Kharbanda (South Australian immuniGENomics Cancer Institute): Evaluating the Prognostic Potential of Transcription Factor Activity in Prostate Cancer
Simisana Nbada (University of Botswana): A Review of the use of R Programming for Data Science Research in Botswana

The best presenter in this session will be awarded a $AU500 cash prize, as judged by the expert panel consisting of:

Julie Cook (Associate of the Australian Actuaries Institute),
https://research.monash.edu/en/persons/julie-cook
Dr. Mahsa Salehi (Monash University)
https://research.monash.edu/en/persons/mahsa-salehi
Dr. Lauren Smith (Walter and Eliza Hall Institute)
https://findaresearcher.wehi.edu.au/smith.la

ORGANIZER: Michael Lydeamore, Monash Business School

CHAIR: Michael Lydeamore, Monash Business School

SPONSOR: Department of Econometrics and Business Statistics, Monash Business School

IDWSDS 2024

Session Info

S12

TECHNICAL SESSION

October 8th, 07:00 - 08:00 UTC

Research and Other Stories of Korean Women Leaders in Data Sciences

CLICK HERE TO ACCESS THE ZOOM ROOM

This session brings together the inspiring work of three Korean women leaders in data science who are using their expertise to address global challenges. Their work spans the development of new methodology for smartphone-aided data collection in conflict zones, the AI application to satellite image and geospatial data for uncovering socioeconomic patterns such as inequality, and biomedical research for understanding the genetic underpinnings of mental health conditions. We also want to inspire and empower the next generation of women data scientists by sharing our personal stories of career paths and research challenges. This session is ideal not only for data scientists but also for those who are curious about the contributions of women leaders in these fields.

ORGANIZER: Yei Eun Shin, Seoul National University

CHAIR: Yei Eun Shin, Seoul National University

Speaker Bios

DR. SEHO PARK

Hongik University

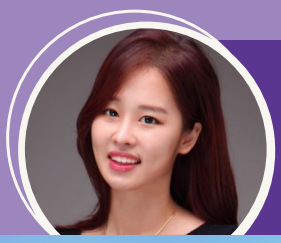
Dr. Seho Park is an Assistant Professor of Department of Industrial and Data Engineering at Hongik University in South Korea, where she specializes in survey sampling and statistical modeling for public health research. She has a PhD in Statistics from Iowa State University and worked at Indiana University School of Medicine and Regenstrief Institute prior to her current position. She has been collaborating with her colleagues on impactful public health research, focusing on advancing community health initiatives and exploring innovative solutions to health disparities. Outside of work, she enjoys hiking and reading.

DR. JIHEE KIM

Korea Advanced Institute of Science & Technologyy

http://www.jiheekim.net/

Jihee Kim is an associate professor at the School of Business and Technology Management, KAIST College of Business, and also holds a joint appointment with the School of Computing and the Graduate School of Data Science at KAIST. As an economist, her primary research focuses on economic growth and inequality, yet her academic and research endeavors have embraced interdisciplinary scholarship. She earned a B.S. in Computer Science at KAIST, then pursued her master's degree in Economics at Stanford University, followed by her PhD in Management Science and Engineering, also at Stanford University. Her academic journey reflects a commitment to interdisciplinary exploration, as evidenced by her collaborative efforts across various disciplines, such as computer science and energy policy, while maintaining a strong foundation in economics. Her latest interdisciplinary research combines artificial intelligence, satellite imagery, and geospatial data to address global socioeconomic challenges.



DR. YOONJUNG JOO

SAIHST

https://gennielab.weebly.com/

Dr. Yoonjung Joo is a biomedical data scientist and informatician with a robust academic background that spans both Korea and the United States. She earned her BS in Life Sciences with a minor in Business Administration from Korea University, followed by MS in Biotechnology from Northwestern University's McCormick School of Engineering. She completed her PhD in Health and Biomedical Informatics at Northwestern University Feinberg School of Medicine, where she conducted multi-institutional GWAS-PheWAS research within the eMERGE network, utilizing nationwide EHR-linked biobanks. After a year of postdoctoral training at Seoul National University, she joined Korea University's Institute of Data Science, where she has taught data science and AI while continuing her research on genomic and neuroimaging data. In September 2023, Dr. Joo began her role as Assistant Professor in the Department of Digital Health at SAIHST, Sungkyunkwan University and Samsung Medical Center.

Abstracts

Modified respondent-driven sampling aided with a smartphone to assess the health and nutrition status of a population in an armed conflict zone

Seho Park, Hongik University

The success of humanitarian programs largely depends on accurate information. However, crises often pose challenges in data collection from the affected populations. The modified Respondent-Driven Sampling method, introduced in our research, offers a potential solution. It helps reduce risks to surveyors and survey participants in conflict zones, and involves survey participants as resources for providing real-time security updates and assists in identifying survey households. This method, which differs from traditional RDS by integrating two-stage cluster sampling, could significantly improve the accuracy of population characteristic estimation, regardless of the estimator type. It is flexible, enabling direct estimation of household and individual characteristics. Open Data Kit is used for data collection, simplifying the survey process by eliminating paper questionnaires and correcting errors. It also allows for remote monitoring through GPS and timestamp recording, with supervisors able to intervene and repeat surveys if necessary. However, this approach requires surveyors skilled in ODK and smartphone use and a reliable telephone network for remote supervision. The definition of 'cluster' based on the catchment areas of functioning health facilities may also leave out regions without such facilities. These challenges and limitations, while significant, do not diminish the potential of this novel methodology for representative data collection in conflict-affected areas.

AI + Satellite Imagery + Geospatial Data for Good

Jihee Kim, Korea Advanced Institute of Science & Technology (KAIST)

This talk explores the transformative potential and challenges in combining AI with satellite imagery and geospatial data to tackle global socioeconomic challenges. Our initiative focuses on providing georeferenced measurements and analyses in innovative and cost-effective ways, supporting sustainable development goals and addressing data gaps in underdeveloped and developing regions.

We highlight a novel approach that leverages daytime satellite images and machine learning, requiring minimal human annotation, to measure economic development at a granular level. This method is especially valuable in regions lacking comprehensive socioeconomic data, offering a scalable and affordable alternative to traditional economic surveys. To illustrate its effectiveness, we present our findings on North Korea, the most isolated country in the world. Our analysis sheds light on how this approach can reveal spatial inequalities and the nation's resource allocation in response to sanctions—a challenging task due to the scarcity of data in sanctioned countries.

Following the North Korean case study, we explore other extensions and applications of this approach, integrating large language models, detecting slums, and providing timely analysis during various disasters. Finally, we discuss the challenges encountered in applying this methodology to real-world policy practices and in extending this work further, outlining future directions for enhancing its effectiveness and impact.

The Unexpected Data Scientist: My Journey Into the Realm of Healthcare Big Data

Yoonjung Joo, Samsung Advanced Institute for Health Sciences & Technology (SAIHST)

A massive number of population-based databases have become available recently, providing novel research opportunities for healthcare informatics on unexplored clinical and genomic landscapes. The extensive biomedical information encoded in large-scale EHR (electronic health records) databases, ranging from diagnosis code, physician reports, brain neuroimaging data to DNA genotype data, are valuable resources for clinical researchers to characterize the pleiotropic architecture of human complex traits.

In this talk, I will share my unconventional path into the world of biomedical data science, with a particular emphasis on healthcare informatics and psychiatric genetics. As a researcher deeply engaged in the analysis of large-scale population datasets, genomic data, and electronic health records, I will discuss the transformative potential of data-driven approaches in advancing precision medicine, particularly in the context of mental health.

The presentation will explore the growth of population-based DNA biobanks and their role in accelerating healthcare research, as well as the integration of neuroimaging and behavioral data to better understand complex psychiatric disorders. By leveraging machine learning and AI techniques, we are uncovering new insights into the genetic underpinnings of mental health conditions like depression, bipolar disorder, and suicidality.

IDWSDS 2024

Session Info

S13

TECHNICAL SESSION

October 8th, 07:00 - 08:00 UTC

Research from Women Statisticians in China

CLICK HERE TO ACCESS THE ZOOM ROOM

China is home to many women statisticians who are making significant contributions to the field. Their work encompasses almost all topics in modern statistics as well as a wide range of application areas. This session highlights the work of three junior women statisticians from China, each presenting cutting-edge research on a different topic.

The first speaker will discuss transcriptome-wide association studies using nonparametric Bayesian methods to integrate multiple functional annotations, offering insights into complex genetic data.

The second presentation will focus on survey designs that address non-compliance with sensitive questions by employing the Poisson item count technique, which enhances data accuracy and participant honesty.

The third topic will explore high-dimensional matrix-variate low-rank approximation through the use of factor analysis and matrix decomposition, providing advanced solutions for handling large-scale matrix-valued data sets.

This session promises to showcase innovative research and methodologies.

ORGANIZER: Xu Zhang, South China Normal University

CHAIR: Xu Zhang, South China Normal University

Speaker Bios

DR. HAN WANG

China Agricultural University

Han Wang is an associate professor in the Department of Mathematics at China Agricultural University. Prior to her current position, she conducted postdoctoral research in the Department of Statistics and Actuarial Science at the University of Hong Kong. Her research interests include statistical genetics and Bayesian analysis. Her research work has been published in journals such as Experimental & Molecular Medicine and Bioinformatics.

DR. QIN WU

South China Normal University

Qin WU is an associate professor in the Department of Data Science at South China Normal University. She got her PHD degree from Hong Kong Baptist University. Her research interests include survey with sentistive question, compositional data analysis and multivariate zero-inflated count data. Her research work has been published in journals Statistics in Medcine, Statistical methods in medical research, etc.

DR. XU ZHANG

South China Normal University

Xu Zhang is an assistant professor in the School of Mathematics at South China Normal University. Her research interests include statistical inference for matrix/tensor-variate data objects. Her work has been published in journals such as Journal of the American Statistical Association, Statistica Sinica, Statistics in Medicine, etc.

Novel nonparametric Bayesian methods for incorporating multiple functional annotations in transcriptome-wide association studies

Han Wang, China Agricultural University

Transcriptome-wide association study (TWAS) has emerged as a powerful tool for translating the myriad variations identified by genome-wide association studies (GWAS) into regulated genes in the post-GWAS era. While integrating annotation information has been shown to enhance power, current annotation-assisted TWAS tools predominantly focus on epigenomic annotations. When including more annotations, the assumption of a positive correlation between annotation scores and SNPs' effect sizes, as adopted by current methods, often falls short. Here, we propose MAAT (multiple annotation-assisted TWAS), expanding the horizons of existing TWAS studies in two pivotal ways:
(i) We propose a non-parametric product partition model with covariates (PPMx) prior to incorporate information from seven multifaceted annotations into TWAS, getting free from the reliance on the assumption of a linear relationship between annotations and effect sizes. The included annotations also extend beyond epigenetic data.
(ii) We introduce an angle-based metric that indicates the most important annotation when a gene influences a trait, providing new perspectives in understanding the biological mechanisms. Through simulations, we demonstrate that MAAT outperforms existing state-of-the-art TWAS methods in terms of imputation R^2 and association power. Applying MAAT to eight psychiatric traits, we identify more gene-trait associations and provide both validation and interpretation of the assigned annotations.

The Poisson Item Count Technique and its non-compliance design for survey with sensitive question

Qin Wu, South China Normal University

The Poisson item count technique (PICT) is a survey method that was recently developed to elicit respondents' truthful answers to sensitive questions. It simplifies the well-known item count technique (ICT) by replacing a list of independent innocuous questions in known proportions with a single innocuous counting question. However, ICT and PICT both rely on the strong "no design effect assumption" (i.e., respondents give the same answers to the innocuous items regardless of the absence or presence of the sensitive item in the list) and "no liar" (i.e., all respondents give truthful answers) assumptions. To address the problem of self-protective behavior and provide more reliable analyses, we introduced a noncompliance parameter into the existing PICT. Based on the survey design of PICT, we considered more practical model assumptions and developed the corresponding statistical inferences. Simulation studies were conducted to evaluate the performance of our method. Finally, a real example of automobile insurance fraud was used to demonstrate our method.

Modeling and Learning on High-Dimensional Matrix-Variate Sequences

Xu Zhang, South China Normal University

We propose a new matrix factor model, named RaDFaM, which is strictly derived from the general rank decomposition and assumes a high-dimensional vector factor model structure for each basis vector. RaDFaM contributes a novel class of low-rank latent structures that trade off between signal intensity and dimension reduction from a tensor subspace perspective. Based on the intrinsic separable covariance structure of RaDFaM, for a collection of matrix-valued observations, we derive a new class of PCA variants for estimating loading matrices, and sequentially the latent factor matrices. The peak signal-to-noise ratio of RaDFaM is proved to be superior in the category of PCA-type estimators. We also establish an asymptotic theory including the consistency, convergence rates, and asymptotic distributions for components in the signal part. Numerically, we demonstrate the performance of RaDFaM in applications such as matrix reconstruction, supervised learning, and clustering, on uncorrelated and correlated data, respectively.

IDWSDS 2024

Session Info

S14

TECHNICAL SESSION

October 8th, 07:00 - 08:00 UTC

Serbian Contribution to Advanced Statistical Methods for Data Analysis: Privacy, Geometry, and Testing

CLICK HERE TO ACCESS THE ZOOM ROOM

The rapid advancement of modern technologies has led to the widespread collection and analysis of data across various domains, raising significant challenges in privacy preservation and data analysis. This session will explore three interconnected themes that are very important in contemporary data science. Firstly, it will examine differential privacy as a leading method for safeguarding individual privacy in the release of aggregated time series data. Secondly, it will delve into the complex realm of multivariate data analysis, focusing on the concept of halfspace depth and its connections with floating bodies in convex geometry. The third part of the session will address recent advancements in goodness-of-fit tests for randomly right-censored data, a common challenge in survival studies. Finally, the session will conclude with a discussion on variable selection problems and solutions, with a special focus on the non-Euclidean setting. This approach will be related to classical methods in the Euclidean setting, paving the way for the adaptation of previously discussed notions of depth measures and privacy concepts, thereby rounding out the session's theme.

ORGANIZER: Marija Cuparić, Faculty of Mathematics, University of Belgrade, Serbia

CHAIR: Bojana Milošević, Faculty of Mathematics, University of Belgrade, Serbia

Speaker Bios

KRISTINA MATOVIĆ

Vlatacom Technology

Kristina Matović completed her BSc and MSc degrees in Statistics at the Faculty of Mathematics, University of Belgrade. She has been enrolled in the PhD program at the Faculty of Mathematics since 2021. Kristina is currently employed at Vlatacom Technology in Abu Dhabi, United Arab Emirates, as a researcher specializing in the development of tracking systems and sensor fusion, encompassing all core software components for radar system control. Her areas of interest include differential privacy techniques for data privatization and the application of statistics in digital signal processing and multitarget tracking systems.

DR. PETRA LAKETA

Nordeus, Belgrade, Serbia

Petra Laketa obtained her PhD degree in Mathematics in 2020 at the Faculty of Sciences and Mathematics in Nis, Serbia, with the topic of time series analysis. After that, she worked as a Postdoctoral Researcher at Charles University in Prague, Czech Republic for over 2 years. switched to the industry and has been working as a Product Manager for the Data Governance and Data Management platform at Ataccama in Prague. Her next job was a Data Scientist role where she worked on the battery health detection for Sentinel Marine in Slovenia. Currently, she is employed at Nordeus in Belgrade, Serbia, as a Data Scientist and she is working on AB testing methodology.

DR. MARIJA CUPARIĆ

Faculty of Mathematics, University of Belgrade, Serbia

http://www.matf.bg.ac.rs/p/-marija-radicevic

Marija Cuparić is an Assistant Professor at the Faculty of Mathematics, University of Belgrade, specializing in probability and statistics. She obtained her PhD in Mathematics in 2021 at the Faculty of Mathematics, University of Belgrade, focusing on goodness-of-fit testing. She has published nine research articles in highly-ranked international journals and has participated in over 20 conferences, four of which were by invitation. She has been a researcher in two projects funded by the Ministry of Education, Science, and Technological Development of the Republic of Serbia. Currently, she is involved in a Bilateral Scientific Project between Serbia and Germany and is also participating in a COST Action as a researcher. Her main research interests include nonparametric statistics, model specification tests, asymptotic efficiency, and distribution theory, with a special focus on censored and other missing data issues.

PROF. BOJANA MILOŠEVIĆ

Faculty of Mathematics, University of Belgrade, Serbia

http://www.matf.bg.ac.rs/p/-bojana-milosevic

Bojana Milošević is an Associate Professor at the Faculty of Mathematics, University of Belgrade, where she chairs Department of Probability and Statistics and serves as Corporate Affairs Coordinator. She earned her PhD in 2016, focusing on characterization-based tests and Bahadur efficiencies. Bojana has published extensively in reputable journals and given numerous invited talks at international conferences. While her primary research is on testing statistical hypotheses, her group also explores other areas, applying concepts across various fields. She serves on the Editorial Boards of the Journal of Applied Statistics, Statistics and Probability Letters, and is the editor of Bernoulli News. Bojana leads the bilateral project "Modeling Complex Data - Selection and Specification," participates in a COST action, and has organized several international events, including EYSM 2019. She has also been involved in student-focused events where she delivers popular talks.

Differential Privacy in Time Series: Balancing Data Protection and Statistical Relevance

Kristina Matović, Vlatacom Technology, Abu Dhabi, UAE, and Faculty of Mathematics, University of Belgrade, Serbia

The advancement of modern technologies has led to the widespread use of various types of sensors for data collection, which are then used to investigate a wide range of phenomena and generate various statistics. A major challenge in such systems is to preserve individual user privacy while maintaining the relevance and accuracy of published statistical results. Differential privacy is one of the most commonly used methods for data perturbation in the release of aggregated time series data, aiming to safeguard individual privacy.
In this talk, we will present several methods for differentially privatizing aggregated time series data and explore their applications across various industries. We will delve into the strengths and weaknesses of each method, providing a comprehensive overview of their practical outcomes. The impact of different factors on the effectiveness of privacy preservation will be analyzed. Through this presentation, we aim to highlight the importance of adopting privacy-preserving measures in the era of big data and pervasive data collection.

Halfspace Depth: Intersection of Statistics and Geometry

Petra Laketa, Nordeus, Belgrade, Serbia

There is no unique definition of multivariate quantiles since it is not obvious how to generalize quantiles from a line to a multi-dimensional setting. Halfspace (or Tuley) depth is a prominent method of nonparametric analysis of multivariate data. There is a concept of floating bodies known in convex geometry that appears to have many connections with the halfspace depth, so this topic lies in the intersection of these two disciplines.

Goodness-of-Fit Testing in Survival Analysis: Imputation vs Adaptation for Censored Data

Marija Cuparić, Faculty of Mathematics, University of Belgrade, Serbia

In survival analysis, the focus is typically on the time until a certain event occurs, such as survival time after surgery or another medical treatment. Since studies are often time-limited and participants may leave the study for various reasons, the issue of randomly right-censored data becomes a significant challenge. When classical complete-case testing procedures are applied in such situations, their stability and reliability are compromised—they may or may not yield accurate results, and the factors affecting their performance are not known. Therefore, modifications are necessary: either the testing procedure itself must be adapted to account for censoring information or missing values must be imputed before applying the standard procedure. In this talk, we will present the latest advancements in both of these approaches for different tests and discuss the pros and cons of each method.

Independence testing and variable selection problems: non-Euclidean perspective

Bojana Milošević, Faculty of Mathematics, University of Belgrade, Serbia

Here we address the challenges associated with independence testing in non-Euclidean spaces, which are increasingly common in modern applications.
Traditional approaches based on Euclidean distance measures often prove inadequate for data with spherical, hyperspherical, or other non-Euclidean structures, necessitating the development of new methodologies. We consider kernel-based generalizations of distance covariance that enable efficient independence testing in such spaces. Moreover, we explore its potential in marginal screening particularly when data components are of different types. Through extensive empirical studies, we demonstrate that our proposed approaches significantly enhance performance and accuracy in comparison to conventional methods.

IDWSDS 2024

Session Info

S15

HISTORY OF INTERNATIONAL BIOMETRIC SOCIETY

October 8th, 08:00 - 08:30 UTC

The International Biometric Society: A Journey Through History and its Impact on Statistical Science

CLICK HERE TO ACCESS THE ZOOM ROOM

This invited session will delve into the rich history and significant contributions of the International Biometric Society (IBS). Established in 1947, the IBS has played a pivotal role in advancing the field of biostatistics and biometrics, fostering collaboration and innovation among statisticians, data scientists, and researchers worldwide. Join us as we explore the origins and evolution of the IBS, highlighting key milestones, influential figures, and groundbreaking conferences that have shaped the discipline. This talk will be presented by Dr. Iris Pigeot, the President of IBS.

Additionally, this session will provide an overview of the International Biometric Conference (IBC), which serves as a leading global forum for discussing the latest advancements in biostatistics. The upcoming IBC 2026, scheduled to take place in the esteemed city of Seoul, South Korea, will be introduced by Dr. Sohee Park, Co-Chair of the Local Organizing Committee for IBC 2026.

ORGANIZER: Sohee Park, Yonsei University

CHAIR: Ho Kim, Seoul National University

SPONSOR: International Biometric Society

Speaker Bios

PROF. IRIS PIGEOT

Leibniz Institute for Prevention Research and Epidemiology

Professor Iris Pigeot has been the director of the today's Leibniz Institute for Prevention Research and Epidemiology – BIPS since March 2004 and has been in charge of the Department of Biometry and Data Management of the institute since September 2001.

Having finished her studies in statistics and sociology at the University of Dortmund in 1985, she worked as a scientific assistant, earned her doctorate with a dissertation on the topic "Estimators of common odds ratios in sparse contingency tables" in 1989, and earned a professorship for statistics with a post-doctoral dissertation entitled "Multiple tests in outlier detection" in 1993.

PROF. SOHEE PARK

Yonsei University

Professor Sohee Park has been the Chair of Department of Health Informatics and Biostatistics at Yonsei University Graduate School of Public Health. She also serves as the Vice Dean of the Graduate School of Public Health. Having finished her Ph.D. in Biostatistics from the University of Southern California, she worked as a Research Associate at Department of Biostatistics, Harvard School of Public Health. She later joined as a faculty at the National Cancer Center in Korea and served as the head of Division of Cancer Prevention and was in charge of various large-scale governmental projects including Korean National Cancer Registry. Since 2012, she has been a professor of Department of Health Informatics and Biostatisics at Yonsei University Graduate School of Public Health, in Seoul, Korea. She is currently the President of International Biometric Society Korean Region, and is the Co-Chair of Local Organizing Committee for International Biometric Conference 2026.

The International Biometric Society: A Journey Through History and its Impact on Statistical Science

Iris Pigeot, Leibniz Institute for Prevention Research and Epidemiology

The International Biometric Society (IBS), founded in 1947, has played a pivotal role in shaping the landscape of biostatistics and biometrics. Originating from the growing recognition of the importance of statistics in the biological sciences, the IBS emerged through the efforts of Chester Bliss and other pioneering statisticians who sought a dedicated platform for the advancement of biometric research. The society's inaugural meeting in Woods Hole, Massachusetts, marked the beginning of an enduring legacy of innovation, collaboration, and excellence.

Over the decades, the IBS has grown from its early days of four regions to encompass a global community with 37 regions and thousands of members. Its flagship journal, Biometrics, established in 1947, and the subsequent launch of the Journal of Agricultural, Biological, and Environmental Statistics (JABES) in 1993, have cemented the society's reputation as a leading publisher of cutting-edge statistical methodologies.

As the society evolved, it has increasingly embraced diversity and inclusion, reflecting these values in its leadership, membership, and activities. The IBS has also maintained strong connections with international organizations, contributing to a broader impact on the global scientific community.

Join Us in Seoul: Shaping the Future of Biostatistics and Biometrics at IBC 2026

Sohee Park, Yonsei University

This talk will provide an overview of International Biometric Conference (IBC) 2026, to be held in Seoul, South Korea, a city renowned for its dynamic fusion of ancient traditions and modern innovation. As the heart of Korea, Seoul offers a unique blend of historical landmarks, cutting-edge technology, and vibrant cultural experiences, making it an ideal setting for a global gathering of the world's leading statisticians, biostatisticians, and data scientists. Organized by the International Biometric Society (IBS), this prestigious conference will bring together experts to share cutting-edge research and explore innovative methodologies. Against the backdrop of Seoul's rich cultural heritage and forward-thinking spirit, IBC 2026 will feature a diverse program of sessions, workshops, and plenary talks, fostering professional growth and collaboration within the global scientific community. We look forward to welcoming you to Seoul, where tradition meets the future, for this landmark event in 2026.

IDWSDS 2024

Session Info

S16

TECHNICAL SESSION

October 8th, 08:00 - 09:00 UTC

New developments in dependent censoring with unknown association

CLICK HERE TO ACCESS THE ZOOM ROOM

In survival analysis it is commonly assumed that the survival time T and censoring time C are stochastically independent. Most commonly used models and methods (like Kaplan-Meier, Cox model, AFT model, log-rank tests,...) are making use of this assumption. However, there are situations in practice where this assumption might be violated. Consider for instance the situation where some patients leave a medical study for reasons related to their health, which will then indirectly be related to their survival time. Recent research starting with Czado and Van Keilegom (2023, Biometrika) has shown that by making use of copulas to describe the relation between T and C, their association can be identified under certain conditions, which is an important step forward. In this session three talks will be given that deal with dependent censoring and that are building further on the aforementioned paper.

ORGANIZER: Ingrid Van Keilegom, KU Leuven

CHAIR: Ingrid Van Keilegom, KU Leuven

Speaker Bios

PROF. INGRID VAN KEILEGOM
KU Leuven
https://feb.kuleuven.be/ingrid.vankeilegom

Ingrid Van Keilegom is a professor of statistics at the KU Leuven, and part-time professor at UCLouvain, both in Belgium. She obtained her PhD in 1998 from Hasselt University, and worked at Penn State University, Eindhoven University of Technology and UCLouvain before starting in Leuven in 2016. Her research focuses on several aspects of survival analysis, non- and semiparametric regression, quantile regression, instrumental regression, bootstrap, and their applications. Ingrid was holder of an Advanced ERC grant (2016-2022). She is a fellow of the Institute of Mathematical Statistics (2008), a fellow of the American Statistical Association (2013), and has been co-editor of the Journal of the Royal Statistical Society - Series B (2012-2015), and associate editor of several other leading journals. She obtained a Honorary Degree from the University of A Coruña in Spain in 2022, and is an elected member of the Royal Flemish Academy of Belgium for Science and the Arts since 2021.

MYRTHE D'HAEN
Data Science Institute & ORSTAT

Myrthe D'Haen completed her Bachelor's degree in Mathematics in 2019 at Hasselt University (Belgium). Afterwards, she obtained a Master of Mathematics degree from KU Leuven (Belgium) in 2021, choosing the research option and pure mathematics profile. She then switched to the field of statistics for the PhD that she is currently pursuing. In a joint PhD project between the Data Science Institute (Hasselt University) and ORSTAT (KU Leuven), she works on copula-based models with a focus on dependent censoring, quantile regression, or their intersection.

Dependent censoring based on copulas with unknown association

Ingrid Van Keilegom, KU Leuven

In this talk we consider survival models in which the survival time and the censoring time are stochastically dependent, which is referred to as dependent censoring. The non-identifiability of a fully nonparametric dependent censoring model leads to challenging problems. A common approach to handle this dependence is based on copulas. To overcome the non-identifiability of the model, the copula can be considered fully known. This is however a heavy assumption in practice, since the strength of the dependence is rarely known. Hence, it results in estimators that can be used for sensitivity analyses but rarely for point estimation of unknown quantities. Recently, a new approach to handle dependent censoring has been proposed, in which the copula is not fully known. The marginal distributions of the survival and censoring time can be modelled parametrically, semiparametrically or even nonparametrically under certain conditions. The talk describes the literature on this second stream of copula based models.

Quantile Regression Under Dependent Censoring with Unknown Association

Myrthe D'Haen, Data Science Institute (Hasselt University) & ORSTAT (KU Leuven)

Censoring is a commonly encountered phenomenon in the study of survival data, leading to the observation of a censoring time rather than the event time of interest for some individuals in the data. Guided by the well-known identifiability issues induced, most existing literature assumes either (conditional) independence between the survival and censoring time, or dependent censoring with a known dependence strength. In the latter case, one often works with a copula model with an association parameter to be specified by the user. However, depending on the envisaged application, both these assumptions may be unrealistic. Importantly, recent work has illustrated the identifiability of some parametric copula models in which the association parameter does not need to be user-specified. We transfer their approach to a quantile context by working with an enriched asymmetric Laplace distribution for the survival time, a distribution closely connected to quantiles and at the same time both parametric and flexible. This approach enables quantile regression for survival data subject to dependent censoring with possibly unknown association strength, which is novel compared to existing quantile literature. Our theoretical results on identifiability, consistency and asymptotic normality are supported by simulation studies and a liver transplant data application.

IDWSDS 2024

Session Info

S17

TECHNICAL SESSION

October 8th, 08:00 - 09:00 UTC

Causal Inference

CLICK HERE TO ACCESS THE ZOOM ROOM

This session will cover four topics in modern causal inference from four experts in the field: causal discovery, partial identification via linear programming, testable inequality constraints implied by the causal model, and semi-parametric estimation of identified causal effects in a longitudinal time-to-event setting. Each of these topics is broad, and each talk will focus on a particular aspect of the topic. First, we will present the usefulness and reliability of causal discovery by comparing it to more "standard" statistical methods. Second, we will present a method for the derivation of testable inequalities implied by given graphical assumptions, a generalization of the classic instrumental variable inequalities. Third, we will demonstrate the use of linear programming for deriving symbolic nonparametric causal bounds in a particularly interesting setting where the difference of a single controversial assumption may make a difference in the information for partial identification. Fourth, we will present on targeted learning for recurrent events analysis. Targeted learning is a rapidly expanding area of research in causal inference, but methods in this area for time-to-event outcomes have been developed at a slower pace, with continuous time methods being relatively new.

ORGANIZER: Erin Gabriel, University of Copenhagen

CHAIR: Erin Gabriel, University of Copenhagen

IDWSDS 2024

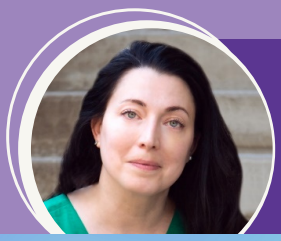
Speaker Bios

S17

DR. ANNE HELBY PETERSEN

University of Copenhagen

My research focuses on the challenges associated with causal inference using observational data. I am particularly interested in causal discovery, the science of inferring causal relationships from empirical data, and how it can be made more applicable within life course studies and epidemiology more broadly. Other topics I have an interest in include: Sibling comparison designs, supervised machine learning, missing information, register data. I also work with R development and have implemented several R packages (dataReporter, PCADSC, causalDisco, geeasy).



DR. ERIN GABRIEL

University of Copenhagen

I am currently working on methodological research in the areas of nonparametric causal bounds, designs and estimation methods for emulated and randomized clinical trials for the evaluation of prediction-based decision rules, and surrogate evaluation. My general statistical areas of interest are in causal inference and randomized trials. Although most of my previous applications have been in infectious disease and vaccination, I have recently started working in common complex diseases.

DR. MARIE BREUM

University of Copenhagen

Postdoctoral researcher working on causal inference methods. My current research focuses on mediation analysis and methods for longitudinal data including time-to-event analysis.

DR. HELENE C. W. RYTGAARD

University of Copenhagen

I work on methodological research in the areas of causal inference, targeted (machine) learning, event history analysis, and efficient nonparametric estimation. My main focus is on the development of statistical machine learning methods for estimating intervention effects in time-varying settings, and their appropriate application in medical and epidemiological studies.

Abstracts

Causal discovery: Data-driven witchcraft or a useful tool for understanding causality?

Anne Helby Petersen, University of Copenhagen

Causal discovery algorithms seek to estimate causal data-generating mechanisms, e.g. a family of directed acyclic graphs (DAGs), by analyzing empirical data. We all know that correlation does not imply causation, so are such methods violating a fundamental rule of scientific inquiry? To address this question, we will provide an introduction to the principles behind causal discovery algorithms, and their underlying assumptions. Moreover, we present recent results from a study comparing causal discovery with traditional theory-driven approaches to constructing causal DAGs in a life-course epidemiological application.

Beyond the Instrumental Inequalities

Erin Gabriel, University of Copenhagen

Instrumental inequalities were first presented in the mid-1990s for the completely binary setting, and they provided a testable set of restrictions on the observed data that could be used to falsify the assumed causal model, i.e., the classic instrumental variable setting. Soon after, it was demonstrated that although these inequalities had straightforward extensions to settings with higher dimension variables, there were additional inequalities. In this work, an algorithm was proposed for deriving the full set of inequalities in extended IV settings, but this algorithm was deemed too computationally intensive to use in practice. More recently, methods for the derivation of more general inequalities have been suggested outside instrumental variable settings, with methods for deriving the complete set once again being deemed too computationally intensive to use in practice. With the development of the R package causaloptim, this is no longer true: we can derive the complete set of inequalities in causal models that include a moderate number of variables each of which is discrete but multicategorial. After outlining the background and usefulness of such inequalities, I will demonstrate the use of an extension of causaloptim for this purpose in a simple but novel setting.

Bounds on separable direct effects in the presence of confounded intermediate variables

Marie Breum, University of Copenhagen

The separable effects framework assumes that the treatment exerts its effect on the outcome through independent mechanisms, which can be intervened upon separately. Unlike the natural direct effect, identification of the separable (interventionist) direct effect requires no cross-world assumptions. If, as is often the case, there are unmeasured confounders for the intermediate variables and the outcome neither the natural nor the separable direct effects are point identified. When a causal effect is not point-identified, one can sometimes derive bounds, i.e. a range of possible values that are consistent with the observed data. Existing work has derived nonparametric bounds on the natural direct effect in the presence of confounded intermediate variables. We extend this work to provide bounds on the separable direct effect.

Targeted learning for recurrent events analysis

Helene Charlotte Wiese Rytgaard, University of Copenhagen

Longitudinal settings involving outcome, competing risks and censoring events occurring and recurring in continuous time are common in medical research, but are often analyzed with methods that do not allow for taking post-baseline information into account. In this work, we define statistical and causal target parameters via the g-computation formula by carrying out interventions directly on the product integral representing the observed data distribution in a continuous-time counting process model framework. In recurrent events settings our target parameter identifies the expected number of recurrent events also in settings where the censoring mechanism or post-baseline treatment decisions depend on past information of post-baseline covariates such as the recurrent event process. We propose a flexible estimation procedure based on targeted maximum likelihood estimation coupled with highly adaptive lasso estimation to provide a novel approach for double robust and nonparametric inference for the considered target parameter.

IDWSDS 2024

Session Info

S18

TECHNICAL SESSION

October 8th, 08:30 - 09:00 UTC

Next Generation: Showcasing Young Portuguese Talent in Biometry and Data Science

CLICK HERE TO ACCESS THE ZOOM ROOM

This session, organized by the Biometrics Section of the Portuguese Statistical Society (SPE-SBIO), features two presentations that illustrate the significant role of statistical techniques in addressing pressing public health issues. The first presentation examines the contributions of women statisticians during the COVID-19 pandemic, focusing on innovative strategies for monitoring health behaviours and understanding vaccine hesitancy. The second presentation delves into spatiotemporal models, specifically designed for time series of counts, with valuable applications to health data analysis. These talks collectively demonstrate the critical impact of biometry in responding to today's global health challenges.

ORGANIZER: Nuno Sepúlveda, Warsaw University of Technology

CHAIR: Clara Cordeiro, University of the Algarve and CEAUL

SPONSOR: Biometry Section of the Portuguese Statistical Society

IDWSDS 2024

Speaker Bios

S18

DR. ANA MARTINS

University of Aveiro

https://orcid.org/0000-0003-4860-7795

Ana Martins is an integrated researcher member at IEETA (Institute of Electronics and Informatics Engineering of Aveiro). She finished her PhD in Applied Mathematics earlier this year at Aveiro University. She has a degree in Biochemistry, a Master's in Public Health and a Master's in Applied Mathematics. She has been selected to attend the European Young Statisticians Meetings (2023) and invited to participate in well-respected conferences like the Bernoulli-IMS 11th World Congress in Probability and Statistics (2024). Recently, she has also been selected to attend the 11th Heidelberg Laureate Forum (2024), a scientific forum to connect young researchers with laureates in the fields of mathematics and science computing.

DR. PATRÍCIA SOARES

Instituto Nacional de Saúde Doutor Ricardo Jorge

Patrícia Soares is a researcher at the National Institute of Health Dr. Ricardo Jorge since 2023. Previously, and during the pandemic, she was a researcher at the National School of Public Health. She also taught Statistics at MSc and PhD levels and advised MSc students. She has a PhD in Genetic Epidemiology (Brighton University, 2018) and an MSc in Biostatistics (University of Lisbon, 2014), focusing on survival, time series, and spatial analysis.

Spatiotemporal models for time series of counts

Ana Martins, University of Aveiro

In this talk, I will provide an introduction to time series of counts and the motivation for the development of models for these types of data. I will briefly discuss the main modelling approaches, focusing on the Integer-valued Autoregressive and Moving Average (INARMA) models. Building on this class, I will introduce a novel class of models that, in addition to the temporal component, also accounts for the spatial dimension of data – the Space-time INARMA (STINARMA) models. Finally, I will illustrate the usage of the models in practical applications, namely in health-related settings.

Women, Statistics, and the COVID-19 pandemic

Patrícia Soares, Instituto Nacional de Saúde Doutor Ricardo Jorge

The COVID-19 pandemic has underscored the crucial role of epidemiology and statistics in addressing global health challenges. Working within a team predominantly composed of women statisticians and epidemiologists, we generated evidence that informed public health responses. This presentation provides an overview of the work undertaken during and after the pandemic, demonstrating how diverse statistical tools were applied to understand and mitigate its impact.

We assessed the spatial distribution of COVID-19 notifications, identifying hotspots. We also developed a bi-weekly questionnaire to track evolving health behaviours and risk perceptions within the population, allowing for real-time monitoring of public risk perception and acceptance of implemented measures. Additionally, we evaluated vaccine hesitancy, both before the vaccine rollout and during the vaccination campaign, offering critical insights into the factors influencing vaccine uptake. Monitoring of COVID-19 vaccine effectiveness is ongoing.

Throughout our research, we used a wide range of statistical techniques, including spatial analysis, survival analysis, and time series analysis, alongside the programming skills necessary to manage vast amounts of data. Through statistics, women have made—and continue to make—significant contributions to public health.

IDWSDS 2024

Session Info

S19

TECHNICAL SESSION

October 8th, 09:00 - 10:00 UTC

Analyzing complex data in biostatistics and public health

CLICK HERE TO ACCESS THE ZOOM ROOM

In this session, three female statisticians from South Korea introduce their recent research to analyze complex data in biostatistics and public health. The first speaker presents a Bayesian spatially-clustered coefficient model with temporal structures to estimate varying risk effects across sub-regions while addressing spatial confounding bias using a two-stage framework. The model is used to analyze hepatitis A data in Korea. The second speaker presents a flexible model to estimate a dynamic trend of treatment effects on survival using the restricted mean survival time, which also incorporates propensity scores to address patient heterogeneity and takes an ensemble approach to improve estimation. The developed method is applied to the study of primary inflammatory breast cancer for assessing the effect of trimodality therapy on survival. The third speaker presents an approach to address missing values through advanced imputation techniques and bias correction using doubly robust estimators, which is shown to effectively uncover meaningful insights while controlling false positives in both single-cell and bulk-cell Alzheimer's Disease proteomic data.

ORGANIZER: Chae Young Lim, Seoul National University

CHAIR: Chae Young Lim, Seoul National University

SPONSOR: Korean Statistical Society

Speaker Bios

DR. JUNGSOON CHOI

Hanyang University

Dr. Jungsoon Choi is a professor in the Department of Mathematics at Hanyang University, South Korea. She is also working as an affiliated faculty position at the Department of Applied Statistics at the graduate school of Hanyang University. She received Ph.D. in Statistics at North Carolina State University in 2008 and worked as a postdoctoral fellow at the Medical University of South Carolina. Her research interest includes spatial statistics, spatial epidemiology, and Bayesian modeling.

DR. CHI HYUN LEE

Yonsei University

Chi Hyun Lee is an Associate Professor in the Department of Applied Statistics at Yonsei University, South Korea. Before joining Yonsei University, she worked in the Department of Biostatistics and Epidemiology at the University of Massachusetts Amherst for six years, following three years of training as a postdoctoral fellow in the Department of Biostatistics at MD Anderson Cancer Center. She earned her PhD in Biostatistics from the University of Minnesota in 2015. Her research interests include developing statistical methodologies for complex survival data and their applications in biomedical research. Chi Hyun's recent research focuses on analyzing recurrent event data and modeling survival data under biased sampling settings. She is also interested in developing statistical methods for clinically interpretable measurements and applying statistical methods to data from cancer, cardiology, and dementia research, oral health and quality of life studies, and child maltreatment.

DR. HAEUN MOON

Seoul National University

Haeun Moon is an incoming assistant professor in the Department of Transdisciplinary Innovations and the Department of Statistics at the Seoul National University, South Korea. Before joining SNU, she was a postdoctoral researcher in the Department of Statistics and Data Science at Carnegie Mellon University. She received Ph.D in Statistics from the University of Pittsburgh. Her research mainly focuses on the development of test of association for modern data, characterized by nonlinear relationships, various forms, incompleteness, or non-iid structures, and its application to variable selection. Her applied work focuses on association studies for genomic data.

Abstracts

A Bayesian Spatially-Clustered Coefficient Model With Temporal Structures for Hepatitis a Data in Korea

Jungsoon Choi, Hanyang University

Hepatitis A, a highly contagious and perilous viral liver infection, is globally widespread, with its data collected across spatial and temporal domains. Also, demographic and socioeconomic covariates, such as population density and per capita income, are gathered over space and time units. Consequently, the association between infectious disease outcomes and risk factors may differ across space and time. Some sub-regions may have a heterogeneous association with others, while a homogeneous temporal structure may exist within certain sub-regions. Acknowledging the potential variability in these associations, this study focused on comprehending the spatio-temporal dynamics of hepatitis A through a statistical model.
In this paper, we analyzed monthly hepatitis A infection counts in the Republic of Korea from January 2020 to December 2021 using a Bayesian spatio-temporal model. Specifically, we employed a Bayesian spatially-clustered coefficient model with temporal structures to estimate sub-regions with the temporally varying risk effects associated with hepatitis A. Our focus lies in utilizing the Bayesian spatio-temporal model to uncover insights into the spatio-temporally varying relationships between covariates and hepatitis A outcomes. Furthermore, we addressed the spatial confounding bias prevalent in common spatial models with spacetime random components by incorporating two-stage framework within our analysis.

Estimating Time-Varying Treatment Effects on Restricted Mean Survival Time in Large Patient Databases

Chi Hyun Lee, Yonsei University

The restricted mean survival time (RMST), which is defined as the life expectancy up to a specific time point, has recently attracted substantial attention as an alternative to the hazard ratio for quantifying the treatment effect in clinical studies. We propose a flexible model to estimate the effect of treatment based on RMST. The effect of treatment is expressed as a function of restriction time to better characterize the dynamic trend of its effect on survival. To account for possible heterogeneity across patients in large databases, we incorporate the propensity scores for receiving treatment into the model. We further introduce an ensemble approach to aggregate estimators constructed based on subsamples of the observed failure times. We evaluate the finite sample performance of the proposed single model and ensemble-based approaches through simulations, and apply the proposed methods to the study of primary inflammatory breast cancer for assessing the effect of trimodality therapy on survival.

Augmented Doubly Robust Post-Imputation Inference for Proteomic Data

Haeun Moon, Seoul National University

Quantitative measurements produced by mass spectrometry proteomics experiments offer a direct way to explore the role of proteins in molecular mechanisms. However, analysis of such data is challenging due to the large proportion of missing values. A common strategy to address this issue is to utilize an imputed dataset, which often introduces systematic bias into downstream analyses if the imputation errors are ignored. In this paper, we propose a statistical framework inspired by doubly robust estimators that offers valid and efficient inference for proteomic data. Our framework combines powerful machine learning tools, such as variational autoencoders, to augment the imputation quality with high-dimensional peptide data, and a parametric model to estimate the propensity score for debiasing imputed outcomes. Our estimator is compatible with the double machine learning framework and has provable properties. Simulation studies verify its empirical superiority over other existing procedures. In application to both single-cell proteomic data and bulk-cell Alzheimer's Disease data our method utilizes the imputed data to gain additional, meaningful discoveries and yet maintains good control of false positives.

IDWSDS 2024

Session Info

S20

TECHNICAL SESSION

October 8th, 09:00 - 10:00 UTC

Measuring stylized facts to bridge gaps, inspire innovation and shape the future

CLICK HERE TO ACCESS THE ZOOM ROOM

Stylized facts—empirical findings that hold true across various contexts—are essential for advancing research and practice in statistics and data science. This session delves into the importance of accurately measuring these facts to bridge knowledge gaps, drive cutting-edge research, and guide future developments. By focusing on consistent patterns across diverse datasets, stylized facts help researchers and practitioners connect theory with practice, leading to more effective and informed decision-making.

The session will present practical methodologies for identifying and validating stylized facts, showcase their impact through real-world examples, and discuss their potential to address global challenges. Participants will gain a deeper understanding of how to harness the power of stylized facts to inspire new ideas, foster collaboration, and shape a more equitable and innovative future in the field.

ORGANIZER: Francesca Greselin, University of Milano Bicocca

CHAIR: Laura Pagani, University of Milano Bicocca

SPONSOR: MUSA – Multilayered Urban Sustainability Action – project, funded by the European Union – NextGenerationEU, under the National Recovery and Resilience Plan (NRRP)

IDWSDS 2024

Speaker Bios

S20

ERIKA GRAMMATICA
University of Milano Bicocca

Erika Grammatica has a master in Statistical and Economic Sciences from the University of Milano-Bicocca. With the results obtained in her master's thesis, she published a scientific article proposing an approach to study the missing data present in social network data. After her Master's Degree, she worked as a Data Analyst and Consultant at a market research company in Milan, mainly in the large-scale retail trade sector. She has been collaborating with the University of Milano-Bicocca for several years as a tutor and trainer for the faculties of Statistics and Economics. She is a member of the scientific committee of the Bicocca Applied Statistics Center at Milano-Bicocca University. She is currently a research fellow at the University of Milano-Bicocca with a project titled: "Quantitative analysis of gender dynamics at the University of Milano-Bicocca". She addresses issues related to the gender gap in various contexts such as: universities, schools, small and large businesses.

PROF. MARIANGELA ZENGA
University of Milano Bicocca

Mariangela Zenga is currently an Associate professor in Social Statistics at the University of Milano-Bicocca (Italy). Her research interests are in models to study the flows of the patients in hospitals, in gender gap in higher education and in labour market studies. She collaborates with the Bicocca Applied Centre at Milano-Bicocca University and with the Centre for Statistical Science and Operational Research in the School of Mathematics and Physics at Queen's University of Belfast.

PROF. ALINA JEDRZEJCZAK
University of Lodz

Professor Alina Jędrzejczak is a distinguished Professor of Statistics at the University of Łódź, where she has established herself as a leading expert in the field.
At the University of Łódź, she is known for her innovative approach to statistical education, mentoring the next generation of statisticians and data scientists. Her research has been published in numerous high-impact journals, and she is frequently invited to speak at international conferences. Professor Jędrzejczak's dedication to her field, combined with her passion for bridging theory and practice, makes her a respected figure in the global statistics community.

Abstracts

When Do Gender Differences in Education Arise? An Analysis of INVALSI Test Scores to Explore the Gender Gap in Mathematics and Italian Performance

Erika Grammatica, University of Milano Bicocca

This research explores the presence of gender stereotypes that are already deeply ingrained in childhood and preadolescence. Using an analysis of INVALSI test scores over time, we investigate gender differences in Math and Italian performance among primary and secondary school students. Furthermore, we examine how students' socio-demographic characteristics influence these gender disparities in performance. The research aims to uncover patterns of disciplinary segregation, revealing that while girls consistently outperform boys in Italian, their performance in mathematics is lower. This disparity reflects ingrained societal stereotypes and poses potential long-term implications for academic and career choices, particularly in fields requiring strong STEM skills. Our findings highlight the need for targeted educational interventions to address these biases and promote gender equity in learning outcomes.

Gender Equality Initiatives in Italian Companies: A Study of the Life Sciences Sector

Mariangela Zenga, University of Milano Bicocca

Promoting gender equality is not just a matter of fairness; it is an essential driver of diverse perspectives that are decisive for innovation, and inclusive economic growth. This study aims to comprehensively examine gender equality measures within Italian life sciences companies, focusing on progress, challenges, and the impact of supportive policies. The main goal is to recognize the significance of empowering individuals regardless of gender, and to contribute insights for sustainable and equitable workplaces. An ad hoc questionnaire was distributed to human resources professionals across pharmaceutical, medical device, biotechnology, and nutraceutical industries in Italy. Statistical analysis, including the creation of a novel indicator considering gender quotas and policies supporting women, was conducted to provide quantitative insights into the effectiveness of gender equality initiatives. Employing a statistical indicator based on the adoption of gender quotas provides valuable insights into the inclination toward gender equality within Italian life sciences companies. Furthermore, transparency in compensation policies emerges as a critical factor in fostering gender equality initiatives. We observed that organizational size, measured by both the number of employees and turnover, proves to be a determining factor influencing companies' inclination towards gender equality. These findings underscore the multifaceted nature of gender equality efforts.

The Gender Gap in the Visegrád Group Countries Based on the Luxembourg Income Study

Alina Jedrzejczak, University of Lodz

Gender equality is a fundamental human right and one of the core values of the European Union (EU). Great efforts have been made to defend this right within the member states and around the world. However, there are still significant differences between men and women, especially in terms of income. The main objective of the paper is to compare income distributions for gender groups across four Central European countries, Poland, Slovakia, Czechia and Hungary, i.e., the members of the Visegrád Group (V4). These countries share similar histories and similar economic development, but there are substantial differences between their approaches to economic reforms, including labour market policy. This, in turn, is reflected in different income distributions and income inequality patterns. There is a debated research issue regarding the methodology of measuring the gender gap - the traditional methods based on comparing means and medians seem unsatisfactory as they do not consider the shape of income distributions. The paper's novelty lies in the application of the relative distribution concept, which goes beyond the typical focus on average income differences toward a full comparison of the entire distribution of women's earnings relative to men's The basis for the calculations was the microdata from the Luxembourg Income Study (LIS). The statistical methods applied in the study were appropriate to describe the gender gap over the entire income range.

IDWSDS 2024

Session Info

S21

TECHNICAL SESSION

October 8th, 09:00 - 09:30 UTC

Round Table: Type 2 Diabetes Mellitus (T2DM) Through the Gender Lens: A Case Study of Ghanaian Community

CLICK HERE TO ACCESS THE ZOOM ROOM

Discussion: Various factors influence a person's vulnerability to diabetes. Some of the factors that influence diabetes are gender, age, hypertension, etc. This dialogue wants to deepen perspective on Gender and share the highs and lows that can be viewed to enhance preventive measures. It is time we take care of ourselves making use of facts (data), especially in comparing if this holds for women from all walks of life.

Bring your beautiful-brainy self and let us compare notes and delve into how we can help ourselves and change the future narratives positively with best practices towards a diabetic-free world.

ORGANIZER: Irene Kafui Vorsah Amponsah, Department of Statistics, University of Cape Coast Ghana

CHAIR: Irene Kafui Vorsah Amponsah, Department of Statistics, University of Cape Coast Ghana

IDWSDS 2024

Speaker Bios

S21

GLADYS BAYELDENG

Ghana Education Service

Gladys Bayeldeng is a Ghanaian pursuing a Master of Philosophy in Statistics at the University of Cape Coast. She also holds a Bachelor of Science in Statistics from the University of Cape Coast, a Diploma in Education from the University of Education, Winneba, and a Higher National Diploma in Statistics from Tamale Technical University in Tamale. She has over fifteen years of teaching experience in teaching mathematics and encouraging girls to develop an interest in the subject. She participated twice in the Help Teachers Teach Mathematics (HTTM) global conference which aimed to find innovative ways of teaching Mathematics and using ICT tools to make the subject more practical and participatory in the classroom. She also worked with the Ghana Statistical Service as a District Trainer and Field Supervisor in data collection procedures for the 2020 Population and Housing Census in Ghana.

DR. IRENE K. V. AMPONSAH

Department of Statistics, University of Cape Coast Ghana

https://directory.ucc.edu.gh/p/irene-kafui-vorsah-amponsah

Dr. Irene Kafui Vorsah Amponsah is a Senior Lecturer with over 15 years of experience (BSc, MPhil, and PhD in Statistics), specializing in computational statistics, data analysis, model selection, and recovery rate. She has published extensively on price models, asymmetry, STEM education, health, and queuing theory. Her passion for statistics is driven by its practical applications and career prospects in academia, industry, government, and medical fields. The 1st president of WIStats Ghana, she has served as a gender and statistics expert on the 2023 Ghana HDR, UN-GWI SDGs Ambassador, and held various academic roles. Dr. Kafui is also a trained GUNSA peer educator, youth advocate, and mentor, guiding students and teaching assistants in Ghana and abroad. A mother of three, she enjoys singing, counselling, taking initiative, spending time with family, reading novels, and exercising. Her inspiration is fuelled by offering support and uplifted by the smiles of her mentees.

IDWSDS 2024

Session Info

S22A

TECHNICAL SESSION

October 8th, 09:30 - 09:45 UTC

Test for Symmetry and Confidence Interval of the Parameter μ of Skew-Symmetric Laplace Uniform Distribution

CLICK HERE TO ACCESS THE ZOOM ROOM

The skew symmetric Laplace uniform distribution SSLUD(μ) is introduced in Lohot and Dixit (2024) using the skewing mechanism of Azzalini (1985). Here we derive the most powerful (MP) test for symmetry of the SSLUD(μ). Since the form of the test statistic is not analytically tractable and it is difficult to obtain its exact distribution, critical values and the power of MP test are obtained using simulation. Further a 100(1-α)% confidence interval (CI) for parameter μ assuming asymptotic normality and empirical distribution of the maximum likelihood estimator of μ.These two methods are compared based on the average length and coverage probability of the CI. Finally the CI of the parameter μ is constructed using data on the "transformed daily percentage change in the price of NIFTY 50, an Indian stock market index" given in Lohot and Dixit(2024).

Key words : Confidence interval, maximum likelihood estimation, most powerful test,simulation, skew-symmetric Laplace-uniform distribution, test for symmetry
AMS classification: 62F03, 62F25,65C10

ORGANIZER: Vaijayanti Dixit, Mumbai University, Mumbai, India

CHAIR: Vaijayanti Dixit, Mumbai University, Mumbai, India

IDWSDS 2024

Speaker Bios

S22A

DR. VAIJAYANTI DIXIT

Mumbai university, Mumbai, India

My name is Dr.(Mrs) Vaijayanti Ulhas Dixit. I am working as an Associate professor in the department of statistics, Mumbai university, Mumbai, India, since 2008. I have 26 years of experience of teaching for post graduation students that is master of science in statistics.
My research areas of interest are theory of estimation, testing of hypotheses, statistical inference, distribution theory, I have published 17 research papers in national and international referred journals.
Two students have completed Ph.D. under my guidance and four are currently working for Ph.D. with me.
My son Anand Dixit has completed M.Sc (statistics) from Mumbai university and Ph. D. from Iowa state university, USA.
My daughter Vaidehi Dixit has completed her M.Sc. (statistics) from Mumbai university and Ph. D. from North Carolina state university, USA.

IDWSDS 2024

Session Info

S22B

TECHNICAL SESSION

October 8th, 09:45 - 10:00 UTC

Evaluating Randomness Assumption: A Novel Graph Theoretic Approach for Linear and Circular Data

CLICK HERE TO ACCESS THE ZOOM ROOM

Randomness or mutual independence is an important underlying assumption for most widely used statistical methods in both linear and circular contexts. However, not many tests are available to check for randomness, particularly for circular data. In this paper, we introduce a new approach for developing non-parametric tests for linear and circular data. We introduce a new concept of Random Circular Arc Graphs (RCAG) for circular data analogous to that of Random Interval Graphs (RIG) for linear data. We examine various properties of the RCAGs, including edge probability, vertex degree distribution, maximum and minimum degrees, and the presence of Hamiltonian cycles. Then, we use them to create randomness tests for circular data. Similar ideas lead to new tests of randomness for linear data. For linear data, we demonstrate that our test outperforms most of the standard parametric and non-parametric tests available in the literature, including the Runs test. Similarly, we substantiate the effectiveness of our tests for circular data through extensive simulations.

ORGANIZER: Shriya Gehlot, Indian Institute of Management Ahmedabad

CHAIR: Shriya Gehlot, Indian Institute of Management Ahmedabad

IDWSDS 2024

Speaker Bios

S22B

SHRIYA GEHLOT

Indian Institute of Management Ahmedabad

https://sites.google.com/view/shriya-gehlot/home

Shriya Gehlot is a PhD student in Operations and Decision Sciences at Indian Institute of Management Ahmedabad India. She has a BS-MS in Mathematics from the Indian Institute of Science Education and Research Bhopal. Her research focuses on Directional Statistics and Random Graphs. Her work on classification with ordinal circular data is accepted to be published in the book Directional and Multivariate Statistics. She has also presented her thesis work at conferences like International Symposium on Non-Parametric Statistics 2024 and Statistical Methods in Finance (StatFin) 2023. She received the best paper award for her work "A Graph Theoretic Test for Independence of Stock Returns" at StatFin 2023 conference. She has been honoured with Chaudhary-Padmanabhan-Pant Award for scholastic performance in the first year of her PhD. She is also a recipient of prestigious S.N. Bose and DST-INSPIRE Scholarships. She has also taught basic mathematics courses to MBA students at IIMA in 2023, 2024.

IDWSDS 2024

Session Info

S23

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

October 8th, 10:00 - 11:00 UTC

Women in Statistics and Data Science: Overcoming Career Hurdles and Leveraging Opportunities in the African context

CLICK HERE TO ACCESS THE ZOOM ROOM

Abstract

This session will feature four accomplished women statisticians and one student discussing the obstacles they navigated in their careers and the strategies they utilized to excel in the field of statistics in African settings.

Rationale

Women continue to face challenges in advancing in the statistical and newly emerging data science field due to various factors compared to their male counterparts. This session will highlight the experiences of successful women statisticians, shedding light on the barriers they encountered and the tactics they employed to progress in their careers in the African context. Including a student in the panel will give an angle from the younger generation we are trying to support in their advancement, we can hear about their current challenges.

ORGANIZER: Cherlynn Dumbura, CeSHHAR Zimbabwe, Place Alert Labs, MSU

CHAIR: Cherlynn Dumbura, CeSHHAR Zimbabwe, Place Alert Labs, MSU

IDWSDS 2024

Speaker Bios

S23

BETTY MAWIRE

Kutsaga Research Center

Betty Mawire is Biometrician/Data analyst working at a Research Institute in Zimbabwe. She is a seasoned statistician with 20 years of experience in designing studies, data analysis including modelling, results interpretation, data visualization and reporting findings. Skilled in leveraging statistical methodologies to derive actionable insights, she is committed to delivering accurate and meaningful results that drive decision-making. She provides statistical consultancy services to Researchers as well as peer review scientific publications. She has been in the leadership for the International Biometric Society representing the Zimbabwean region and presented her work on different international platforms.

CAROLINE MUGO

JKUT

Caroline is currently a lecturer in the Department of Statistics and Actuarial Sciences, School of Mathematics and Physical Sciences (JKUAT). She holds a master's degree in applied statistics and another in Biostatistics from the University of Hasselt, Belgium and is currently undertaking her doctoral research on Hierarchical Models for Infectious Disease Dynamics. Caroline's responsibilities include teaching, research, and consultancies, mainly on the areas of statistics, biostatistics, and data science. She participates in various research studies providing statistical support in study design and data analysis. She has taken part in capacity building in several research institutes and government institutions in a bid to ensure quality use of data and statistics. She has served as a trainer for Statistics Softwares and others including R, SPSS, STATA among others.

MAKOMBORERO NYONI

National University of Science and Technology

Makomborero Nyoni is a passionate statistics student currently pursuing her bachelor's degree in Operations Research and Statistics at National University of Science and Technology. With a keen interest in data analysis and statistical modelling. She is actively involved in the university's Statistics Club, where they collaborate with peers on research projects and participates in data competitions and she is part of a consultancy which conduct part time project evaluation.

SHUVAI MASVIMBO

TelOne Technical Center

Shuvai Bridget Masvimbo is an educator and a qualified Statistician and Data Scientist. She has years experience as an educator and is an advocate of making quality education available to minorities. with an Honours Degree in Operation Research and Statistics and a Masters in Big Data Analytics, she pursues her passions as a Lecturer at TelOne Centre for Learning and as the CEO of Ocean Crest Institute an education facility.

IDWSDS 2024

Session Info

S24

TECHNICAL SESSION

October 8th, 10:00 - 11:00 UTC

Methodologies in Time Series and Spatial Statistics With Applications

CLICK HERE TO ACCESS THE ZOOM ROOM

We have three talks in our session, relating to nonstationary and nonlinear modeling and inference for time series and spatial data. The first talk introduces a structured state-space diffusion model in the neural network framework with uncertainty quantification (presented by Yuting Fan, PhD candidate at the Institute of Statistics, Yang Ming Chiao Tung University). Empirical study shows the advantages of this approach in enhancing short-term forecasts for load time series data in power systems. The second talk introduces a covariate-dependent spatial-temporal covariance model with applications to environmental data (presented by Yen-Shiu Chin, Postdoc at the Institute of Statistics Science, Academia Sinica). This novel approach provides a way of exploring the impact of climate covariates on spatial covariance for better spatial predictions. The third talk introduces an integrated SVM-ARMA-GARCH model to forecast and detect structural changes for nonlinear heteroscedastic time series (presented by Meihui Guo, Professor at the Department of Applied Mathematics, National Sun Yat-sen University).

ORGANIZER: Nan-Jung Hsu, Institute of Statistics, National Tsing-Hua University, Taiwan

CHAIR: Nan-Jung Hsu, Institute of Statistics, National Tsing-Hua University, Taiwan

SPONSOR: The Chinese Institute of Probability and Statistics, Taiwan

IDWSDS 2024

Speaker Bios

S24

YUTING FAN

Institute of Statistics, Yang Ming Chiao Tung University

Yuting Fan is a PhD candidate at the Institute of Statistics at Yang Ming Chiao Tung University in Taiwan. During her PhD studies, she focused on spatial data analysis and developed a spatial statistical approach to analyze complex spatial data. She then further investigated the use of generative models, especially diffusion models, in deep learning to analyze time series data and spatio-temporal data.

DR. YEN-SHIU CHIN

Institute of Statistics Science, Academia Sinica

Yen-Shiu Chin is a postdoctoral researcher at the Institute of Statistical Science, Academia Sinica, Taiwan. She received her PhD in Statistics from National Tsing Hua University, Taiwan, in 2024. Her research interests include spatial-temporal statistics, statistics of extremes, and variable selection.

PROF. MEIHUI GUO

Dept. of Applied Math., National Sun Yat-sen University

Meihui Guo was awarded the B.A. in mathematics from National Tsing Hua University (Hsinchu City, Taiwan) in 1983, and the Ph.D. in statistics from the University of Maryland (College Park, Maryland, USA) in 1989. She was an Assistant Professor in the Department of Mathematical Sciences, Worcester Polytechnic Institute, (Worcester, Massachusetts, USA) from 1989–1992. She has been on the faculty of the Department of Applied Mathematics, National Sun Yat-sen University (Kaohsiung, Taiwan) since 1992, and is now a Professor. Her research interests are time series, estimation theory, data science, and machine learning.

Abstracts

Enhancing Short-Term Forecasting in Power Systems with Generative Models

Yuting Fan, Institute of Statistics, Yang Ming Chiao Tung University in Taiwan

The power system is a critical backbone for economic stability, which makes precise short-term forecasting indispensable in the field. Recent advancements in diffusion models have set new benchmarks in deep learning, particularly for time-series and spatio-temporal analysis. This study explores the application of Structured State Space Diffusion (SSSD) models to load data obtained from the New York Independent System Operator (NYISO). By incorporating structured state-space models into the neural network architecture of conditional diffusion models, SSSD effectively captures time dependencies and can be applied to spatio-temporal data. This approach demonstrates a promising avenue for improving the accuracy of day-ahead forecasts in power systems. Moreover, this study also considers uncertainty quantification by applying conformal prediction to adjust the prediction interval.

Covariate-Dependent Spatio-Temporal Covariance Models With Applications to Environmental Data

Yen-Shiu Chin, Institute of Statistics Science, Academia Sinica

Many meteorological variables are known to play a crucial role in affecting environmental and geophysical processes. However, in the literature, these covariates are mainly used to model the mean structure rather than the spatio-temporal covariance structure. We propose a novel covariate-dependent covariance model built on the spatio-temporal random-effects model framework. The covariates are incorporated into the spatial covariance function via a Cholesky-type decomposition to ensure the positive-definite property. We apply the maximum likelihood for parameter estimation, computed via an expectation conditional maximization algorithm. Simulation studies and real data applications demonstrate that the proposed method outperforms stationary approaches without considering the impact of covariates in the spatio-temporal covariances.

Forecasting and Change Point Test for Nonlinear Heteroscedastic Time Series Based on Support Vector Regression

Meihui Guo, Department of Applied Mathematics, National Sun Yat-sen University

SVR-ARMA-GARCH models provide flexible model fitting and good predictive powers for nonlinear heteroscedastic time series datasets. In this study, we explore the change point detection problem in the SVR-ARMA-GARCH model using the residual-based CUSUM test. For this task, we propose an alternating recursive estimation (ARE) method to improve the estimation accuracy of residuals. Moreover, we suggest using a new testing method with a time-varying control limit that significantly improves the detection power of the CUSUM test. Our numerical analysis exhibits the merits of the proposed methods in SVR-ARMA-GARCH models. A real data example is also conducted using BDI data for illustration, which also confirms the validity of our methods.

IDWSDS 2024

Session Info

S25

TECHNICAL SESSION

October 8th, 10:00 - 10:30 UTC

Data Science for Health Equity

CLICK HERE TO ACCESS THE ZOOM ROOM

This session delves into the pivotal role of data science in advancing health equity, emphasizing the potential of data science to address and mitigate disparities in healthcare outcomes. Health equity requires that all individuals, regardless of socioeconomic status, race, or geographic location, have access to the highest quality of care. However, traditional healthcare systems and algorithms often reflect and perpetuate existing biases, leading to unequal outcomes. By integrating rigorous statistical approaches, we can uncover and correct these biases, ensuring that health innovations benefit everyone fairly. Statistical methods offer the tools needed to analyze and understand the complex factors contributing to health disparities, from data collection practices to algorithm design and implementation. This session will explore how these methods can be applied to identify inequities, assess their impact, and develop strategies to promote fairness in healthcare delivery. Hearing from academic experts and representatives of a volunteer organization, attendees will gain a deeper understanding of the challenges and opportunities in making healthcare more just and inclusive through Data Science.

ORGANIZER: Leandra Braeuninger, University College London, Alan Turing Institute
CHAIR: Leandra Braeuninger, University College London, Alan Turing Institute
SPONSOR: Data Science for Health Equity (DSxHE)

Speaker Bios

LEANDRA BRAEUNINGER

University College London, Alan Turing Institute

https://leandrabraeuninger.github.io/

Leandra Bräuninger is a doctoral student developing statistical methods to mitigate social bias in personalised medicine at the University College London supervised by Dr Brieuc Lehmann. Their research interests include interdisciplinary approaches to algorithmic fairness, genomic fairness, uncertainty quantification, counterfactual frameworks and causal inference to name a few.

Previously, Leandra held research and/or teaching positions in the areas of statistical fairness in genomics, mathematical malaria prediction, and fundamental biology at the Alan Turing Institute (London, UK), the Mathematical Biology Group at the University of Melbourne (Australia), and the Centre for Active Learning in the Department of Biology at ETH Zurich (Switzerland).

DR. ELINOR LAW

University of Birmingham

https://www.birmingham.ac.uk/staff/profiles/inflammation-ageing/laws-elinor

Dr Elinor Laws is a Public Health doctor and academic with the AI and Digital Health Research and Policy Group at the University of Birmingham.
The AI and Digital Health group conduct research that seeks to ensure AI technologies are safe, effective and equitable. The group works in collaboration with academic, industry and policy institutions around the world, bringing diverse and interdisciplinary teams together to build best practices that can be translated internationally. Dr Laws co-led the STANDING Together project to release recommendations to encourage transparent reporting of health data.
Dr Laws is interested in understanding how we can achieve digital health equity, looking to academic disciplines such as gender studies, politics and modern languages to see how they inform our practice as healthcare scientists.

CLAIRE COFFEY

University of Cambridge

https://www.hdruk.ac.uk/people/claire-coffey/

Claire Coffey is a PhD Candidate and Researcher in Health Data Science at the University of Cambridge, supported by Health Data Research UK, The Alan Turing Institute, and the Wellcome Trust. Her research centres on the fairness and equity of medical AI and predictive algorithms in healthcare. Previously, she was a DeepMind Scholar in MPhil Advanced Computer Science at the University of Cambridge. She received her BSc in Computer Science from the University of Birmingham, with placements at the University of British Columbia and the University of Waterloo. She has industrial experience in Research and Development, building AI software for autonomous vehicles; work for which she has multiple patents.

Revealed: Gaps in the World Map of Available Healthcare Datasets (STANDING Together)

Elinor Law, University of Birmingham

In this talk, Elinor Law will present the STANDING together project and its recommendations.
Artificial intelligence (AI) health technologies have the potential to transform healthcare help address unmet health needs worldwide. Concerningly however, a growing corpus of literature demonstrates the ability of these tools to cause or contribute to health inequity. There is growing evidence of a disconnect between those who are represented in health data and those who have the greatest unmet need.
We have built recommendations encouraging transparency around who is represented in datasets (and who has been left out), how they are represented, and how data is used when developing AI technologies for healthcare. These recommendations are the product of a multi-stakeholder international consensus process involving representatives from 58 countries.
One of the priorities of the STANDING together project is to ensure the diverse voices of minoritised populations are heard. We have woven Patient and Public Involvement and Engagement (PPIE) input throughout our research, from commissioning to dissemination.

Heartfelt Algorithms: Exploring Equity in Cardiovascular Disease Risk Prediction Through Algorithmic Fairness

Claire Coffey, University of Cambridge

Accurate and equitable cardiovascular disease (CVD) risk prediction is crucial for effective screening, diagnosis, and treatment decisions. However, many clinical risk prediction models (CPMs) fall short, particularly for minority and intersectional groups, due to unrepresentative training data and underlying societal inequalities. In this talk, I will explore the algorithmic fairness of widely used CPMs, including the Pooled Cohort Equations and QRISK2 & 3, investigating their performance for different population subgroups. Using UK Biobank data, I evaluate the algorithmic group fairness of these models across protected characteristics (sex, age, ethnicity, deprivation, and their intersections). My findings reveal widespread calibration issues and high levels of unfairness, especially for intersectional and minority groups. Notably, while sex- and age-based recalibration improves calibration in some cases, it is insufficient for addressing disparities in ethnicity groups. I propose ethnicity-based recalibration, which shows promise in reducing unfairness and improving calibration. Yet, no approach can address all fairness metric differences as each metric gives insight into different errors. This work illustrates the importance of investigating CPMs beyond population-level metrics, to uncover and address hidden disparities. My findings aim to guide the development of more equitable models that do not leave historically disadvantaged groups behind.

IDWSDS 2024

Session Info

S26

TECHNICAL SESSION

October 8th, 10:30 - 11:00 UTC

Modeling the Impact: Non-Pharmaceutical Interventions (NPIs) and COVID-19 Transmission

CLICK HERE TO ACCESS THE ZOOM ROOM

In the early stage of the COVID-19 pandemic, governments globally enacted various non-pharmaceutical interventions (NPIs) to curb the spread of the virus. This research scrutinized the impact of these measures within the United States throughout the initial surge of cases, utilizing three distinct analytical approaches. The prototypical Bayesian hierarchical model is employed to measure the impact of five NPIs: gathering restrictions, restaurant capacity restrictions, business closures, school closure, and stay-at-home orders — in 42 states that reported over 100 fatalities by the end of the study period. Additionally, the impact of mask-wearing mandates, as the 6th NPI, was evaluated through counterfactual modeling. This specialized version of the Bayesian hierarchical model was designed to explore hypothetical scenarios, estimating the outcomes if states had either implemented or not implemented such a mandate. The investigation was completed with an advanced Bayesian hierarchical model that assessed the effectiveness of all six NPIs across the nation, encompassing all 50 states and the District of Columbia. The findings of this study affirm previous conclusions about the overall efficacy of NPIs in mitigating the proliferation of the virus from a Bayesian perspective, underscoring the significance of statistical analysis and data science in providing data-driven and evidence-based public health insights.

ORGANIZER: Yuhang Liu, Moderna, Inc.

CHAIR: Yuhang Liu, Moderna, Inc.

Speaker Bios

DR. YUHANG LIU
Moderna, Inc.
https://www.linkedin.com/in/yuhang-liu-67426b90/

Dr. Yuhang Liu completed her Master's in Applied Mathematics and Statistics at Stony Brook University in 2020. Subsequently, she became a part of Dr. Wei Zhu's research team at the Center of Excellence Wireless and Information Technology (CEWIT) at Stony Brook University and earned her Ph.D. in 2022. With a solid foundation in statistics, Dr. Liu is deeply involved in clinical research and public health initiatives. Her expertise in statistical analysis and machine learning is central to her research efforts, where she focuses on deriving data-driven insights that enhance clinical studies, public health improvement, and the management of disease. After receiving her doctorate, Dr. Liu joined Moderna, Inc., where she currently works as a Biostatistician.

IDWSDS 2024

Session Info

S27

TECHNICAL SESSION

October 8th, 11:00 - 11:30 UTC

Predicting Recurrent Events in a Survival Framework: Development of a Machine Learning Approach and an Application in Oncology

CLICK HERE TO ACCESS THE ZOOM ROOM

In medical research, individuals often encounter multiple instances of the same event over time, such as repeated hospitalizations or cancer relapses. While survival analysis traditionally models the time to the first event, this approach does not fully capture the complexity inherent in recurrent events. To address this, several statistical models have been developed, yet a consensus on learning approaches for high-dimensional data remains elusive. A systematic literature review was conducted to synthesize state-of-the-art methodologies and compare existing methods, revealing a significant gap in modeling recurrent events with machine learning techniques. In response, this research project introduces an enhanced version of the Random Survival Forest algorithm, specifically designed for recurrent event data. This extension, RecForest, leverages survival analysis principles and ensemble learning. Furthermore, this thesis addresses the critical need for interpretability and explainability in algorithms, particularly when integrated into medical devices. Recommendations are provided to ensure compliance with health authority standards, underscoring the necessity for transparent and accountable AI-based medical devices. These guidelines aim to enhance the evaluation of algorithmic performance, fostering trust and reliability in medical decision-making tools.

ORGANIZER: Juliette Murris, Inria, Pierre Fabre R&D

CHAIR: Juliette Murris, Inria, Pierre Fabre R&D

Speaker Bios

JULIETTE MURRIS

Inria, Pierre Fabre R&D

Juliette Murris is a PhD candidate specializing in biomathematics and biostatistics amongst the HeKA team (Inria – Inserm initiative). Her research focuses on the prediction of multiple clinical events in oncology. She also has a strong interest in any innovative technologies, specifically when related to machine learning/artificial intelligence issues. In addition to her doctoral studies, Juliette works as a biostatistician in the R&D department of a pharmaceutical company named Pierre Fabre.

IDWSDS 2024

Session Info

S28

TECHNICAL SESSION

October 8th, 11:00 - 12:00 UTC

SEIO Women in Statistics and Data Science: A Research Sample From Different Perspectives and Career Stages

CLICK HERE TO ACCESS THE ZOOM ROOM

This session, organised by the SEIO Women Comission of the Spanish Society of Statistics and Operations Research (SEIO), presents 3 talks of relevant women researchers who are at different career stages:

- Lola Ruiz Medina, full professor
- Vanesa Guerrero, associate professor
- María Jaenada, assistant professor

The topics are focused on functional time series analysis, optimisation applied to data science for interpretable models, and robust estimation in reliability with censored data.

ORGANIZER: Begoña Vitoriano, Spanish Society of Statistics and Operations Research (SEIO)

CHAIR: Eva Vallada, Technical University of Valencia

SPONSOR: Spanish Society of Statistics and Operations Research (SEIO)

IDWSDS 2024

Speaker Bios

S28

PROF. M. DOLORES RUIZ MEDINA

University of Granada

Full professor since 2006 in the Department of Statistics and Operational Research and researcher at the Institute of Mathematics (IMAG) of the University of Granada. She holds a bachelor's degree in Mathematics (1990) and obtained a PhD in Mathematics (2003) from the University of Granada. She collaborates with researchers all around the world, being a member of Editorial Boards of international journals, as Spatial Statistics and TEST.
Her research interests are random fields, functional time series and manifolds, leading research projects since 2003 with more than 100 papers published in prestigious journals. She supervised more than 10 PhD theses. Currently, she also leads the Thematic Network on Stochastic Processes and Applications, and the Granada University and the SEIO research groups on the topic. She is a member of the Academy of Mathematical, Physicochemical and Natural Sciences in Granada. She was Vicepresident of SEIO (2019-22) and coordinates the SEIO Women Commission.

https://researchportal.uc3m.es/display/inv45738

DR. VANESA GUERRERO

Carlos III University of Madrid

Associate Professor since 2024 in the Department of Statistics at the Carlos III University of Madrid. She holds a degree in Mathematics (2012), a master's degree in Advanced Mathematics (2013), and obtained a PhD in Mathematics (with an Extraordinary Doctorate Award) from the University of Seville in 2017. She has carried out research stays at the Copenhagen Business School (Denmark) and at the École Polytechnique (France).
Her research focuses on the use of mathematical optimization tools for problems related to the analysis of complex data. She has received the 2018 Vicent Caselles Award from the Royal Spanish Mathematical Society and the BBVA Foundation, the 2018 Ramiro Melendreras Award from the Spanish Society of Statistics and Operations Research (SEIO), and the L'Óreal-UNESCO "For Women in Science" award in 2024. She participates in and leads research projects, and collaborates in mathematics outreach activities to encourage scientific vocations among young people.

https://produccioncientifica.ucm.es/investigadores/278458/detalle

DR. MARÍA JEANADA

Complutense University of Madrid

Assistant Professor in the Department of Statistics and Operational Research and researcher of the Interdisciplinary Mathematics Institute, at Complutense University of Madrid. She holds a degree in Mathematics (2019), a degree in Mathematics and Statistics (2019) and a master's degree in Computational Statistics (2020) from Complutense University of Madrid. She obtained a PhD in Mathematical Engineering, Statistics and OR from Complutense University of Madrid and Technical University of Madrid in 2024. She has carried out research stays at the McMaster University (Canada) and for teaching in the Universidade Pedagógica de Maputo (Mozambique).
Her research interests include information theory, generalized regression models, high dimensional data, reliability analysis and robust statistics. She is co-author of several research articles and has presented her work at national and international conferences. She participates in several research and cooperation for development projects.

Non-Euclidean Functional Time Series Analysis

M. Dolores Ruiz Medina, University of Granada

Manifolds, in particular spherical functional time series analysis, help in understanding the dynamics of spatial patterns of data that are inherently spherical, or embedded into a manifold, providing valuable insights for prediction, monitoring, and decision-making. Spherical functional time series analysis is applied in several fields where data are collected over time on a spherical domain (respectively, on a general manifold). These applications include the statistical analysis of climate variables over the Earth's surface to study climate changes in Climate Sciences and Meteorology; the analysis and prediction of marine variables over time in Oceanography; the detection of changes in the Earth's magnetic field over time in Astrophysics; the brain activity patterns analysis over time in Neuroimaging; the geographical analysis of global economic indicators such as GDP growth rates or unemployment rates across countries in Economics, among others. These applied fields leverage manifold, and, in particular, spherical functional time series analysis for their ability to represent complex functional data sets in non-euclidean domains, providing crucial insights, and precision improvements in prediction based on regression, machine learning, and computer sciences. The present talk reviews some recent advances in the analysis of non-euclidean functional data correlated in time. The methodological approaches are illustrated in terms of simulations and real-data applications.

Enhancing Interpretability in Additive Models via Mathematical Optimization

Vanesa Guerrero, Carlos III University of Madrid

In an era when the decision-making process is often based on the analysis of complex and evolving data, it is crucial to have systems which allow us to incorporate human knowledge and provide valuable support to the decider. During this talk, statistical modelling and mathematical optimization paradigms merge to address the problems of, first, estimating smooth curves and hypersurfaces which verify structural properties (e.g. about sign, monotonicity or curvature), and second, perform feature selection in additive models. In both cases, we assume that the smooth functions to be estimated are defined through a reduced-rank basis (B−splines) and fitted via a penalized splines approach (P−splines). Conic Optimization and Mixed Integer Quadratic Programming are used to address these problems. The proposed methodologies are tested in both simulated and real datasets, and they are shown to be competitive against other approaches in the literature.

Robust Estimation for Interval-Censored Reliability Experiments

María Jeanada, Complutense University of Madrid

Handling censored data in reliability analyses is a key concern in practice. Interval-censored data emerges in experiments where failure times are only known to fall within a specific interval rather than being observed precisely. Additionally, many modern devices are extremely reliable with the large lifetimes to failure, necessitating long experimental durations for inference under normal operating conditions. Instead, accelerated life tests (ALTs) are used to reduce the lifetime of devices by increasing one or more stress factors, which induces failure. After suitable inference, results can be extrapolated to working conditions. Classical estimation methods relying on the likelihood function of the lifetime distribution can be significantly influenced by data contamination. As an alternative, robust estimators based on distance measures are developed.

IDWSDS 2024

Session Info

S29

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

October 8th, 11:00 - 11:30 UTC

Working as a Statistician in the Healthcare Industry

CLICK HERE TO ACCESS THE ZOOM ROOM

In the dynamic landscape of modern industry, statisticians play a pivotal role in transforming data into actionable insights and driving strategic decisions. This panel brings together seasoned professionals from different countries to share their experiences and provide valuable perspectives on the multifaceted role of statisticians in industry settings. Attendees will gain insights into the evolving demands of the healthcare industry and the essential skills needed to thrive in this field. Through interactive discussions, this session aims to equip current and aspiring statisticians with a deeper understanding of the industry landscape and practical strategies for success.

ORGANIZER: Umut Ozbek, Eli Lilly and Company

CHAIR: Hélène Sapin, Eli Lilly and Company

Speaker Bios

ZHIHONG CAI

Eli Lilly and Company

Zhihong Cai received her PhD in biostatistics at Kyoto University, and joined Eli Lilly Japan in 2008. She worked on phase 1-3 clinical trials and new drug application (NDA) submission for 10 years, and then transitioned to real world analytics and works on post launch activities including Health technology assessment (HTA), real world evidence (RWE) research, post-market safety studies, medical affairs publications across all therapeutical areas.

HÉLÈNE SAPIN

Eli Lilly and Company

I am based in France. I joined Lilly 17 years ago after working for about 10 years in a CRO.
Since joining Lilly, I have worked primarily in Diabetes, for post regulatory support : Medical affairs, Real World Evidence, Health Technology Assessments around the globe. I have supported compounds in diabetes area, but also insulins and connected care. This vast experience is helpful in my daily work as I can adapt to diverse needs and asks from partners.

YING LOU

Eli Lilly and Company

Ying Lou received her Ph.D. degree in Statistics in 2014 from Southern Methodist University. While pursuing her Ph.D. degree, she worked as a research assistant in UT Southwestern Medical Center for about three years. Ying joined China affiliate of Eli Lilly and Company as a project statistician soon after her graduation. She has worked on a variety of phase III studies, including pain, diabetes, obesity, and ulcerative colitis.

TATINI CHAKRABORTY

Eli Lilly and Company

Tatini Chakraborty is a Senior Principal Statistician in Eli Lilly and Company, Bangalore, primarily working in early phase studies across various therapeutic areas. Prior to this she worked at GSK, specializing in the HIV portfolio and has five years of experience in the pharmaceutical industry.
Tatini was born in Kolkata, India. Fluent in Bengali, English and Hindi, she pursued her passion for data and analysis, earning a master's degree in Statistics from Presidency University, Kolkata.
Since moving to Bangalore in 2019, Tatini has embraced her love for traveling and cooking, and she also enjoys exploring her creative side as an amateur writer.

IDWSDS 2024

Session Info

S30

TECHNICAL SESSION

October 8th, 11:30 - 12:00 UTC

Using Functional Data Analysis for Surrogate Model Development

CLICK HERE TO ACCESS THE ZOOM ROOM

JHU/APL has developed a physics-based modeling and simulation software suite that simulates the performance of a system. That physics-based model has dozens of input and output variables and can take several minutes to run for each test case. To support risk assessment, concept studies, and system trades, development of a surrogate model would simplify model usage, reduce run time, and allow opportunity to share with other organizations. An effort was conducted to develop a surrogate model of that physics-based model by 1) exploring a robust down selection of critical input parameters to cover sampling of the analysis space and 2) leveraging functional data analysis to model two transient output curves to those critical input parameters. The goal of this brief will be to showcase the analysis process taken to develop such a surrogate model.

ORGANIZER: Elizabeth Murrin, Johns Hopkins Applied Physics Laboratory
CHAIR: Joseph Warfield, Johns Hopkins Applied Physics Laboratory
SPONSOR: Johns Hopkins University Applied Physics Laboratory (JHU/APL)

IDWSDS 2024

Speaker Bios

S30

ELIZABETH MURRIN

JHU/APL

Elizabeth Murrin is a Senior Statistician, Project Manager, and Section Supervisor at the Johns Hopkins Applied Physics Laboratory. She has been working at the lab for over 14 years with experience on various multi-domain projects alongside engineers, chemists, physicists, and other scientists. Her expertise includes design of experiments, regression analysis, surrogate modeling, anomaly detection, test and evaluation, and requirements development. She holds both a BS and MS in Statistics from the University of Maryland, Baltimore County.

CATHERINE CHALIKIAN

JHU/APL

Catherine Chalikian is an Associate Data Scientist at The Johns Hopkins Applied Physic Lab, where she supports projects across the lab with machine learning, text and data mining, surrogate modeling, regression analysis, and analytics development. In a previous career, Catherine worked as a business analyst and financial manager for Nexleaf Analytics. She holds an MS in Statistics from Wake Forest University, a Post-Baccalaureate Certificate in Mathematics from Smith, an MBA in Finance from the University of Pittsburgh, and College BA in French from Emory University.

IDWSDS 2024

Session Info

S31

TECHNICAL SESSION

October 8th, 11:30 - 12:00 UTC

Use of Artificial Intelligence and Statistics in the World of Mental Health

CLICK HERE TO ACCESS THE ZOOM ROOM

The use of AI and statistics in mental health has the potential to revolutionize diagnostic, treatment, and support processes, leading to improved patient outcomes and public health impact. This proposal outlines strategies for leveraging AI and statistical methods to enhance mental health care, including diagnostic support, personalized treatment, predictive analytics, mental health monitoring, therapy support, research, and public health initiatives.

Objectives:

- Develop AI-driven diagnostic tools to improve accuracy and timeliness of mental health disorder identification.
- Utilize statistical modeling for personalized treatment planning based on individual patient data.
- Implement AI and statistical methods to predict and prevent the onset of mental health disorders.
- Integrate AI-powered mental health monitoring tools to enable early detection of changes in mental well-being.
- Explore the use of AI-driven chatbots and virtual assistants to provide accessible mental health support and therapy.
- Apply statistical and AI methods to advance mental health research and drug development.
- Utilize statistics and AI to inform public health policies and interventions targeting mental health.

Leveraging AI and statistics in mental health holds significant promise for advancing diagnostic accuracy, treatment efficacy, and public health strategies. With this session we aim to improve mental health outcomes, reduce stigma, and enhance accessibility.

ORGANIZER: Arinjita Bhattacharyya, Merck

CHAIR: Arinjita Bhattacharyya, Merck

Speaker Bios

DR. ANNIE J. LEE

Columbia University Irving Medical Center

Annie J. Lee, PhD is an assistant professor of neurological science (in Neurology, the Sergievsky Center, and the Taub Institute). Dr. Lee is a graduate of Ewha Womans University in Seoul, Korea. She received an MS and PhD in Biostatistics from Columbia, where she focused on developing statistical methods to understand neurodegenerative diseases using genomic and family history data. In research, Dr. Lee is focused on understanding neurodegenerative diseases, in particular Alzheimer's disease, multiple sclerosis, and Parkinson's disease, using advanced statistical approaches and large-scale multi-omics data.

DR. SHARON-LISE NORMAND

Harvard Chan School of Public Health

Sharon-Lise Normand, Ph.D., is S. James Adelstein Professor of Health Care Policy (Biostatistics) in the Department of Health Care Policy at Harvard Medical School and Professor in the Department of Biostatistics at the Harvard School of Public Health. Dr. Normand's research focuses on the development of statistical methods for health services and outcomes research, primarily using Bayesian approaches, including the evaluation of medical devices in randomized and non-randomized settings for pre- and post-market assessments, causal inference, provider profiling, evidence synthesis, item response theory, and latent variables analyses. Her application areas include cardiovascular disease, severe mental illness, medical device safety and effectiveness, and medical technology diffusion. Dr. Normand is Director of the Medical Device Epidemiology Network's (MDEpiNet) Methodology Center.

Alzheimer's Disease Subtyping through Integration of Longitudinal Cognitive Function, Vascular Risk Factors, and Omics Data Using Novel Latent Mixture Model

Annie J. Lee, Columbia University Irving Medical Center

Treating and preventing neurodegenerative diseases is challenging due to the heterogeneity in older individuals, suggesting subgroups with shared biological features but varying responses to disease risk factors. High-throughput sequencing and conventional unsupervised clustering methods may not capture clinically relevant subtypes due to confounding factors and a lack of integration with clinical outcomes. To identify disease subtypes guided by a clinical outcome, existing supervised clustering methods use mixture models that focus on cross-sectional clinical outcomes and covariates. We propose a novel latent mixture model that incorporates longitudinal clinical outcomes and time-varying covariates to identify outcome-guided disease subtypes from high-dimensional omics data. Our approach identifies clinically meaningful subtypes based on the association of longitudinal clinical outcome and longitudinal risk factors while incorporating genetic pathway information for regularizing gene section. Applied to the ROSMAP study using brain transcriptomic profiles, longitudinal cognitive function, and vascular risk factors (e.g., hypertension, diabetes, stroke, and frailty), our method reveals Alzheimer's disease subtypes with clinicopathologic and neurobiological relevance. This work is crucial for designing targeted therapeutics and personalized clinical trials.

Causal Inference and Evidence on Drug Outcomes Among Elderly Schizophrenia Patients

Sharon-Lise Normand, Harvard Chan School of Public Health

Large observational databases from usual care settings provide opportunities to generate evidence on antipsychotic treatment outcomes for elderly persons with schizophrenia. However, comparing outcomes is complicated by multiple competing antipsychotic drugs; different longitudinal treatment patterns ranging from monotherapy to sequential or concurrent use of different antipsychotic drugs; lack of randomized trial evidence for many treatment regimens; and potential treatment effect heterogeneity associated with race/ethnicity and social determinants of health (SDH). This talk describes a program of causal inference research for mental health services and demonstrates results using targeted based minimum loss-based estimation (TMLE) for multi-valued treatment effect estimation in the cross-sectional setting, TMLE for binary effect estimation in the longitudinal setting, and transfer learning for subgroup effect estimation.

This work is joint with Denis Agniel, PhD, Max Rubinstein, PhD, and Marcela Horvitz-Lennon, MD, from the RAND Corporation; Larry Han, PhD from Northeastern University, and is funded by Grant R01MH130213 from the US National Institute of Mental Health.

IDWSDS 2024

Session Info

S32

TECHNICAL SESSION

October 8th, 12:00 - 13:00 UTC

Next Generation: Showcasing Young Portuguese Talent in Statistics and Data Science

CLICK HERE TO ACCESS THE ZOOM ROOM

This session is dedicated to highlighting the remarkable work of three emerging Portuguese researchers in the fields of statistics and data science. Each speaker will present their cutting-edge research, showcasing advancements in statistical methodologies, data analytics, and their applications across various industries.
The first presentation introduces a novel model integrating multiple data sources to enhance understanding of marine species distribution. The second explores a smart metering approach for water utilities, aimed at improving water loss management. The final presentation delves into multivariate time series analysis using advanced multilayer graph methodologies.
Attendees will gain insights into the latest advancements and applications in statistical modeling, data analysis, and time series analysis The session aims to foster dialogue, collaboration, and networking among professionals, academics, and students. Together, these talks showcase cutting-edge methodologies and their practical applications in addressing complex real-world challenges.
Join us to celebrate and support the promising contributions of the next generation of Portuguese statisticians and data scientists.

ORGANIZER: Lígia Henriques-Rodrigues, University of Évora and CIMA , Évora, Portugal
CHAIR: Lígia Henriques-Rodrigues, University of Évora and CIMA , Évora, Portugal
SPONSOR: Portuguese Statistical Society (SPE) and CWS

IDWSDS 2024

Speaker Bios

S32

DR. DANIELA SILVA

Daniela Silva is a PhD researcher specializing in applied mathematics, with a focus on spatial statistics and its applications in marine ecology. As part of a multidisciplinary team, her research integrates advancements in spatial statistics with real-world challenges, particularly in marine species dynamics and environmental interactions. Daniela has developed species distribution models to understand marine species dynamics and optimize sampling designs for research surveys. Recently, she has been developing spatio-temporal species distribution models to integrate different data sources. Daniela holds a PhD in Applied Mathematics and possesses key skills in spatial statistics, data integration, preferential sampling, and geostatistical modeling, with a focus on fish data. She was a researcher in the PREFERENTIAL Project and collaborates on the SARDINHA2020 and SARDINHA2030 projects.

DR. MARIA ALMEIDA SILVA

DEISI & COPELABS

Maria Almeida Silva is an Assistant Professor in the Department of Computer Engineering and Information Systems (DEISI) and a researcher at the Association for Research and Development in Cognition and People-centric Computing (COPELABS) from Lusófona University, Portugal. Maria is also a researcher at the Computational and Stochastic Mathematics (CEMAT) from Instituto Superior Técnico, Lisbon University, Portugal. She has a PhD in Statistics and Stochastic Processes from Instituto Superior Técnico. Maria works with the National Laboratory for Civil Engineering (LNEC) in Portugal to manage water losses efficiently. She also worked with Alcoitão Rehabilitation Medicine Center. The work presented is a joint work with Conceição Amado from IST and Dália Loureiro from LNEC.

DR. VANESSA SILVA

INESC TEC - CRACS & University of Porto, Portugal

Vanessa Alexandra Freitas da Silva is an Assistant Researcher at INESC TEC – CRACS and an Invited Assistant Professor at the Faculty of Sciences of the University of Porto. She obtained a Ph.D. in Computer Science from the University of Porto in 2023, and a M.Sc. in Networks and Informatics Systems Engineering from the University of Porto in 2018. Her main areas of research are Network Science and Time Series Analysis. Specifically, she is interested in analyzing time series data using complex network methodologies and graph theory. She is currently working on developing new methods for analyzing multivariate time series data through multilayer networks as well as on their exploration and analysis. She supervised two master's theses within this research area.

Joint Model for Zero-Inflated Data Combining Fishery-Dependent and Fishery-Independent Sources

Daniela Silva

Accurately identifying species distribution patterns is crucial for both scientific insight and societal impact. The increasing quantity and quality of ecological datasets present heightened statistical challenges, complicating spatio-temporal species dynamics comprehension. This study introduces a pioneering five-layer Joint model to address the task of integrating multiple data sources to enhance spatial fish distribution understanding in marine ecology. The model integrates fishery-independent and fishery-dependent data, accommodating zero-inflated data and distinct sampling processes. A comprehensive simulation study evaluates the model performance across various preferential sampling scenarios and sample sizes, elucidating its advantages and challenges. Our findings highlight the model's robustness in estimating preferential parameters, emphasizing differentiation between presence-absence and biomass observations. Evaluation of spatio-temporal covariance estimation, intercept parameters, and prediction performance underscores the model's reliability. Assessing the contribution of each data source reveals successful integration, providing a comprehensive representation of biomass patterns. Empirical validation within a real-world context further solidifies the model's efficacy in capturing species' spatio-temporal distribution. This research advances methodologies for integrating diverse datasets and contributes to a more informed understanding of marine species dynamics.

Stratified Sampling of Billing Data and Singular Spectrum Analysis for Smart Meter Installation and Flow Time Series Decomposition in Drinking Water Distribution Systems

Maria Almeida Silva, Department of Computer Engineering and Information Systems (DEISI) and Association for Research and Development in Cognition and People-centric Computing (COPELABS), Lusófona University, Portugal

Water utilities face significant challenges in managing water losses from leaks, bursts, and unauthorized consumption. In Portugal, the average non-revenue water for distribution systems was 27.1% in 2022. Smart metering technology is crucial for monitoring consumption and managing water losses. However, it is costly to acquire, install, operate, and maintain. This study supports water utilities by inferring total consumption using a representative sample of customers with smart meters rather than smart metering data from all customers. By combining billed metered consumption from smart meters with network flow time series, water utilities can obtain a complete view of non-revenue water over time. In a predominantly domestic zone, eight strata were identified through clustering analysis based only on customers' billing time series. Stratified sampling was employed, and a representative sample of 53% of the population was selected to infer essential consumption statistics with minimal error. Singular Spectrum Analysis revealed hidden non-revenue water components, enabling water utilities to develop strategies to reduce water losses. The successful outcomes demonstrated that a representative sample of customers provides accurate and meaningful consumption data essential for effective network management and water loss control. Moreover, using only billing data for sample selection benefits water utilities that may face challenges in obtaining additional consumer information.

Multivariate Time Series Analysis via Multilayer Graphs

Vanessa Silva, INESC TEC - CRACS and Faculty of Science, University of Porto, Portugal

Nowadays, we are surrounded by sensing devices that collect and store data over time, measuring different, often interdependent, variables. Multivariate time series analysis is not as well-established or mature a field as univariate analysis, primarily due to the challenges posed by serial and cross dependencies, as well as high dimensionality. Consequently, numerous studies have focused on developing new approaches and methodologies. Network science has emerged as a promising approach to address these challenges. These methods involve transforming a time series dataset into one or more complex networks, which can then be analyzed in depth to gain insight into the original time series. This talk provides insights into how to model multivariate time series data using multilayer graph structures. Specifically, it introduces two new multivariate time series mappings: Multilayer Horizontal Visibility Graphs and Multilayer Quantile Graphs. Additionally, a feature-based approach for multivariate time series analysis based on the resulting multilayer graphs will be presented. Finally, this talk will present a framework for multivariate time series data mining using multilayer graph methodologies, with practical applications.

IDWSDS 2024

Session Info

S33

TECHNICAL SESSION

October 8th, 12:00 - 13:00 UTC

Statistics and data science with machine learning and AI

CLICK HERE TO ACCESS THE ZOOM ROOM

This session explores innovative applications of advanced statistical methodologies and machine learning to understand human behavior and performance across different domains. Our three speakers will present cutting-edge research leveraging large-scale assessment data, clinical data, and online behavior data to gain new insights.

Dr. Minjeong Jeon from UCLA, USA will discuss novel network-based approaches for analyzing assessment data, moving beyond traditional psychometric methods. Her work applies network modeling and latent space analysis to uncover complex relationships between assessment items and respondent characteristics.

Dr. Minyoung Yun from ENSAM, France will present a machine learning method for predicting depression risk using limited clinical data. By applying graph convolutional networks to small datasets, this work demonstrates how advanced analytics can extract meaningful insights even with constrained data resources.

Finally, Dr. Heyoung Yang from KISTI, Korea will share research on identifying meaningful actions in online behavior sequences using natural language processing techniques. By analyzing log data from large-scale assessments, this work develops new methods to understand problem-solving processes and performance.

ORGANIZER: Heyoung Yang, Korea Institute of Science and Technology Information

CHAIR: Heyoung Yang, Korea Institute of Science and Technology Information

SPONSOR: Korea Institute of Science and Technology Information

Speaker Bios

PROF. MINJEONG JEON
UCLA

Minjeong Jeon, Ph.D., is a Professor of Advanced Quantitative Methods at the UCLA Department of Education.
Dr. Jeon received her Ph.D. in Quantitative Methods from UC Berkeley.
Prior to joining the UCLA faculty, she was an Assistant Professor of Quantitative Psychology at Ohio State University.
Her research revolves around developing, applying, and estimating latent variable models for studying measurement and growth.
Her recent research topics include latent space modeling, process modeling, and joint analysis.

DR. MINYOUNG YUN
ENSAM

Degree)
Master in chemical engineering.
PhD in mechanical engineering.
Experience)
2 years of experience as a Postdoctoral at ENSAM paris france, in a digital twin lab.
3 years of experience as a research engineer at Korea Institute of Science and Technology Information (KISTI).
Current research focus is digital twin in manufacturing process.

DR. HEYOUNG YANG
KISTI

Heyoung Yang, Ph.D., is a Principal Researcher at Korea Institute of Science and Technology Information. Dr. Yang received her Ph.D. and Master in Physics from Seoul National University.
Her main research field is using artificial intelligence to analyze various data to find insights, and she is particularly interested in science and technology big and small data including biomedical data and human behavioral data.

Next Generation Assessment: Methodological Advancements and Future Directions

Minjeong Jeon, UCLA

In this talk, I will explore emerging topics in psychometrics and how they contrast with traditional measurement research. I will present an example of applying social network analysis to clinical assessment data to identify individual patients' strength and weakness profiles. Furthermore, I will address new types of data generated by modern assessment techniques and emphasize the need for methodological and statistical innovations, along with potential areas for further research and development.

A Novel Machine Learning-Based Prediction Method for Patients at Risk of Developing Depressive Symptoms Using a Small Data

Minyoung Yun, ENSAM

This study introduces an innovative algorithm that leverages natural language processing (NLP) machine learning methods, specifically Word2Vec and Doc2Vec, to identify and verify meaningful actions within action sequences, utilizing the 2012 PIAAC data. The employed machine learning tools, Word2Vec and Doc2Vec, enable the visualization of action sequences in a 2D space, facilitating the discovery and verification of meaningful actions. Additionally, a neural network trained with the data was employed to predict participant scores based on modified action sequences, including or excluding identified meaningful actions. The proposed algorithm, incorporating these three machine learning components, successfully identified and verified meaningful actions in two example cases–'party invitation' and 'club membership' among PIAAC problems. This insight suggests the potential development of an interactive feedback system, providing necessary feedback for individuals not performing these actions in an educational assessment setting. Furthermore, the methodology's application can be extended to other fields such as medical diagnosis and marketing, like identifying disease conditions of depression or dementia by observing behavioral patterns, or predicting the likelihood of purchasing a product through consumer behavior.

Human Action Sequence Logdata-Based Classification and Prediction Study

Heyoung Yang, KISTI

This study presents an innovative algorithm that uses natural language processing (NLP) and machine learning methods, specifically Word2Vec and Doc2Vec, to identify and verify meaningful actions within action sequences using 2012 PIAAC data. These tools visualize action sequences in 2D space, aiding in the discovery of meaningful actions. A neural network predicts participant scores based on modified sequences, highlighting meaningful actions. The algorithm successfully verified actions in two cases–'party invitation' and 'club membership'–suggesting the potential for an interactive feedback system in educational assessments. Additionally, the methodology could extend to fields like medical diagnosis and marketing, identifying conditions like depression or predicting product purchases through behavioral patterns.

IDWSDS 2024

Session Info

S34

PROMOTING PHD STUDENTS AND THEIR RESEARCH

October 8th, 12:00 - 13:00 UTC

Showcasing Research by PhD Students from MRC Biostatistics Unit (Efficient Study Design)

CLICK HERE TO ACCESS THE ZOOM ROOM

This session will allow 3 female researchers in their final year to showcase the major results from the PhD work as students based at the MRC Biostatistics Unit. These students are part of the bigger research group on the theme of Efficient Study Design, which develops and applies novel statistical designs to the improvement of clinical trials.

ORGANIZER: Sofia S. Villar, MRC Biostatistics Unit (University of Cambridge)
CHAIR: Sofia S. Villar, MRC Biostatistics Unit (University of Cambridge)
SPONSOR: MRC Biostatistics Unit (University of Cambridge)

Speaker Bios

JULIETTE LIMOZIN
MRC Biostatistics Unit (University of Cambridge)

Juliette is a second-year PhD student supervised by Dr Li Su and Dr Shaun Seaman. Her research is primarily focused on causal inference in longitudinal studies, utilising target trial emulation to improve the robustness and reliability of study results. Juliette's work aims to bridge the gap between theoretical statistical methodologies and practical applications in public health and clinical research.

BETHANY HEATH
MRC Biostatistics Unit (University of Cambridge)

Bethany is a third year PhD student in the MRC Biostatistics Unit. Her research focuses on testing policies for epidemics particularly focusing on pooled testing policies. Her work involves developing agent-based models for evaluating these different testing methods.
Bethany has received one of six travel grants from the IBS British and Irish Region for career-young biostatisticians to attend the International Biometric Conference. The conference is taking place in Atlanta in December. She will be giving a talk on her work on evaluating testing policies for epidemic management particularly focusing on the use of pooled testing.

XIJIN CHEN
MRC Biostatistics Unit (University of Cambridge)

I am a PhD student with a research focus on dose-finding studies in early-phase clinical trial designs. Previously, I worked on addressing missing data problems in response-adaptive designs. Passionate about the challenges in adaptive clinical trial designs, I aim to contribute to advancing innovative approaches that enhance patient outcomes and accelerate the drug development process.

Inference Procedures in Sequential Trial Emulation With Survival Outcomes: Comparing Confidence Intervals Based on the Sandwich Variance Estimator, Bootstrap and Jackknife

Juliette Limozin, MRC Biostatistics Unit (University of Cambridge)

Sequential trial emulation (STE) is an approach to estimating causal treatment effects by emulating a sequence of target trials from observational data. In STE, inverse probability weighting is commonly utilised to address time-varying confounding and/or dependent censoring. Then structural models for potential outcomes are applied to the weighted data to estimate treatment effects. For inference, the simple sandwich variance estimator is popular but conservative, while nonparametric bootstrap is computationally expensive, and a more efficient alternative, linearised estimating function (LEF) bootstrap, has not been adapted to STE. We evaluated the performance of various methods for constructing confidence intervals (CIs) of marginal risk differences in STE with survival outcomes by comparing the coverage of CIs based on nonparametric/LEF bootstrap, jackknife, and the sandwich variance estimator through simulations. LEF bootstrap CIs demonstrated the best coverage with small/moderate sample sizes, low event rates and low treatment prevalence, which were the motivating scenarios for STE. They were less affected by treatment group imbalance and faster to compute than nonparametric bootstrap CIs. With large sample sizes and medium/high event rates, the sandwich-variance-estimator-based CIs had the best coverage and were the fastest to compute. These findings offer guidance in constructing CIs in causal survival analysis using STE.

Evaluating Testing Policies for Managing Emerging Epidemics in Resource-Constrained Settings'

Bethany Heath, MRC Biostatistics Unit (University of Cambridge)

Pooled testing, where multiple samples are combined for testing, can reduce costs but was underutilized during the COVID-19 pandemic. This study investigates the effectiveness of pooled testing in community and hospital settings using a dynamic model. Factors such as testing capacity, symptom prevalence, compliance, and test type were considered. Results show pooled testing can significantly reduce infections, especially when compliance is low. However, its effectiveness varies based on specific settings and conditions. This research provides valuable insights for policymakers considering pooled testing strategies in future epidemics.

Using ctDNA as a Novel Biomarker of Efficacy for Dose-Finding Designs in Oncology

Xijin Chen, MRC Biostatistics Unit (University of Cambridge)

Dose-finding trials are designed to identify a safe and potentially effective drug dose and schedule during the early phase of clinical trials. Historically, Bayesian adaptive dose-escalation methods in Phase I trials in cancer have mainly focused on toxicity endpoints rather than efficacy endpoints. This is partly because efficacy readouts are often not available soon enough for dose escalation decisions. In the last decade, `liquid biopsy' technologies have been developed, which may provide a readout of treatment response much earlier than conventional endpoints. This project develops a novel design that uses a biomarker, circulating tumor DNA (ctDNA), with toxicity and activity outcomes in dose-finding studies. Simulation results show that the proposed approach can yield significantly shorter trial duration and may improve identification of the target dose. In addition, this approach has the potential to minimise the time individual patients spend on potentially inactive trial therapies.

IDWSDS 2024

Session Info

S35

TECHNICAL SESSION

October 8th, 14:00 - 15:30 UTC

Pioneering Women in Statistics and Data Science: Bridging Global Research and Innovation

CLICK HERE TO ACCESS THE ZOOM ROOM

This session brings together three distinguished women statisticians from around the world, each a leader in advancing statistical science and its applications across various domains. Dr. Lynne Billard (University of Georgia, USA), Dr. Anuška Ferligoj (University of Ljubljana, Slovenia), and Dr. Mihoko Minami (Keio University, Japan) will present cutting-edge research that spans symbolic data analysis, multivariate clustering methods, and statistical modeling for biological and environmental sciences.

This session not only celebrates the contributions of these women to statistical research but also fosters international collaboration and encourages the participation of women in the field of statistics and data science worldwide.

ORGANIZER: Taerim Lee, Korea National Open University

CHAIR: Taerim Lee, Korea National Open University

SPONSOR: Women in Statistics in Korea (WISK)

Speaker Bios

PROF. LYNNE BILLARD

University of Georgia

https://www.stat.uga.edu/directory/people/lynne-billard

Dr. Lynne Billard is an Australian statistician and Distinguished Research Professor at the University of Georgia, known for her contributions to statistics, leadership in professional societies, and advocacy for women in science. She has served as President of both the American Statistical Association (1996) and the International Biometric Society (1994-1995), one of the few people to have led both organizations.

Her research spans areas such as epidemic theory including AIDS research, stochastic processes, sequential analysis and symbolic data analysis. Dr. Billard's work has earned her numerous accolades, including the COPSS F.N. David Award, Elizabeth Scott Award, and the Janet L. Norwood Award. She is also a Fellow of the American Statistical Association and the Institute of Mathematical Statistics, and an elected member of the International Statistical Institute.

PROF. ANUŠKA FERLIGOJ

University of Ljubljana, Ljubljana, Slovenia

https://en.wikipedia.org/wiki/Anu%C5%A1ka_Ferligoj

Anuška Ferligoj is an Emeritus Professor at the University of Ljubljana. Her research interests include multivariate analysis (clustering with constraints, multicriteria clustering) and social network analysis (blockmodeling, quality of network measurement).
She has published over 100 papers, several book chapters, and books. For the monograph co-authored with Patrick Doreian and Vladimir Batagelj "Generalized Blockmodeling" published by Cambridge University Press (2005), they obtained the Harrison White Outstanding Book Award 2007, given by the Mathematical Sociology Section at the American Sociological Association.
She is an elected member of the European Academy of Sociology and the International Statistical Institute. She received several awards. In 2010, she received Doctor et Professor Honoris Causa at Eotvos Lorand University in Budapest.

PROF. MIHOKO MINAMI

Keio University, Japan

https://mminami.math.keio.ac.jp/

Mihoko Minami is a Professor, Department of Mathematics, Faculty of Science and Technology, Keio University, Japan. She received her BA from Ochanomizu University in 1982, MA and Ph.D. from University of California, San Diego in 1990 and 1993, respectively.
Her research interests include distribution theories (Multivariate Inverse Gaussian distribution, Lagrange family of distributions) and multivariate analysis (independent component analysis, zero-inflated models, clustering for distributions, repeated measure/mixed effect models, non-parametric models).
She currently serves as the director of Japanese society of applied statistics, and the chair of the special committee for the promotion of diversity, Japan statistical society.

Distributions as Numbers

Lynne Billard, University of Georgia

Today, massively large data sets are routine and ubiquitous. What is not so routine is how to analyze these data. One approach is to aggregate the data according to some scientific criteria. The resultant data are perforce symbolic data, i.e., lists, intervals, histograms, and so on. Applications abound, especially in the physical, medical and social sciences. Other data sets (small or large in size) are naturally symbolic valued.

Unlike classical data which are points in p-dimensional space, symbolic data are hypercubes in p-dimensional space. We describe such data and how they arise. We illustrate briefly some of the differences between classical and symbolic data. In particular, we note that the often-seen approach of taking classical surrogates such as aggregated means is an incorrect approach since this discards a lot of information inherent to the data set.

Clustering of Attribute And/or Network Data

Anuška Ferligoj, University of Ljubljana, Ljubljana, Slovenia

A large class of clustering problems can be formulated as an optimizational problem in which the best clustering is searched for among all feasible clustering according to a selected criterion function. This clustering approach can be applied to a variety of very interesting clustering problems, as it is possible to adapt it to a concrete clustering problem by an appropriate specification of the criterion function and/or by the definition of the set of feasible clusterings.

Both, the blockmodeling problem (clustering of the network data) and the clustering with relational constraint problem (clustering of the attribute and network data) can be very successfully treated by this approach. It also opens many new developments in these areas.

Model for Bycatch and Clustering Method for Distributions

Mihoko Minami, Keio University, Japan

Bycatch refers unwanted marine creature that are caught in the net while fishing for another species. Count data on bycatch, and catch of some target species, can have many zero-valued observations, but also include large values when aggregations of animals are caught. We applied zero-inflated negative regression models and a few other models with spline smoothing to shark bycatch data from the eastern Pacific Ocean tuna purse-seine fishery over 11 years.

Comparison of trends among models suggests that the negative binomial regression model may overestimate model coefficients and thus the temporal trend when fitted to data with many zero-valued observations. We theoretically investigated why this phenomena happens.

We also introduce a clustering method for distributions and a testing procedure to find homogeneous subpopulations that are strongly desired for building stock assessment models for the target species.

IDWSDS 2024

Session Info

S36

TECHNICAL SESSION

October 8th, 14:00 - 15:00 UTC

Integration of Mobile and Wearable Data to Study the Interrelationships Between Biological Processes and Mood in Real Time

CLICK HERE TO ACCESS THE ZOOM ROOM

This session will explore how digital health technologies can enhance our understanding of the relationships and implications in biological functions of motor activity and mood patterns. To comprehensively study the role of motor activity, the cross-site collaboration entitled Motor Activity Research Consortium for Health (mMARCH) was established to standardize methods and analytic approaches to investigate associations between motor activity, mood, and related disorders across multiple studies and cultures. The mMARCH initiative enabled the groups to efficiently share and combine data to learn more about how activity affects different disorders and diseases across many populations, including mood disorders, sleep patterns, circadian rhythms, genetic studies, emotion, eating, and other disorders that impact public health which can also define targets for prevention and intervention studies. The content will include three presentations chosen to showcase (1) Processing of Accelerometry Data with GGIR in mMARCH, (2) Compliance with Ecological Momentary Assessment (EMA), and (3) Integrative Modeling of Accelerometry-Derived Sleep, Physical Activity and Circadian Rhythm Domains with current and remitted Major Depression.

ORGANIZER: Sun Kang, NIH

CHAIR: Sun Kang, NIH

Speaker Bios

DR. SUN KANG
NIH

Sun Jung Kang, PhD, is a Statistician at the National Institute of Mental Health. She focuses on the development of translational studies to identify the regulatory systems underlying motor activity and sleep across species by joint analysis of multiple domains. She worked at Albany Stratton VA Medical Center and SUNY Downstate Medical Center before she joined NIMH. She received her BA in Mathematics from the University of Virginia, an MS in Applied Mathematics from New York University, a PhD in Applied Mathematics and Statistics from State University of New York Stony Brook, and post-doctoral training from Duke University and Case Western Reserve University.

DR. WEI GUO
NIH

Wei Guo, PhD, is a Biostatistician at the National Institute of Mental Health. She is responsible for data analysis for several projects including research on mobile technologies in mental and physical health. She received her Bachelor's and Master's degree in mathematics and statistics from Northeast Normal University, and a PhD in Biostatistics from University of Hong Kong. She has established a pipeline to process high level multi-level repeated-measure data from several thousand participants that will facilitate cross-study comparability and increase statistical power. She has provided documentation to allow investigators from multiple international sites to process the data independently and guided them to visualize and analyze the data. She has led the transition to using novel statistical programs in R for multi-level repeated-measure data and analyses of data from different sites with different devices to promote data harmonization and cross-site validation.

ANANYA SWAMINATHAN
NIH

Ananya Swaminathan is a post-baccalaureate IRTA fellow at the National Institute of Mental Health. She received her Bachelor's and Master's degree in biomedical engineering from Johns Hopkins University. She hopes to pursue a PhD in biostatistics or biomedical informatics.

Abstracts

Integrative Modeling of Associations between Accelerometry-Derived Sleep, Physical Activity and Circadian Rhythms Domains with Current or Remitted Major Depression in a Community Sample

Sun Kang, NIH

Accelerometry has been increasingly used as an objective index of sleep (SL), physical activity (PA), and circadian rhythms (CR) in people with mood disorders. However, most prior research has focused on SL or PA alone without consideration of the strong within- and cross-domain intercorrelations. Moreover, few studies have distinguished between trait and state profiles of accelerometry domains in Major Depressive Disorder (MDD). The aims of the study are (1) to identify joint and individual components of the domains derived from accelerometry, including SL, PA, and CR using the Joint and Individual Variation Explained method (JIVE), a novel multimodal integrative dimension reduction technique; and (2) to examine associations between joint and individual components with current and remitted MDD. The sample included 2317 participants from a cohort study (Lausanne, Switzerland). The results show that both current and remitted depression were associated with the first two joint components that were distinguished by the salience of physical activity and sleep timing, respectively. Application of a novel multi-modal dimension reduction technique demonstrates the importance of joint influences of SL, PA, and their timing on MDD. This work illustrates the value of accelerometry as a potential biomarker for subtypes of depression and highlights the importance of consideration of the full 24-hour sleep-wake cycle in future studies.

Processing of Accelerometry Data with GGIR in Motor Activity Research Consortium for Health

Wei Guo, NIH

Motor activity has received more attention to study mood and related disorders. The R package GGIR had been used as a popular tool to process the raw accelerometer data. After running GGIR, researchers had to load the GGIR output, clean the data and extract important features of interest to study the association between the features and health outcomes. For this purpose, the R package postGGIR was developed for all GGIR users to examine and summarize the output of GGIR, clean activity data and extract features of three domains of sleep, physical activity and circadian rhythmicity. Further, the JIVE program decomposed all feature matrix into joint variation across three domains, individual variation to each domain and residual noise to study the interaction between features. This package will generate a few comprehensive reports of data processing and graphical presentation of all extracted features in the .html format by R Markdown. This package has been widely used in the mMARCH Consortium recently.

Compliance with Simultaneous Collection of Ecological Momentary Assessment (EMA) and Biologic Measures

Ananya Swaminathan, NIH

Ecological Momentary Assessment (EMA) is a tool that can be used to capture in-the-moment data on mood states and their corresponding contexts. Several previous studies have simultaneously collected EMA or mood diary data and biological measures. For example, cortisol, which is a commonly collected marker, can be used to examine associations between momentary mood states and the stress system. However, most studies that involve collection of cortisol in conjunction with EMA have short collection periods (<5 days). As a result, we were motivated to examine the feasibility of simultaneous collection of cortisol and EMA in a naturalistic setting over a longer period of time. We found that the overall compliance for EMA was 77.4%, while the overall compliance for cortisol was about 65.9%. We also found that compliance was affected by age group, day of week, and time of day.

IDWSDS 2024

Session Info

S37

TECHNICAL AND CAREER DEVELOPMENT SESSION

October 8th, 14:00 - 15:00 UTC

The Role of Women in Pharma Analytics: End-to-End Analytics for Dynamic Targeting

CLICK HERE TO ACCESS THE ZOOM ROOM

The Business Insights & Analytics (BI&A) group at Eli Lilly is a group of data engineers, analysts, scientists, and integrators who use analytics to drive better commercial decisions. The presenters will first highlight their career experiences and the roles that women play at all levels within BI&A. We will then demonstrate the range of analytics innovation that women drive in BI&A in the groundbreaking project “Dynamic Targeting”.

The goal of the project was to use advanced analytics to dynamically identify and engage with customers. We will discuss the unique challenges of the pharmaceutical marketing space and the role of women in analytics in each stage.

We will then discuss the extensive data engineering, ML model development, and deployment requirements of the final analytics solution. We will first also discuss the organizational change management (OCM) needed to strategize and deliver this transformational capability.

ORGANIZER: Helena Baptista, Eli Lilly & Co.

CHAIR: Helena Baptista, Eli Lilly & Co.

SPONSOR: Eli Lilly & Co.

Speaker Bios

LAURA CHEETHAM
Eli Lilly & Co.

Laura is an Executive Director of BI&A Capabilities, Innovation & Media at Eli Lilly & Company. She holds a Bachelor's degree in Media Arts & Sciences from Indiana University in Bloomington.

SUMEGHA SETIA
Eli Lilly & Co.

Sumegha is a Senior Data Engineer at Eli Lilly, bringing over eight years of experience across various domains. She holds a bachelor's degree in computer science engineering and an MBA in International Business from IIFT.

DR. STEPHANIE CHEN
Eli Lilly & Co.

Stephanie is currently a Sr Director of BI&A HCP Data Science at Eli Lilly and Company. She leads a team of data engineers, data scientists, and ML engineers who use advanced analytics to drive better decision-making. She holds a Ph.D in Statistics from the North Carolina State University at Raleigh and a BS in Ecology from the University of Michigan.

MARIA VARCHENKO
Eli Lilly & Co.

Maria is a Senior Director of Data Engineering at Eli Lilly & Company. Maria holds a Master's degree in Statistics from Purdue University in Indiana.

Organizational Change Management (OCM)

Laura Cheetham, Eli Lilly & Co.

Laura will discuss leading through change, and transforming the organization to become more systematically data driven via Dynamic Targeting. Laura will also share her career experiences and the types of roles that women play within BI&A.

Data Engineering

Sumegha Setia, Eli Lilly & Co.

Sumegha will explore key aspects of data engineering, including data collection, transformation using business rules, and the delivery to downstream systems. She will also delve into the tools and technologies utilized throughout these processes.

ML Model Solution

Stephanie Chen, Eli Lilly & Co.

Stephanie will touch on her career development inside Lilly and the role of women in BI&A. She will also present the development of the Machine Learning models solution.

IDWSDS 2024

Session Info

S38

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

October 8th, 15:00 - 16:00 UTC

The Art of the Invite: Crafting Successful Invited Session Proposals

CLICK HERE TO ACCESS THE ZOOM ROOM

Invited sessions at conferences provide important opportunities for the exchange of ideas. But how do we get invited? And how can we do the inviting? In this panel, we will bring together experienced women in statistics from all career stages to share their tips on organizing invited sessions. Our panelists have planned and participated in numerous successful invited sessions at statistical conferences and have served on program committees to plan and select these sessions on a large scale. This panel is intended to demystify the invited session proposal process and to empower researchers to submit their ideas in the future.

ORGANIZER: Ashley Mullan, Vanderbilt University

CHAIR: Ashley Mullan, Vanderbilt University

SPONSOR: Caucus for Women in Statistics and Data Science

Speaker Bios

DR. SUHWON LEE

University of Missouri

https://leesuh.mufaculty.umsystem.edu/

Suhwon Lee is a teaching professor and Director of the Center for Applied Statistics and Data Analysis in the Department of Statistics at the University of Missouri, which she joined in 2004. She also serves as an adjunct faculty member for the Master of Public Health program since 2009.
Her research interests focus on interdisciplinary work, fusing statistical methods with other fields to tackle complex issues and providing quantitative insights for innovative solutions. She collaborates with experts across disciplines to adapt statistical techniques to diverse datasets.
Passionate about statistical education, she teaches various undergraduate and graduate courses, including Applied Statistical Models, Categorical Data Analysis, Nonparametrics, Statistical Software & Data Analysis, Sampling, and Statistical Methods in Health Sciences.

DR. LUCY D. MCGOWAN

Wake Forest University

https://www.lucymcgowan.com/

Lucy D'Agostino McGowan is an assistant professor in the Department of Statistical Sciences at Wake Forest University. She received her PhD in Biostatistics from Vanderbilt University and completed her postdoctoral training at Johns Hopkins University Bloomberg School of Public Health. Her research focuses on analytic design theory, statistical communication, causal inference, and data science pedagogy. Dr. D'Agostino McGowan can be found blogging at livefreeordichotomize.com, on Twitter @LucyStats, and podcasting on the American Journal of Epidemiology partner podcast, Casual Inference.

DR. ANA ORTEGA-VILLA

National Institute of Allergy and Infectious Diseases

https://www.niaid.nih.gov/about/brb-staff-ortega-villa

Dr. Ana M. Ortega-Villa joined the Biostatistics Research Branch (BRB) in 2018 and serves as a mathematical statistician. Prior to joining the BRB, Dr. Ortega-Villa obtained her Ph.D. in Statistics from Virginia Tech and completed post-doctoral fellowships at both the Eunice Kennedy Shriver National institute of Child Health and Human Development and the National Cancer Institute. Her interests include longitudinal data, mixed models, vaccines, immunology, research capacity building in developing countries, statistics education, and diversity and inclusion initiatives.

IDWSDS 2024

Session Info

S39

TECHNICAL SESSION

October 8th, 15:00 - 16:00 UTC

Navigating the Noise: Statistical Methods for Measurement Error in Data

CLICK HERE TO ACCESS THE ZOOM ROOM

Measurement error is an inevitable challenge in data collection across various fields, from healthcare to social sciences. These errors can significantly impact the validity and reliability of research findings if not properly addressed. This session brings together experts to discuss the latest statistical methods and best practices for managing and correcting measurement error. Attendees will gain insights into techniques such as error modeling, data validation, and bias correction. Through practical examples and case studies, this session aims to equip researchers with the tools needed to enhance data accuracy and improve the robustness of their analytical results.

ORGANIZER: Sarah Lotspeich, Wake Forest University

CHAIR: Sarah Lotspeich, Wake Forest University

Speaker Bios

DR. LILIAN BOE

Memorial Sloan Kettering Cancer Center

https://www.mskcc.org/profile/lillian-boe

Lillian (Lily) Boe is a Principal Biostatistician in the Department of Epidemiology & Biostatistics at Memorial Sloan Kettering Cancer Center (MSK). In this role, she engages in collaborative research with investigators from the Plastic & Reconstructive Surgery Service and the Head & Neck Service in the Department of Surgery, as well as with the Department of Radiation Oncology. Her PhD dissertation
research, completed in the Department of Biostatistics, Epidemiology, and Informatics at the University of Pennsylvania Perelman School of Medicine, was focused on statistical approaches for reducing bias and improving variance estimation in the presence of covariate and outcome measurement error in time-to-event settings.

DR. SARAH PESKOE

Duke University

https://biostat.duke.edu/profile/sarah-peskoe

Sarah Peskoe is an assistant professor of Biostatistics and Bioinformatics and the Director of the Data Science and Statistics Lab (DSSL) in the Aging Center at Duke University. Her research involves evaluating and correcting biases that arise in epidemiologic studies, with a particular focus on applications to aging-related research. She earned her PhD in Biostatistics from Harvard University.

DR. GRACE YI

University of Western Ontario

http://fisher.stats.uwo.ca/faculty/yyi/

Grace Y. Yi is a Professor and Tier I Canada Research Chair in Data Science at the University of Western Ontario. Her research focuses on statistical methodology for measurement error, causal inference, missing data, high-dimensional data, and statistical machine learning. She authored "Statistical Analysis with Measurement Error or Misclassification" (2017) and co-edited "Handbook of Measurement Error Models" (2021). She is a Fellow of the ASA, IMS, and an Elected Member of the ISI. In 2010, she received the Centre de Recherches Mathématiques and the Statistical Society of Canada (CRM-SSC) Prize. She has served key editorial roles for the Electronic Journal of Statistics, New England Journal of Statistics in Data Science, and Canadian Journal of Statistics. She was the chair of the ASA Lifetime Data Science Section, President of the Statistical Society of Canada, and she founded the first chapter (Canada Chapter, established in 2012) of the ICSA.

DR. LI TANG

St. Jude Children's Research Hospital

https://www.stjude.org/directory/t/li-tang.html

Dr. Tang is an Associate Professor of Biostatistics, currently at St. Jude Children's research Hospital. She also co-leads the Biostatistics Shared Resource. Her expertise is in measurement error, missing data, longitudinal modeling, study design and recently on machine and deep learning methods. She serves on the Statistical Advisory Panel for Nature Medicine and has published numersouly on leading medical journals such as JAMA and major statistical journals as lead author.

Practical Considerations for Sandwich Variance Estimation in Two-Stage Regression Settings With Applications in Regression Calibration

Lilian Boe, Memorial Sloan Kettering Cancer Center

We present a practical approach for computing the sandwich variance estimator in two-stage regression model settings. As a motivating example for two-stage regression, we consider regression calibration, a popular approach for addressing covariate measurement error. The sandwich variance approach has been rarely applied in regression calibration, despite it requiring less computation time than popular resampling approaches for variance estimation, specifically the bootstrap. This is likely due to requiring specialized statistical coding. We first outline the steps needed to compute the sandwich variance estimator. We then develop a convenient method of computation in R for sandwich variance estimation, which leverages standard regression model outputs and existing R functions and can be applied in the case of a simple random sample or complex survey design. Our package, sandwich2stage, available on GitHub, can be used directly to compute sandwich variance estimates for the two-stage setting of regression calibration. We use a simulation study to compare the sandwich to a resampling variance approach for both settings. Finally, we further compare these two variance estimation approaches in the Women's Health Initiative (WHI) and Hispanic Community Health Study/Study of Latinos (HCHS/SOL). The sandwich variance estimator typically had good numerical performance, but simple Wald bootstrap confidence intervals were unstable or over-covered in certain settings.

Missing Data in EHR: A Measurement Error Problem in Disguise

Sarah Peskoe, Duke University

Electronic Health Records (EHR) are a valuable resource for clinical research, yet they are often plagued by missing data. This issue arises for various reasons, such as unperformed tests, patients not seeking medical care, or receiving care outside the recorded health system. Traditional methods like imputation and inverse probability weighting are commonly used to address missing data. However, EHR presents a unique challenge where the absence of data is not always apparent, complicating the identification of missing data. This presentation explores the concept that many missing data issues in EHR can be reframed as measurement error problems. For instance, healthier patients, with fewer health system interactions, often have outdated and potentially inaccurate measures (e.g., cholesterol, A1C), leading to measurement errors. Conversely, patients with frequent health system interactions are more likely to have conditions diagnosed, resulting in differential misclassification of exposures and outcomes. Finally, we will discuss how leveraging this understanding can enhance clinical research using EHR, including our recent methodological developments to addressing this missing data in EHR through a measurement error lens.

Learning with Crowdsourced Noisy Annotations under Instance-Dependent Transition Models

Grace Yi, University of Western Ontario

The predictive ability of supervised learning algorithms hinges on the quality of annotated examples, whose labels often come from multiple crowdsourced annotators with diverse experiences. To aggregate noisy crowdsourced annotations, many existing methods employ an annotator-specific instance-independent noise transition matrix to characterize the labeling skills of each annotator. Learning an instance-dependent noise transition model, however, is challenging and remains relatively less explored. To address this problem, in this paper, we formulate the noise transition model in a Bayesian framework and subsequently design a new label correction algorithm. Specifically, we approximate the instance-dependent noise transition matrices using a Bayesian network with a hierarchical spike and slab prior. To theoretically characterize the distance between the noise transition model and the true instance-dependent noise transition matrix, we provide a posterior-concentration theorem that ensures the posterior consistency in terms of the Hellinger distance. We further formulate the label correction process as a hypothesis testing problem and propose a novel algorithm to infer the true label from the noisy annotations based on the pairwise likelihood ratio test. Moreover, we establish an information-theoretic bound on the Bayes error for the proposed method. We validate the effectiveness of our approach through experiments on benchmark and real-world datasets.

Misclassification Mechanism Revisit: Common Misconception and Insight

Li Tang, St. Jude Children's Research Hospital

In many epidemiological and clinical studies, misclassification can occur in one or several variables, leading to potentially invalid analytical results, such as biased estimates of odds ratios, if not properly corrected. This study revisits the misclassification mechanism, addressing common misconceptions and providing new insights. We specifically focus on situations where correlated binary response variables are subject to misclassification. Building upon previous research, we introduce an approach to adjust for potentially complex differential misclassification using internal validation sampling at multiple study time points. To illustrate the misconception, we present a case study using longitudinal data on bacterial vaginosis from the HIV Epidemiology Research (HER) Study.

IDWSDS 2024

Session Info

S40

SHOWCASE OF APPLIED STATISTICAL WORK WITHIN A NEW RESEARCH CENTRE

October 8th, 15:30 - 16:30 UTC

Perspectives on Local and Global Health: Showcasing Work by Women and Non-Binary People at the London School of Hygiene & Tropical Medicine's New Data Science and Statistics Centre

CLICK HERE TO ACCESS THE ZOOM ROOM

In this session, we will introduce the newly launched Centre for Data and Statistical Science for Health (DASH Centre) which aims to bring together data science expertise from across the London School of Hygiene & Tropical Medicine's four sites to generate new opportunities for research, training and impact. We will spotlight the research of four individuals who work with observational health datasets both locally and globally with applications that span infectious and genetic diseases, maternal health, and artificial intelligence. We will close the session with some reflections on inclusion and gender in our academic environment.

ORGANIZER: Poppy Mallinson, London School of Hygiene & Tropical Medicine

CHAIR: Poppy Mallinson, London School of Hygiene & Tropical Medicine

SPONSOR: London School of Hygiene & Tropical Medicine

Speaker Bios

DR. POPPY MALLINSON

University of Lisbon

https://www.lshtm.ac.uk/aboutus/people/mallinson.poppy

Poppy Mallinson is an assistant professor at the London School of Hygiene & Tropical Medicine and a Challenge Lead for the new Centre for Data and Statistical Science for Health (DASH Centre). Her research sits at the interface of data science and epidemiology, with a substantive focus on non-communicable diseases in low-resource settings. She aims to leverage high-resolution data from various modalities (e.g. medical imaging, metabolomics, surveys) to better understand and diagnose non-communicable diseases. She teaches epidemiology and data science for Masters and Research degree programmes at LSHTM.

DR. EMILY GRANGER

London School of Hygiene & Tropical Medicine

https://www.lshtm.ac.uk/aboutus/people/granger.emily

Emily is a research fellow at the London School of Hygiene and Tropical Medicine and a member of the Centre for Data and Statistical Science for Health. Her current research focuses on the use of target trial emulation for estimating treatment effects using observational data, with a focus on treatments in cystic fibrosis. She is a member of the Cystic Fibrosis Trial Emulation Network (CF-TEN), an international collaborative focussed on using data from cystic fibrosis patient registries available across the world for target trial emulations. Emily's other research interests include causal inference methods for the analysis of longitudinal data and methods for handling missing data. She has an MSci in Mathematics and Statistics from Lancaster University, and completed her PhD in Medical Statistics at the University of Manchester.

DR. ORLAGH CARROLL

London School of Hygiene & Tropical Medicine

https://www.lshtm.ac.uk/aboutus/people/carroll.orlagh

Dr. Orlagh Carroll is a research fellow in Statistics and Health Data Science at the London School of Hygiene and Tropical Medicine (LSHTM) and a co-lead of the Early Career Researcher group in the Centre for Data and Statistical Science for Health (DASH) at LSHTM. She currently works for Dr Enny Paixao Cruz on a project looking at the impact of in utero infectious disease exposure on childhood outcomes. Dr Carroll's research interests include missing data and causal inference.

EM PRESTIGE

London School of Hygiene & Tropical Medicine

https://www.lshtm.ac.uk/aboutus/people/prestige.em

Em Prestige is a research fellow and PhD student at the London School of Hygiene & Tropical Medicine, where they have been based for the past three years. They have a particular interest in combining statistical and mathematical modelling methods, and learning to use tools from both fields to answer questions surrounding health inequalities. Their current work focuses on using Electronic Health Records (EHRs) to explore disparities in respiratory viruses in England. Outside of their research, they focus on staff and student advocacy as a student representative, a member of the Gender Equity Taskforce and as someone working towards disability inclusion at their institution.

Abstracts

Artificial Intelligence for Low Cost and High Resolution Phenotyping of Metabolic Diseases: Some Examples of Recent Findings From India

Poppy Mallinson, London School of Hygiene & Tropical Medicine

In this talk I will share some highlights from four different analyses in which we have used low-cost and readily available data collection devices such as smartphones to proxy more costly clinical measures that we commonly need in epidemiological research or clinical practice. In each case we use artificial-intelligence based algorithms (such as convolutional neural networks) to predict the 'ground-truth' clinical measures based on the low-cost data inputs, and test the performance on a held-out portion of the data. The data I will show comes mostly from a longitudinal cohort study called APCAPS in Telangana State in India. Use cases include estimating body composition from photographs captured on smart phones and cheap anthropometric devices, estimating age and ageing related phenotypes from smartphone videos of people walking, and automated classification of fatty liver disease from ultrasound images. In the short term we plan to use these to improve the efficiency and cost of our research data collection, and in the longer term they could be externally validated for clinical use in rural or remote populations.

Target Trial Emulation in Cystic Fibrosis: Optimising the Use of Real-World Data to Estimate Treatment Effects

Emily Granger, London School of Hygiene & Tropical Medicine

Cystic fibrosis (CF) is an inherited disease affecting over 11,000 people in the UK. A key challenge for the CF community is that there are important clinical research questions about the effects of treatments used in practice that remain unanswered but may never be addressed in randomised controlled trials (RCTs), due to feasibility and cost. These include questions about effects of long-term treatments and treatments used in combination. An alternative to RCTs in this situation is to use real-world data. Target trial emulation provides a framework for making best use of real-world data to study treatment effects, while helping to avoid common biases that often occur in real-world data analyses. Target trial emulation involves first describing the RCT we would like to conduct – the "target trial" – and then specifying how each element of the target trial protocol will be emulated using real-world data. This approach is increasingly used across many disease areas, but its use in CF is so far limited.
In this talk, I will discuss some of the ongoing work on target trial emulation in the field of CF. This includes studies which aim to emulate existing trials in CF and assess to what extent the results from the trial can be replicated using registry data; and studies which aim to use target trial emulation to address questions that are unlikely to be answered in RCTs.

The Exposure of Syphilis During Pregnancy on Childhood Hospital Admissions in Brazil

Orlagh Carroll, London School of Hygiene & Tropical Medicine

Syphilis is a bacterial infection which can be transmitted sexually or from mother to child during pregnancy. There are high levels of syphilis in Brazil and its impact on pregnancy is a public health concern as resulting adverse birth outcomes include miscarriage, pre-term birth, and death. Using live births from routinely collected Brazilian data, we compared hospital admissions in children under five years old who had been exposed to syphilis in pregnancy and those who were not. This included time to first hospitalisation, length of stay of their admission and the ICD-10 codes associated with these records. We also looked at time to repeat hospitalisation and death. We found that children who were exposed during pregnancy and born with syphilis (congenital syphilis) had over a 6-fold hazard of first hospitalisation while those who were exposed and not infected had over a 2-fold increase of hospitalisation, compared to those not exposed during pregnancy. Any syphilis exposure during pregnancy was associated with longer hospital stays, an increased hazard of repeat admissions and an increased hazard of death when compared to children who were not exposed. This highlights the importance of syphilis prevention in women pre-conception and the need for careful monitoring of exposed children in their early years.

Modelling Disparities in Maternal RSV Vaccines

Em Prestige, London School of Hygiene & Tropical Medicine

Respiratory Syncytial Virus (RSV) is a respiratory virus which leads to a large healthcare burden amongst infants, especially those below six months old. This burden has been shown in other populations to be disproportionately distributed amongst certain ethnic and socioeconomic groups. One way this burden can be relieved is through maternal vaccinations, which are due to be implemented in England in Autumn 2024. However, maternal vaccines are shown to have varying uptake, with those of white ethnicity and those from higher socioeconomic groups being more likely to be vaccinated. The combination of higher burden and lower likelihood of maternal vaccination could lead to the widening of health disparities in a number of groups. Throughout my PhD I will first quantify the burden of RSV infections in infants in England, and the disparities in this burden focusing on ethnicity, socioeconomic status, and household composition. From these results I will then explore how these burdens may be exacerbated by the upcoming rollout of the maternal RSV vaccine. I will use mathematical modelling to explore a number of scenarios to explore how this can be averted, specifically looking at the required uptake needed in groups of interest in order to mitigate worsening disparities.

IDWSDS 2024

Session Info

S41

EARLY CAREER SESSION AIMED AT SHOWCASING POSTDOCTORAL RESEARCH

October 8th, 16:00 - 17:00 UTC

Showcasing Research by Postodoctoral researchers from MRC Biostatiscts Unit (Efficient Study Design)

CLICK HERE TO ACCESS THE ZOOM ROOM

This session will allow 3 female researchers to showcase outcomes from the work as postdoctoral researchers at the MRC Biostatistics Unit. These researchers are part of the bigger research group on the theme of Efficient Study Design (led by Dr. Villar), which develops and applies novel statistical designs to the improvement of clinical trials. The titles of the talks are provided below:

The talks included are 3:
Implementing Bayesian Response Adaptive Randomisation in a rare disease setting. by Dr. Rajenki Das
Eight Methodological Questions for bringing Digital Outcome Measures into Clinical Trials BY Dr. Mia Tackney
On the finite-sample and asymptotic error control of a randomization-probability test for response-adaptive designs by Dr. Nina Deliu

ORGANIZER: Sofia S. Villar, MRC Biostatistics Unit (University of Cambridge)
CHAIR: Sofia S. Villar, MRC Biostatistics Unit (University of Cambridge)
SPONSOR: MRC Biostatistics Unit - University of Cambridge

Speaker Bios

DR. RAJENKI DAS

MRC Biostatistics Unit (University of Cambridge)

Rajenki Das is a research associate at the MRC Biostatistics Unit, University of Cambridge where she is working on the development of the statistical analysis plan and implementation of a Bayesian Response Adaptive Randomisation design for the StratosPHere study. Prior to working on clinical trials, she focussed on modelling longitudinal mobile health data understanding underlying digital phenotypes. Her current key interest lies in implementing and testing trial designs in real life scenarios accounting for different challenges. She is particularly interested about the intersection of Mathematics and Statistics with Health, with an inclination towards advancing research in mental health.

DR. MIA TACKNEY

MRC Biostatistics Unit (University of Cambridge)

Mia Tackney is a Research Associate at the MRC-Biostatistics Unit. Her research focuses on statistical methodology for using digital devices to measure outcomes in trials, and she has a particular interest in missing data methods.

DR. NINA DELIU

MEMOTEF, Sapienza University of Rome

Nina is a Researcher in Statistics at MEMOTEF, Sapienza University of Rome, and a Visiting Researcher at the MRC - BSU, University of Cambridge. She is part of the editorial board of YoungStatS, the blog of Young Statisticians Europe, and has curated the Women in Statistics and Data Science Twitter account. Nina's primary interest is devoted to sequential decision-making problems, intersecting areas of inference, Bayesian statistics, multi-armed bandits, and the design of experiments. Her research is inspired by challenges arising in real-life and she strongly believes that the methodological progress should go along with the concrete real-world needs, and be "not simply good, but also good for something".

DR. SOFIA S. VILLAR

MRC Biostatistics Unit (University of Cambridge)

https://www.mrc-bsu.cam.ac.uk/staff/sofia-villar

Since September 2020, I have been an MRC Investigator (Programme Leader) working as part of the Efficient Study Design (ESD) theme. My research aims to improve clinical trial design through the development of innovative methods that lie in the intersection between optimisation, machine learning and statistics. These methods may result in efficiency gains (i.e. smaller or faster trials) but face several practical barriers (e.g. a high computational cost) to be widely adopted. I have a Ph. D. in Business Administration and Quantitative Methods at Universidad Carlos III de Madrid in July 2012, with a focus on Stochastic Dynamic Optimization. In 2013, I joined the MRC Biostatistics Unit (BSU) in Cambridge as part of a project on the design of multi-arm multi-stage clinical trials. In 2014, I was awarded the first ever Biometrika post-doctoral fellowship. I co-lead the Adaptive Designs Working Group part of the MRC-NIHR Trials Methodology Research Partnership (TMRP).

Implementing Bayesian Response Adaptive Randomisation in a Rare Disease Setting

Rajenki Das, MRC Biostatistics Unit (University of Cambridge)

Response-adaptive randomisation (RAR) designs are valuable as they increase likelihood of allocations to the most promising arm but keeping randomisation as the allocation method. However, this remains a challenge when the trial size is very small. Motivated by a phase 2 trial in a rare disease setting, we propose a Mapping strategy to address the question - how can we avoid undesirable treatment allocations per randomisation stage while staying true to the essence of RAR? To tackle this, we add an intermediate step of “Mapping” to the trial design. This involves a decision rule by introducing probability thresholds, to map the continuous randomisation probabilities produced at the interim stage to a target vector of discrete randomisation ratios. On comparing the performance of mapped designs with the original design of the trial StatosPHere 2 by evaluating power, type I error and expected allocation probabilities under the alternative hypothesis of an optimal arm, we find that implementing mapping helps us achieve efficiency both in terms of statistical performance and practical patient benefits. Thus, the Mapping approach not only helps avoid unexpected allocations, which is essential especially when the sample size is small, but also serves as a strategy to improve small-sample trials while staying true to the RAR design.

Eight Methodological Questions for bringing Digital Outcome Measures into Clinical Trials

Mia Tackney, MRC Biostatistics Unit (University of Cambridge)

The use of Digital Health Technologies to measure outcomes in clinical trials opens new opportunities as well as methodological challenges. Digital outcome measures may provide more sensitive and higher-frequency measurements, but pose vital statistical challenges around how such outcomes should be defined and validated and how trials incorporating digital outcome measures should be designed and analysed. This talk will present eight methodological questions, exploring issues such as the length of measurement period, choice of summary statistic and definition and handling of missing data, as well as the potential for new estimands and new analyses to leverage the time series data from digital devices.

On the Finite-Sample and Asymptotic Error Control of a Randomization-Probability Test for Response-Adaptive Designs

Nina Deliu, MEMOTEF, Sapienza University of Rome

This work addresses existing inferential challenges in response-adaptive designs, specifically the lack of type-I error guarantees and power efficiency, especially in finite samples. We show how an innovative test statistic, defined on the randomization probabilities of the adaptive design, can achieve improved frequentist error control while also preserving the expected outcome optimality. The finite-sample and asymptotic guarantees of the test are studied both in general settings and under a Bayesian response-adaptive design, which is commonly used both in clinical trials and beyond (e.g., recommendation systems or mobile health).

IDWSDS 2024

Session Info

S42

TECHNICAL SESSION

October 8th, 16:00 - 17:00 UTC

From Animal Health & Welfare to Plant & Crop Science: Contributions from Statisticians and Modellers at BioSS

CLICK HERE TO ACCESS THE ZOOM ROOM

BioSS data scientists contribute quantitative skills to a wide range of scientific fields, mostly in four broad activity areas: animal health and welfare, plant and crop science, ecology and environment, and human health and nutrition. Three experts from the organisation will present their latest work.

ORGANIZER: Altea Lorenzo-Arribas, Biomathematics and Statistics Scotland (BioSS)

CHAIR: Altea Lorenzo-Arribas, Biomathematics and Statistics Scotland (BioSS)

SPONSOR: Biomathematics and Statistics Scotland (BioSS)

Speaker Bios

KAREN KEEGAN
Moredun, The Rowett Institute, BioSS

https://www.bioss.ac.uk/people/kkeegan

I am a final year PhD student based at the Moredun Research Institute and my supervisors are Dr Eleanor Watson, Dr Nuno Silva and Dr David Longbottom (Moredun), Dr Karen Scott at the University of Aberdeen and Dr Nick Schurch at Biomathematics and Statistics Scotland (BioSS). With my PhD on 'Nanopore sequencing to investigate zoonosis and antibiotic resistance at the wildlife – livestock interface', I am focussing on evaluating the use of Nanopore sequencing (a third-generation sequencing technology) to help detect and monitor potential transfer of foodborne zoonotic pathogens and antimicrobial resistance (AMR) between wildlife and livestock. The project aims to evaluate Nanopore sequencing as a tool to monitor 1) potential on-farm transmission of zoonotic pathogens and AMR between cattle and wild geese and 2) microbial contamination and AMR in the marine environment, through the sampling of grey seals which can act as a sentinel species for marine ecosystem health.

GRACIELA MARTINEZ
BioSS

https://www.bioss.ac.uk/people/gmartinez

Graciela did her master's degree in mathematics at National Autonomous University of Mexico (2018). During her career she worked as Social Researcher in a private company. Her recent work focuses on providing statistical advice in Social Research with the collaboration of researchers at the James Hutton Institute.

Optimising Avian Faecal Samples for Gut Microbiome Research via Nanopore Sequencing

Karen Keegan, Moredun, The Rowett Institute, BioSS

Karen will present this important work based on data from the Orkney Isles and part of her PhD on "Nanopore sequencing to investigate zoonosis and antibiotic resistance at the wildlife-livestock interface."

Understanding By-Product Material Flows in the UK Seafood Industry to Estimate the Potential for Adding Value

Graciela Martinez, BioSS

In this study, we quantify the processing of by-product flows in the UK, which is the prerequisite for exploring how current and future uses and markets can be. To this end, we use annual landing statistics and UK trade data published in the UK Sea Fisheries Annual Statistics Report 2022 to estimate the amount of unprocessed seafood that remains in the UK. With historical information from 2012 to 2022, we calculate the annual average of imports, exports and landed weight for the main seafood species caught within the UK. Then, we use technical literature to estimate the amount of by-product from each key species under standard processing practices. For each of the main by-product categories that would be suitable for different uses, we discuss the potential for adding value under different scenarios and discuss the potential share that remains in Scotland using secondary information sources. Our results may inform the waste management debate and the investment decisions of innovative facilities, and other initiatives that would add value to seafood by-products, thus contributing to a more circular economy.

IDWSDS 2024

Session Info

S43

TECHNICAL SESSION

October 8th, 16:30 - 17:00 UTC

Chart-Toppers and Cliffhangers: Stats in Pop Culture

CLICK HERE TO ACCESS THE ZOOM ROOM

Join us for an exciting exploration of how statistics illuminate the world of pop culture! First, Anh Nguyen will dive into the meteoric rise of Taylor Swift through a survival analysis of her Billboard Hot 100 charting times, revealing her journey from an aspiring small-town country singer to a pop icon. Then, switch gears with to the small screen as Ashley Mullan analyzes the impact of pivotal romantic moments in television, like the much-anticipated kiss between Nick and Jess on New Girl, on show ratings. Discover how data can quantify audience reactions and uncover the trends behind your favorite songs and TV moments. Whether you're a Swiftie, a sitcom fan, or a data enthusiast, this session blends music, television, and statistics.

ORGANIZER: Sarah Lotspeich, Wake Forest University

CHAIR: Daniel Beavers, Wake Forest University

SPONSOR: Wake Forest University Department of Statistical Sciences

Speaker Bios

ANH NGUYEN
Wake Forest University

Anh Nguyen is currently a Visiting Assistant Professor at Wake Forest University. She earned her M.S. in Statistical Sciences from Wake Forest University in 2022. She has enjoyed working on managing clinical trial data. She is currently working on applying Bayesian approach to estimating missing distance data in a two-phase design with Dr. Sarah Lotspeich.

ASHLEY MULLAN
Vanderbilt University Medical Center
https://ashleymullan.github.io/

Ashley Mullan is a Ph.D. student in the Department of Biostatistics at Vanderbilt University Medical Center. She earned her master's degree in Statistics and a certificate in Data Science from Wake Forest University, and she has bachelor's degrees in Applied Mathematics and Philosophy from the University of Scranton. Ashley's current research interests include developing methods for missing or misclassified data and applying them to questions in public health. In her free time, she enjoys reading, watching, and writing about romcoms.

Look What You Made Me Do: A Taylor Swift's Singles Charting Time Analysis

Anh Nguyen, Wake Forest University

The Billboard Hot 100, released every Saturday, tracks sales, media plays, and streaming for songs every week. We know all too well that Taylor Swift is its chart-topping queen, with 188 charted songs and 24 top-5 hits. 'Midnight,' a recent album release, dominated the Billboard Hot 100, occupying all the top-10 positions and charting all 20 songs. As Taylor continues her iconic 'Eras Tour,' we can reflect on her eras through the lens of survival analysis. By analyzing the time from song release to charting on the Billboard Hot 100, we can better understand Taylor's rise from an aspiring small-town country singer to a pop icon. With data from the 'TayloR' R package, we use a Kaplan-Meier estimator to capture the overall time to charting for all songs and by album era. The era-specific analysis suggests that more recent eras had shorter median times to charting. Since Taylor is famous for her relatable and immaculate songwriting, we also consider whether "speechiness" (a Spotify measurement for the amount of spoken words in a song) impacted time to charting. Using a Cox hazard proportional model, we found that songs with more spoken words tend to chart more quickly. With her recent rise in popularity as evidenced by shorter charting time and her impactful lyrics, we see that Taylor Swift is on her way to cementing herself as a legend in the contemporary music sphere.

For Better or For Worse: The First Kiss Effect on Television Ratings

Ashley Mullan, Vanderbilt University Medical Center

Picture your favorite “will they or won’t they" television couple. From the first episode, they’ve got insane chemistry. Their dynamic constantly shifts between almost-romance and friendship, and you’re just sitting there, binge watching, waiting for them to just KISS ALREADY! One such duo is Nick Miller and Jess Day from the show New Girl, and that big moment comes in Season 2. For these couples, the first kiss is a pivotal moment in their character arcs, and it defines the trajectory of the plot. Some viewers are hooked and excited to see the upcoming drama, so they grab the popcorn and don’t leave the couch for at least another half season. Others might be susceptible to the Zeigarnik effect and see that kiss as the completion of the story arc. They’ll then get bored and either immediately abandon the show or slowly fizzle out when they lose interest. This dichotomy of behavior poses an interesting question to those data-minded Netflix fans. Luckily, ratings can quantify how the audience feels about these kisses! In this talk, we seek to discover the effect of a couple’s first kiss on a show’s ratings.

IDWSDS 2024

Session Info

S44

TECHNICAL SESSION

October 8th, 17:00 - 17:30 UTC

Causal Inference Methods for Evaluating Surrogate Markers

CLICK HERE TO ACCESS THE ZOOM ROOM

This session will focus on distinct methods for surrogate marker evaluation. In clinical trials, it's common to assess the impact of a treatment on an intermediate outcome, especially when the primary outcome is challenging or expensive to measure. Speakers will focus on formal ways to evaluate these outcomes as potential surrogates using causal inference methods. Each will describe their research of novel methods including principal stratification and individual causal association metrics.

ORGANIZER: Emily Roberts, University of Iowa

CHAIR: Emily Roberts, University of Iowa

Speaker Bios

FENNY ONG

Hasselt University

https://researchers.mq.edu.au/en/persons/ayse-bilgin

Fenny Ong is a PhD student and researcher at Hasselt University in Belgium. Her dissertation work focuses on causal inference methods to evaluate surrogate endpoints.

EMILY ROBERTS

University of Iowa

https://emilykroberts.github.io

Emily Roberts is an assistant professor at the University of Iowa. She completed her PhD in biostatistics from the University of Michigan in 2022.

An Information-Theory Approach for the Evaluation of Continuous Surrogate for Binary True Endpoints Based on Causal Inference

Fenny Ong, Hasselt University

The development of methods to validate surrogate endpoints remains an active and intensely researched areadue to its importance in expediting the process of clinical trials of a large number of new promising treatments. A general definition and quantification of surrogacy based on the concept of information theory has been introduced to provide a unified framework for such evaluation. First introduced in the meta-analytic framework, this approach has been expanded to the single-trial setting within the causal inference framework. In particular, it has been developed in the setting where both surrogate and true endpoints are either continuous or binary outcomes. The current study aims to extend the approach to the setting where the true and surrogate endpoints are binary and continuous, respectively.

The surrogacy metric developed in this context, the individual causal association (ICA), quantifies the association between the individual causal treatment effects on the true and surrogate endpoints. The metric is generic and model-independent, but in this study, we propose to define the underlying causal inference model as the decomposition of the joint distribution of the potential outcomes associated with the putative surrogate and the true endpoint of interest. The identifiability issue inherent to this type of model is handled via sensitivity analysis. The methodology is evaluated via simulations and is further illustrated in a real case study.

Using Principal Stratification for Surrogate Evaluation with Mixed Models

Emily Roberts, University of Iowa

Before taking advantage of a potential surrogate endpoint in a clinical trial, it is critical to carefully determine whether a candidate marker is valid for future use. This work builds on established causal inference approaches for evaluating a candidate surrogate using potential outcomes when measures of both outcomes are repeatedly collected. We do so by using the principal stratification framework and linear mixed models. This proposed framework allows us to calculate a causal effect predictiveness curve based on the distribution of random effects for both the surrogate and clinical endpoints and observed and counterfactual outcomes. Our work is primarily motivated by a diabetes clinical trial which investigated a potential treatment to protect beta cells in new onset Type 1 Diabetes patients.

IDWSDS 2024

Session Info

S45

TECHNICAL SESSION

October 8th, 17:00 - 18:00 UTC

Recent advances in Bayesian methods for covariance estimation and network data

CLICK HERE TO ACCESS THE ZOOM ROOM

Modern datasets often involve structured and high-dimensional data showcasing complex dependencies, posing challenges for standard statistical methods. On the contrary, Bayesian approaches offer significant advantages in addressing these complexities. This session features three talks on the latest advances in Bayesian methods and computational techniques for complex data.

Dr. Elizabeth Bersson will discuss covariance estimation for high-dimensional data, incorporating dimension reduction via latent factors and allowing for shrinkage towards structures learned from feature-level explanatory covariates.

Prof. Deborah Sulem will introduce a fully Bayesian method for estimating Gaussian graphical models. Graphical modeling is a widely used technique for analyzing the partial dependence structure among variables. In undirected models, the partial dependency structure informs on direct associations between variables.

Dr. Martina Amongero will present a probabilistic framework for analyzing network data. In network analysis, community detection is one of the most interesting problems. From a model-based perspective, the most studied approach is the one of stochastic block models, which typically do not entail the assortative behavior Martina explores in her work.

ORGANIZER: Beatrice Franzolini, Bocconi University

CHAIR: Beatrice Franzolini, Bocconi University

SPONSOR: j-ISBA The junior section of the International Society for Bayesian Analysis (ISBA)

Speaker Bios

DR. ELIZABETH BERSSON

Duke Univeristy and MIT

https://betsybersson.github.io/

Elizabeth (Betsy) received her PhD in May 2024 from the Department of Statistical Science at Duke University under the supervision of Peter Hoff. She is joining Tamara Broderick's group at MIT as a postdoctoral fellow in Fall 2024. Her work is related to hierarchical modeling, covariance estimation, small area estimation, and conformal prediction.

PROF. DEBORAH SULEM

Università della Svizzera italiana

https://dsulem.github.io/

Deborah is an Assistant Professor of Data and Statistical Sciences at the Universita' della Svizzera Italiana. She received her Ph.D in 2023 from the Department of Statistics at the University of Oxford and was previously a postdoctoral researcher at the Barcelona School of Economics (Universitat Pompeu Fabra). Her research interests include Bayesian inference, network analysis, graph deep learning, and algorithmic and statistical fairness.

DR. MARTINA AMONGERO

University of Torino

https://www.esomas.unito.it/do/docenti.pl/Alias?martina.amongero

Martina is a statistician in Torino (Italy). She graduated in Mathematics for Engineering in 2017 and in Mathematical Engineering in 2019 at Politecnico di Torino. She obtained her Ph.D. in Pure and Applied Mathematics in 2024 at the Department of Mathematical Sciences of the Politecnico di Torino working in biostatistics, with a scholarship sponsored by GSK Vaccines. She is now working as a Postdoc at the Department of Economics, Social Studies, Applied Mathematics, and Statistics of the University of Torino, focusing on Bayesian mixture models for community detection.

Covariance Meta Regression

Elizabeth Bersson, Duke Univeristy and MIT

This work considers the task of covariance estimation for high dimensional data consisting of a large amount of features relative to the number of samples. Standard approaches to such covariance modeling include imposing structural assumptions based on auxiliary information regarding the features or utilizing unsupervised dimension reduction with latent factors. In this work, we present a prior distribution for a covariance matrix that incorporates dimension reduction via latent factors and flexibly allows for shrinkage towards a structure learned from feature-level explanatory covariates, or, meta covariates. As with classical regression analysis, the proposed prior can flexibly utilize covariates of mixed types such as categorical and continuous. If no meta covariates are available, we show the proposed prior more accurately estimates non-diagonal population covariance matrices than standard alternatives. One natural application of this work is in modeling chemical exposures where broad chemical classes may be labeled for convenience rather than scientific differentiation. In this application, a researcher may wish to allow for robustness to class label in order to more accurately represent an across-chemical covariance matrix instead of assuming conditional independence. We demonstrate the utility of this prior in jointly modeling chemical exposures in the NHANES data set.

A Fast Bayesian Approach to Sparse Graphical Modelling With Many Variables

Deborah Sulem, Università della Svizzera italiana

Sparse graphical models provide an interpretable approach to recover the most significant partial dependencies in contexts with many observed variables. Bayesian methods can enforce sparsity and incorporate information on the variables through the prior distribution. They also provide uncertainty on the estimated graphical models, a feature that is particularly useful in applications where the sample size is small and the number of observed variables is high. However, in this context, even under the assumption of joint Gaussianity of the variables, the computational demands of Bayesian algorithms have limited their scope of applications. In this work, we introduce a scalable, interpretable, and fully Bayesian method for estimating a Gaussian graphical model. Our method capitalises on a discrete spike-and-slab parametrisation of the prior distribution, leading to a truly sparse estimated graphical model. We propose an efficient algorithm to sample from the posterior distribution, together with an almost-parallel version that exploits the relationship between the conditional dependence structure and a linear regression model. This strategy facilitates decomposing the high-dimensional estimation problem into sub-components, allowing the application of efficient methodologies originally developed for linear regression. We empirically demonstrate that statistical and computational efficiencies are improved by our discrete parametrisation.

Bayesian Community Detection for Assortative Networks

Martina Amongero, University of Torino

Data available in the form of networks are gaining increasing attention in modern research, with a key interest in community detection, which involves dividing the nodes into clusters. For this purpose, the Stochastic Block Model (SBM) provides a well-suited generative process for explaining the formation of communities. A recent line of work uses Bayesian methods for the recovery of communities in classical SBM by placing a prior distribution on the number of clusters k and estimating cluster assignments with collapsed Gibbs samplers. This work focuses on improving the prior for community detection by including assortativity in SBM. In particular, an SBM is assortative when the probability of a connection is higher for nodes belonging to the same cluster rather than different clusters. In Bayesian SBM, assortativity is usually sacrificed to preserve conjugacy, which is key to devising collapsed Gibbs samplers. While minimax theory suggests that estimation and community detection in assortative SBM are no more difficult than in SBM, it is worth exploring if posterior inference benefits from an assortative prior, considering the computational cost of losing conjugacy. We propose a probabilistic framework for the simultaneous estimation of k and the labels for assortative SBM and illustrate posterior inference by suitably modifying existing collapsed Gibbs samplers. Synthetic data are considered in comparison with existing algorithms.

IDWSDS 2024

Session Info

S46

TECHNICAL SESSION

October 8th, 17:00 - 18:00 UTC

Innovations in Precision Medicine and Optimal Treatment Strategies

CLICK HERE TO ACCESS THE ZOOM ROOM

Historically, healthcare recommendations have followed a 'one-size-fits-all' approach, treating patients as if they were the 'average' individual. In recent years, there has been an increase in the popularity and development of precision medicine techniques, which tailor prevention and treatment strategies to each patient's unique characteristics, leading to more effective and personalized care. During our session, we will delve into novel approaches that bridge both statistical and medical gaps in the field of precision medicine. Christina Zhou will begin by sharing her precision medicine methodology for survival data in the presence of competing risks, applied to peripheral artery disease patients. Dr. Daiqi Gao will then share her development of a new reinforcement learning algorithm for a mobile health clinical trial to improve patient commitment to physical activity. Finally, Dr. Yuanjia Wang will share her work on incorporating intermediate outcomes into precision medicine rules in the context of mental health disorders.

ORGANIZER: Marissa Ashner, Duke University

CHAIR: Marissa Ashner, Duke University

Speaker Bios

CHRISTINA ZHOU
University of North Carolina at Chapel Hill

Christina Zhou is a doctoral student in the Department of Biostatistics at the University of North Carolina at Chapel Hill. She is a member of the Precision Health and AI Research (PHAIR) Lab and is working under the advisement of Dr. Michael Kosorok. Her research focuses on developing precision medicine methods for survival data to estimate optimal individualized treatment regimes. In today's talk, she will present her research for novel precision medicine methodology for survival data with competing risks.

DR. DAIQI GAO
Harvard University

Daiqi Gao is a postdoctoral fellow in the Department of Statistics at Harvard University, working with Professor Susan Murphy. She received her Ph.D. from the Department of Statistics and Operations Research at the University of North Carolina at Chapel Hill, where she was advised by Professors Yufeng Liu and Donglin Zeng. Previously, she earned her B.S. in Industrial Engineering and Statistics from Tsinghua University. Her primary research interests are in statistical reinforcement learning and machine learning, with applications in mobile health and personalized medicine.

WENBO FEI
Columbia University

Wenbo Fei is a PhD candidate in the Department of Biostatistics at Columbia University. She works with her advisor, Yuanjia Wang, to develop statistical methods that facilitates precision medicine by leveraging evidence from multiple data sources.

Estimating Optimal Individualized Treatment Regimes for Survival Data With Competing Risks

Christina Zhou, University of North Carolina at Chapel Hill

For more than a decade, using precision medicine (PM) to determine a patient's optimal treatment has risen in popularity over the traditional "one-size-fits-all" treatment assignment. Extensive methodology for estimating individualized treatment regimes (ITRs) developed to account for individual heterogeneity. Although PM methods for survival data have become more abundant in recent years, less focus is given to estimating ITRs in the presence of competing risks (CR). CR are events where their occurrence precludes the occurrence of other events, and not accounting for them can lead to biased results. Because CR are prevalent in healthcare settings, it is crucial to examine the risk of a specific event for treatment planning in addition to overall survival for accurate prognosis and effective treatment planning. Thus, we develop nonparametric ITR estimation methodology by extending generalized random survival forests into the CR setting and introducing a multi-utility value function. We propose a two-phase method that accounts for both overall survival from all events as well as the cumulative incidence of the main event of interest (i.e., a priority cause). Simulation studies show that our proposed method works well, and we apply the proposed method to a cohort of peripheral artery disease patients.

Harnessing Causality in Reinforcement Learning for Slowly Evolving Rewards

Daiqi Gao, Harvard University

Mobile health (mHealth) provides effective ongoing support in everyday life to help users sustain a healthy lifestyle. In a new mHealth clinical trial, we aim to improve users' commitment to physical activity (PA), which is a slowly evolving reward. A personalized policy decides whether and when notifications are pushed to mobile devices to prompt short bouts of activity. However, the reward is sparsely observed through weekly surveys and the effect size of each action is relatively small, creating challenges for efficiently learning the policy. We prepare for the new trial by developing a reinforcement learning algorithm that sequentially updates the personalized policy based on these sparsely observed rewards. To speed up learning, causal information provided by domain experts is leveraged for imputing missing rewards, reducing state space, constructing a prior, and reward engineering. We build a simulation testbed using real data from the HeartSteps clinical trial and evaluate the proposed method on different variations of the testbed.

Machine Learning Methods for Optimal Early Decision Treatment Rules with Multi-domain Intermediate Outcomes

Wenbo Fei, Columbia University

Adopting precision medicine for mental disorders presents challenges due to disease complexity and heterogeneity in patient responses. Empirical studies suggest that early indicators, such as interim measures (e.g., patient self-reports) of disease improvement or relapse, can predict longer-term outcomes, serving as proxies when final outcomes (e.g., in-clinic assessments) are hard to obtain. However, existing approaches for deriving individualized treatment rules (ITRs) often ignore these early indicators, focusing instead on a final outcome as the reward. In this work, we propose a new method that incorporates intermediate outcomes from various domains into a personalized composite outcome, serving as the reward for learning ITRs. This composite is a weighted sum of inferred latent states from observed measures, with weights personalized for each patient, ensuring it mirrors the long-term response. Our simulations show that this approach not only provides timely detection of non-responders at an early stage but can also improve long-term treatment outcomes. Applying our framework to a randomized clinical trial on major depressive disorder (MDD) demonstrates its utility and advantages in ITR learning.

IDWSDS 2024

Session Info

S47

TECHNICAL SESSION

October 8th, 17:30 - 18:00 UTC

Statistical Analysis Plans: An Overview and Practical Guidance for Enhancing Research Rigor

CLICK HERE TO ACCESS THE ZOOM ROOM

Statistical Analysis Plans (SAPs) are vital for enhancing the rigor and reproducibility of scientific studies. This session offers an overview of SAPs, focusing on their role and importance in research, with a particular emphasis on observational study designs.

The session will begin by defining SAPs and explaining how they serve as roadmaps for the statistical methods of research studies. The session will cover the structure and key components of SAPs. A template created by the presenter, drawing from established guidelines, best practices, and real-world experiences, will be demonstrated with examples that highlight its application in observational studies. Additionally, the current landscape of SAPs will be examined, highlighting emerging trends and variations in their adoption across different research settings.

By the end of this session, attendees will gain an understanding of the role of SAPs in promoting research rigor and reproducibility and will gain practical insights into how to use SAPs to enhance the quality of research.

ORGANIZER: Hunna Watson, Department of Psychiatry, School of Medicine, University of North Carolina at Chapel Hill

CHAIR: Hunna Watson, Department of Psychiatry, School of Medicine, University of North Carolina at Chapel Hill

SPONSOR: Department of Psychiatry, University of North Carolina at Chapel Hill

Speaker Bios

DR. HUNNA WATSON
University of North Carolina at Chapel Hill
https://www.linkedin.com/in/hunna-watson-52030352/

Hunna Watson is a Research Associate Professor in the School of Medicine at the University of North Carolina at Chapel Hill and biostatistician at the Center of Excellence for Eating Disorders. She also holds Adjunct Research Fellow appointments at The University of Western Australia and Curtin University in Australia. Her research focuses on eating disorders, including diagnostic nosology, prevention, treatment, molecular genetics, and randomized clinical trials. She has authored over 120 peer-reviewed articles and book chapters, and her work has been presented globally through lectures, papers, and workshops. Associate Professor Watson serves on the editorial board of the International Journal of Eating Disorders and is a member of the Alumni Group of the National Eating Disorders Collaboration Steering Committee. She co-founded the Helping to Outline Paediatric Eating Disorders (HOPE) Project, a clinical eating disorder registry at Perth Children's Hospital in Australia.

IDWSDS 2024

Session Info

S48

TECHNICAL SESSION

October 8th, 19:00 - 20:00 UTC

Infection Insights: Women's Contributions to Infectious Disease Modeling

CLICK HERE TO ACCESS THE ZOOM ROOM

This session showcases pioneering work by women in infectious disease modeling, emphasizing innovative statistical methods that enhance our understanding and management of diseases. First, Dr. Natalie Dean will highlight the Emory Center for Infectious Disease Modeling and Analytics and Training Hub (CIDMATH), exploring how the center's initiatives in wastewater surveillance, contact pattern monitoring, and machine learning contribute to advancing public health. Her focus will be on CIDMATH's role in training the next generation of infectious disease modelers through collaborative programs and partnerships. Then, Dr. Staci Hepler will introduce a Bayesian spatio-temporal model for accurately estimating Coccidioides endemicity amidst challenges in disease detection. By linking data on Valley fever cases to environmental and geographical factors, her model provides insights into disease distribution and highlights areas needing improved surveillance. Finally, Dr. Lucy D'Agostino McGowan will address the challenges of pathogen transmission analysis, presenting methods for quantifying differential transmission using phylogenetic data. She will focus is on sample size calculations and statistical corrections to enhance the reliability of transmission studies, with practical tools provided via the R package phylosamp.

ORGANIZER: Sarah Lotspeich, Wake Forest University

CHAIR: Sarah Lotspeich, Wake Forest University

Speaker Bios

DR. FAN BU

University of Michigan

https://fanbu1995.github.io/

Dr. Fan Bu is a tenure-track assistant professor in Biostatistics at the University of Michigan, Ann Arbor (started January 2024). Previously, she was a postdoctoral research fellow at University of California – Los Angeles, working with Dr. Marc Suchard to develop Bayesian statistical methods for analyzing large-scale observational health data. Dr. Bu obtained my Ph.D. degree in Statistics from Duke University under the supervision of Dr. Alexander Volfovsky in Fall 2021. Her research interests include Bayesian statistics and computation for emerging and complex data, stochastic processes and dynamic models, and health data science and quantitative social science.

DR. STACI HEPLER

Wake Forest University

https://sites.google.com/a/wfu.edu/hepler/

Dr. Staci Hepler is Associate Professor and Associate Chair in the Department of Statistical Sciences at Wake Forest University. Her research focuses on spatio-temporal modeling and Bayesian computation, and her methodological work is motivated by applications in environmental science, ecology, and epidemiology.

DR. LUCY D. MCGOWAN

Wake Forest University

https://www.lucymcgowan.com/

Lucy D'Agostino McGowan is an assistant professor in the Department of Statistical Sciences at Wake Forest University. She received her PhD in Biostatistics from Vanderbilt University and completed her postdoctoral training at Johns Hopkins University Bloomberg School of Public Health. Her research focuses on analytic design theory, statistical communication, causal inference, and data science pedagogy. Dr. D'Agostino McGowan can be found blogging at livefreeordichotomize.com, on Twitter @LucyStats, and podcasting on the American Journal of Epidemiology partner podcast, Casual Inference.

DR. NATALIE DEAN

Emory University

https://nataliexdean.net/

Dr. Natalie Dean is Associate Professor of Biostatistics at Emory Rollins School of Public Health. She is an expert in infectious disease epidemiology and vaccine study design. She co-directs the Summer Institute in Statistics and Modeling in Infectious Diseases (SISMID) and the Emory Alliance for Vaccine Epidemiology (EAVE). Dr. Dean is a co-PI on the Emory Center for Infectious Disease Modeling and Analytics and Training Hub (CIDMATH) – one of thirteen centers funded by the CDC's Center for Forecasting and Outbreak Analytics to advance infectious disease modeling capacity in the US.

Abstracts

Inferring HIV Transmission Patterns from Viral Deep-Sequence Data via Latent Spatial Poisson Processes

Fan Bu, University of Michigan

Viral deep-sequencing technologies play a crucial role toward understanding disease transmission patterns, because the higher resolution of these data provide evidence on transmission direction. To better utilize these data and account for uncertainty in phylogenetic analysis, we propose a spatial Poisson process model to uncover HIV transmission flow patterns at the population level. We represent pairings of two individuals with viral sequence data as typed points, with coordinates representing covariates such as sex and age, and the point type representing the unobserved transmission statuses (linkage and direction). Points are associated with deep-sequence phylogenetic analysis summary scores that reflect the strength of evidence for each transmission status. Our method jointly infers the latent transmission status for all pairings and the transmission flow surface on the source-recipient covariate space. In contrast to existing methods, our framework does not require pre-classification of the transmission statuses of data points, instead learning them probabilistically through fully Bayesian inference. By directly modeling continuous spatial processes with smooth densities, our method enjoys significant computational advantages over previous methods that discretize the covariate space. In a HIV transmission study from Rakai, Uganda, we demonstrate that our framework can capture age structures in HIV transmission at high resolution and bring valuable insights.

Estimating Coccidioidomycosis Endemicity While Accounting for Imperfect Detection Using Spatio-Temporal Occupancy Modeling

Staci Hepler, Wake Forest University

Coccidioidomycosis, or Valley fever, is an infectious disease caused by inhaling Coccidioides fungal spores. Incidence has risen in recent years, and it is believed the endemic region for Coccidioides is expanding in response to climate change. While Valley fever case data can help us understand trends in disease risk, using case data as a proxy for Coccidioides endemicity is not ideal because case data suffers from imperfect detection, including false positives (e.g., travel-related cases reported outside of endemic area) and false negatives (e.g., misdiagnosis or underreporting). We proposed a Bayesian, spatio-temporal occupancy model to relate monthly, county-level presence/absence data on Valley fever cases to latent endemicity of Coccidioides, accounting for imperfect detection. We used our model to estimate endemicity in the western United States. We estimate high probability of endemicity in southern California, Arizona, and New Mexico, but also in regions without mandated reporting, including western Texas, eastern Colorado, and southeastern Washington. We also quantified spatio-temporal variability in detectability of Valley fever, given an area is endemic to Coccidioides. We estimated an inverse relationship between lagged 3- and 9-month precipitation and case detection, and a positive association with agriculture. This work can help inform public health surveillance needs and identify areas that would benefit from mandatory case reporting.

Power and Sample Size Calculations for Testing the Ratio of Reproductive Values in Phylogenetic Samples

Lucy D'Agostino McGowan, Wake Forest University

The quality of the inferences we make from pathogen sequence data is determined by the number and composition of pathogen sequences that make up the sample used to drive that inference. However, there remains limited guidance on how to best structure and power studies when the end goal is phylogenetic inference. One question that we can attempt to answer with molecular data is whether some people are more likely to transmit a pathogen than others. In this talk we will present an estimator to quantify differential transmission, as measured by the ratio of reproductive numbers between people with different characteristics, using transmission pairs linked by molecular data, along with a sample size calculation for this estimator. We will also provide extensions to our method to correct for imperfect identification of transmission linked pairs, overdispersion in the transmission process, and group imbalance. We validate this method via simulation and provide tools to implement it in an R package, phylosamp.

Building Community in Infectious Disease Training and Research

Natalie Dean, Emory University

During this presentation, I will talk about work from the Emory Center for Infectious Disease Modeling and Analytics and Training Hub (CIDMATH). This recently funded center within CDC's Insight Network supports innovation in forecasting and outbreak analytics. CIDMATH's areas of focus include wastewater surveillance, monitoring human contact patterns, machine learning for forecasting, and vaccine evaluation. This work is achieved via partnerships in the public sector (Georgia Department of Public Health) and the private sector (Kaiser Permanente Georgia). CIDMATH further supports training activities, particularly the Summer Institute in Statistics and Modeling in Infectious Diseases (SISMID). I will talk about building and sustaining a pipeline of infectious disease modelers and analysts.

IDWSDS 2024

Session Info

S49

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

October 8th, 19:00 - 19:30 UTC

Bridging the Gender Gap in Statistics Education: Strategies to Encourage More Women to Pursue Studies in Statistics and Data Science

CLICK HERE TO ACCESS THE ZOOM ROOM

The fields of statistics and data science are vital for the advancement of technology and society. Yet, despite their significance, women are still underrepresented in these fields. Statistics is one of the most dreaded courses for students around the globe especially in Nigeria, where men are more likely to pursue it more than women. Many women are discouraged by the perception that statistics is too difficult or not suited for them. This inequality is not limited to academic difficulty; it is also a result of cultural and societal norms, parental policies that subtly deter women from pursuing careers in data science and statistics. We can start to shift the narrative and encourage more women to pursue the study of statistics by addressing these underlying causes and developing more encouraging techniques for inspiring and supporting more women to pursue studies and careers in statistics and data science.

ORGANIZER: Christianah Olanrewaju, Teaching Assistant at Bowen University, Iwo, Osun State, Nigeria

CHAIR: Christianah Olanrewaju, Teaching Assistant at Bowen University, Iwo, Osun State, Nigeria

Speaker Bios

CHRISTIANAH OLANREWAJU

Bowen University, Iwo, Osun State, Nigeria

Christianah Olanrewaju is a First Class Graduate of Statistics from Ladoke Akintola University of Technology, Nigeria, recognized as the Second Best Graduating Student in her department. She is currently a Teaching Assistant at Bowen University, Nigeria, and a skilled Data Analyst with a passion for leveraging data to drive impactful decisions. Her research includes a publication on COVID-19 vaccine efficacy, highlighting her interest in Biostatistics. Christianah's research interest focuses on statistics education, statistical modeling, data visualization, and machine learning, particularly in healthcare. She founded "The Beauty of Statistics" to promote the relevance of statistics across sectors. At the International Day for Women in Statistics and Data Science Conference, she will discuss strategies to encourage more women to pursue studies in statistics and data science.

IDWSDS 2024

Session Info

S50

TECHNICAL SESSION

October 8th, 19:00 - 20:00 UTC

Empowered Voices: Brazilian Award Winning Women Share Their Research

CLICK HERE TO ACCESS THE ZOOM ROOM

Together with new, more efficient and powerful computing capabilities, within the context of Data Science, Statistics is playing a fast-growing role in business and society. This session intends to show how this may be done, through the presentation of some interesting case studies.

ORGANIZER: Gisela Tunes, USP - University of São Paulo

CHAIR: Renata Guerra, UFSM - University of Santa Maria

SPONSOR: Brazilian Statistical Association (ABE - Associação Brasileira de Estatística)

Speaker Bios

DR. DAIANE ZUANETTI

UFSCar - Federal University of São Carlos

Daiane holds a Bachelor's Degree in Statistics from the Federal University of São Carlos (2003), a Master's Degree in Statistics from the Federal University of São Carlos (2006) and a PhD in Statistics from the Inter-institutional Postgraduate Program of the Department of Statistics of the Federal University of São Carlos and the Institute of Mathematical Sciences and Computing of the University of São Paulo – PIPGEs (2016). She is currently an adjunct professor in the Department of Statistics of the Federal University of São Carlos and coordinator of PIPGEs. She has worked in the area of Probability and Statistics, with emphasis on the selection and estimation of independent or dependent mixture models (HMM), QTL mapping with familial dependence, parametric and non-parametric Bayesian inference and MCMC computational methods with application mainly in Genomics and Molecular Biology.

DR. ANA GABRIELA SILVA

IBGE

Ana Gabriela Faria da Silva works as Technologist at the Brazilian Institute of Geography and Statistics (IBGE). She holds Graduation in Statistics and Actuarial Sciences from the Federal University of Rio de Janeiro (UFRJ) and MSc in Statistics from the University of São Paulo (USP). Her studies are related to Machine Learning, Natural Language Processing, and the Classification of Economic Activities. She currently works on Statistical Methods for Economic Surveys. In 2023, Ana Gabriela was honored with the 2023 ISI Jan Tinbergen Award in Division B (Statistical Systems).

DR. CLARICE DEMÉTRIO

USP - University of São Paulo

Clarice G. B. Demétrio is a Professor at ESALQ, University of São Paulo (USP), Brazil. She has a BS in Agronomy, Masters and PhD in Applied Statistics in Agriculture from USP, and a Post-Doctoral training at Imperial College of Science and Technology, England, under David Cox supervision. She also has a Doctor Honoris Causa from Hasselt University, Belgium, 2019. Clarice teaches undergraduate and graduate courses and has published papers on various topics including generalized linear models and extensions applied to Agriculture. She got the “Herman Callaert Leadership Award in Biostatistical Education”, Hasselt University, Belgium in 2006; the award “Best Contributed Paper from a Special Circumstance for the Americas”, during the IBC2008; the “Premio Anual del Proyecto Juárez Lincoln Marti”, in 2009, the “Rob Kempton Award for Outstanding Contribution to the Development of Biometry in the Developing World, IBC2010, the Pesquisador 2022 Award, ABE, and the Destaque RBras Award, 2024.

The Efficiency of the Data-Driven Reversible Jump Algorithms for Model Selection

Daiane Zuanetti, UFSCar - Federal University of São Carlos

Despite the advantage of reversible jump algorithms in quantifying the uncertainty over different models tested and not requiring all potential models to be estimated and compared in a second stage through selection criteria, they have still been little used in recent decades due to their difficulty of implementation, bad mixing and slow convergence. In this talk, we discuss how data-driven (or guided or informed) reversible jump algorithms have shown easier and improved implementation and convergence both for selecting models with different structures (mixture models, for example) and for selecting variables in regression models.

Exploring the Use of Web Pages Texts, in Brazilian Portuguese, for Classifying Main Economic Activity of Companies

Ana Gabriela Silva, IBGE

This work has been developed over the years I have spent at the Brazilian Institute of Geography and Statistics (IBGE), the National Statistical Office of Brazil. In my MSc studies in Statistics, part of this was systematized and it was documented on paper. I cannot recall a step of this work that was carried out without the involvement of other women. As a result of this study, I was honored with the 2023 ISI Jan Tinbergen Award in Division B (Statistical Systems). Its purpose was to evaluate the use of supervised learning, in the context of text mining, to achieve the ational Classification of Economic Activities (CNAE) corresponding to the companies's main economic activity. Keeping the CNAE updated ensures better quality for the statistics produced by IBGE. The study used texts as predictor variables, obtained via web scraping, from business websites and URLs. Both URLs and the response variable, the CNAE, were derived from the Annual Business Surveys, from IBGE. Due to the hierarchical structure of the classification, two approaches were tested to fit the models: Flat classification and hierarchical classification. In both cases, the Logistic Regression classifier presented the best performance, being able to extract patterns fit to identify the classification. Despite both approaches' results were similar when considering all classes, the flat classifier apparently performed better in categories that tended to be more difficult to characterize in the higher levels.

Overdispersion Models for Clustered Toxicological Data in a Bioassay of Entomopathogenic Fungus

Clarice Demétrio, USP - University of São Paulo

We consider discrete mortality data for groups of individuals observed over time. The fitting cumulative mortality curves as a function of time involves the longitudinal modelling of the multinomial response. Typically such data exhibit overdispersion, that is greater variation than predicted by the multinomial distribution. To model the extra-multinomial variation (overdispersion) we consider a Dirichlet-multinomial model, a random intercept model and a random intercept and slope model. We construct asymptotic and robust covariance matrix estimators for the regression parameter standard errors. Applying this model to a specific insect bioassay of the fungus {\it Beauveria bassiana}, we note some simple relationships in the results and explore why these are simply a consequence of the data structure. Fitted models are used to make inferences on the effectiveness of different isolates of the fungus and results are compared with a simple empirical analysis to provide recommendations for the field use of this fungus as a biological control.
Note: joint work with Silvia M. Freitas, Lida Fallah, Clarice G.B. Demétrio (Speaker), John Hinde

IDWSDS 2024

Session Info

S51

HISTORY OF WOMEN STATISTICIANS SESSION

October 8th, 19:30 - 20:00 UTC

The Stories Behind the History of Women in Statistics

CLICK HERE TO ACCESS THE ZOOM ROOM

We will take a look at some of the hidden stories of women who changed the world and statistics.

ORGANIZER: Altea Lorenzo-Arribas, Biomathematics and Statistics Scotland (BioSS)
CHAIR: Cecilia Lanata Briones, Warwick University

SPONSOR: Royal Statistical Society Celebrating Diversity SIG and History of Statistics Section

Speaker Bios

PROF. PENNY REYNOLDS

University of Florida

https://expertfile.com/experts/penny.reynolds/penny-reynolds

Penny Reynolds is a statistician whose research focuses on statistical experimental design, methodology and sample size estimation. She is an expert on animal welfare and research reproducibility. She is an assistant professor of Anesthesiology in the College of Medicine.

DR. ALTEA LORENZO-ARRIBAS

BioSSi

https://www.bioss.ac.uk/people/altea

I work as a socio-economic statistician in interdisciplinary projects with researchers at the Scottish Environment, Food and Agriculture Research Institutions. My research interests include causal inference (as part of BioSS Large-scale and Systems Modelling research stream) and social responsibility of Artificial Intelligence (AI).
I am a panel member of the UKRI Interdisciplinary Assessment College and the Spanish Government Interdisciplinary Assessment Board of interdisciplinary and cross-disciplinary research programme in Artificial Intelligence. I am an elected council member of the Royal Statistical Society, a member of the Society's AI Task-Force, chair of the Celebrating Diversity Special Interest Group, and secretary of the History of Statistics Section. I am also a member of the Women Committee of the Spanish Society of Statistics and Operations Research, and the Spanish Biostatistics Network (Biostatnet).

"Well-Behaved Women Seldom Make History"....or Do They? Six Women You Never Heard of Who Changed the Statistics & Data Science Landscape Forever

Penny Reynolds, University of Florida

The history of statistics has neglected the contributions of women who fail to tick the conventional academic resumé checkboxes. The following six women nevertheless revolutionised the practices and procedures of statistics now taken for granted today. Beatrice Cave Brown Cave (1874-1947; UK) produced landmark papers on child development then switched to aeronautical engineering & revolutionised fixed-wing aircraft design. Kirstine Smith (1878-1939; Denmark) invented the field of optimal designs. Jessamine Whitney (1880-1940; USA), a pioneer of tuberculosis epidemiology and baseball sabermetrics. Frances Elizabeth (Betty) Allan (1905-1952; Australia) the first consulting statistician in Australia, whose work elevated CSIRO research to world class status. Mary Eleanor Spear (1897-1986; USA), a pioneer of data visualization methods & inventor of the box plot. Mary Gibbons Natrella (1922–1988; USA) a National Bureau of Standards statistician, her monumental Handbook of Experimental Statistics is still a key scientific and engineering reference on experimental design for researchers.

Suffrage, Statistics, and Spurious Correlations

Altea Lorenzo-Arribas, BioSSi

This talk will look at the connection between the Women's Suffrage and Statistics, and the important work of a number of suffragists who were directly involved in statistical organisations both in the United Kingdom and the United States.

IDWSDS 2024

Session Info

S52

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

October 8th, 20:00 - 21:00 UTC

Resilience, Innovation, and Impact: Women's Journeys in Statistics and Data Science

CLICK HERE TO ACCESS THE ZOOM ROOM

In this session, we will explore the remarkable journeys of three women who have forged successful careers in the male-dominated statistics and data science fields. Each presenter will share their unique path, from overcoming significant personal and professional challenges to becoming successful in their respective areas of expertise. Through their stories, we will hear how perseverance, adaptability, and a fearless approach to innovation enabled them to navigate transitions across industries, countries, and disciplines.

Join us for an inspiring session highlighting how resilience, determination, and innovation can lead to transformative careers in statistics and data science while creating a more inclusive and impactful future for all.

ORGANIZER: Shirin Jabbarzadeh, Emory University

CHAIR: Hyun-Joo Kim, Truman State University

Speaker Bios

DR. SHIRIN JABBARZADEH
Emory University

Dr. Shirin Jabbarzadeh is a public health professional with a comprehensive background in health research, communications, education, program evaluation, and clinical practice. With a Master of Science in Public Health Informatics from the Rollins School of Public Health at Emory University and a Medical Doctorate from Iran University of Medical Sciences, Dr. Jabbarzadeh combines clinical expertise with advanced public health informatics.
Currently serving as a Project Data Manager in the Department of Biostatistics and Bioinformatics at Emory University, Dr. Jabbarzadeh has led the data management team for numerous high-impact projects. Her role involves designing data collection tools and databases, supervising data collection, data quality control, and collaborating with investigators on data analysis and publishing results.

RUBA SHALHOUB
NIH

Ruba Shalhoub is a mathematical statistician at the Office of Biostatistics Research at the National Heart, Lung, and Blood Institute of the NIH. After receiving her BS in Biomedical Engineering from the University of Virginia in 2015, she decided to change course and pursue a degree in Biostatistics. She graduated with an MS in Biostatistics from Georgetown University in 2018, after which she completed a research fellowship at the NIH. In 2020, she officially joined the Office of Biostatistics Research, where she works today.

DR. MINA AMINGHAFARI
University of Calgary

Dr. Mina Aminghafari is an Associate Professor of Machine Learning at the University of Calgary and a professional statistician of the Statistical Society of Canada. Before this role, she served as an Associate Professor at Amirkabir University of Technology (2013 to 2023) and as an assistant professor at the same university (2006-2013). She has also held positions as a Senior Data Scientist in the education and insurance industries (2018-2023) in Canada.

PROF. HYUN-JOO KIM
Truman State University

Dr. Hyun-Joo Kim is a Professor of Statistics and currently serves as the Chair of the Health Science Department at Truman University. With a passion for education and student success, Dr. Kim has previously led the Statistics Department from 2016 to 2021 and the Computer Science Department from 2023 to 2024. As the Director of the Data Science Program, Dr. Kim has worked to develop a comprehensive graduate and undergraduate data science curriculum that meets the needs of students. Additionally, Dr. Kim is a co-founder of the Center for Applied Statistics and Evaluation, where they continue to contribute to the field through research and practical applications.

Abstracts

Bridging Clinical Medicine and Data-Driven Research: A Journey in Biostatistics and Public Health Informatics

Shirin Jabbarzadeh, Emory University

My career path has been a journey of continuous learning and adaptation, driven by a deep commitment to improving public health. I began my professional journey as a General Practitioner in Tehran, Iran, where I concentrated on mother and child health. This role was both rewarding and challenging, as it required me to address diverse health issues while navigating the complexities of a busy clinical practice. In addition to my medical duties, I served as a science editor for national newspapers and participated in evaluating medical education programs. These experiences not only enhanced my clinical skills but also provided me with valuable insights into health communication and the broader healthcare landscape.
Seeking to expand my impact, I moved to the United States to pursue a Master of Science in Public Health Informatics at Emory University. Transitioning to a new educational system and diving into the rapidly evolving field of public health informatics presented significant challenges. However, my determination to bridge the gap between clinical practice and data-driven public health initiatives kept me motivated. At Emory, I developed expertise in data management and biostatistics, which paved the way for my current role as a Project Data Manager in the Department of Biostatistics and Bioinformatics. Managing complex, multi-center studies has been both demanding and fulfilling, allowing me to contribute to influential research projects.

Unlocking Potential: Navigating Uncertainty and Exploring a Career in Biostatistics

Ruba Shalhoub, NIH

Finding a fulfilling career path as a young professional is a challenging undertaking, often compounded by uncertainty and indecision. That was certainly the case for me, having just graduated from college with an engineering degree and unsure what to do next. Through trial and error, I stumbled upon the field of biostatistics, and I haven't looked back since. It was a significant shift in focus but one driven by a deep desire to leverage quantitative skills in a new context. It was a move that allowed for engagement with a broad range of medical research questions and simultaneously scratched my itch for problem-solving.
My search for job opportunities consistent with these qualities eventually led me to a career in the federal government, collaborating with large teams of investigators on clinical trials and epidemiological studies. As a mathematical statistician working at the National Institutes of Health, I am involved in designing novel studies, analyzing complex data, interpreting results to support clinical and epidemiological research, and exploring different avenues for innovative statistical solutions to complex research questions. My experience working in the public sector has given me a profound sense of personal and professional fulfillment through advancing clinical research and health outcomes.
Join me as I discuss how I navigated uncertainty and leveraged existing skills to turn challenges into opportunities for growth.

Embracing the Journey: Perseverance, Innovation, and Breaking Barriers in Machine/ Statistical Learning

Mina Aminghafari, University of Calgary

In this session, I will share my personal and professional journey, emphasizing the importance of perseverance, fearlessness, and innovation in machine/statistical learning and beyond. Having navigated the challenges of transitioning to a new country and industry, I have learned that everything is possible when we embrace our destiny, think outside the box, and remain resilient in fear, disappointment, and discouragement. My journey is a testament to the power of resilience, and I want to inspire you to face your challenges with the same determination.

Through my experiences in academia and industry, I have developed a deep commitment to rejecting assumptions and exploring new possibilities. This perspective includes questioning established hypotheses or applying machine learning and statistics in novel ways to benefit our communities. We can create a better world by testing different approaches and embracing uncertainty. Join me in this session to explore the exciting world of innovation in machine learning.

Moreover, I am passionate about mentoring women and equity-deserving communities to pursue their goals without fear. In this session, I will discuss how we can tackle our capabilities, overcome challenges, and use our knowledge to make a meaningful impact. Together, we can push the boundaries of what's possible in machine/statistical learning and create a more inclusive and innovative future.

IDWSDS 2024

Session Info

S53

TECHNICAL SESSION

October 8th, 20:00 - 21:00 UTC

Bayesian vs. Frequentist Approaches in Group Sequential Clinical Trial Design

CLICK HERE TO ACCESS THE ZOOM ROOM

Bayesian scholars argue that their approach allows for multiple interim trial decisions without adjustments, while frequentist methods require adjusting stopping boundaries to control Type 1 error, which increases with more interim analyses. Also Bayesian guidelines are thought to result in faster decisions on efficacy or futility compared to traditional frequentist approaches, like spending functions and conditional power.

This invited session offers a critical and thorough examination of Bayesian and frequentist methods in clinical trial design, with a focus on their performance and practical implications.

ORGANIZER: Ruba Shalhoub, NHLIB/NIH

CHAIR: Dong-Yun Kim, NHLIB/NIH

SPONSOR: Caucus for Women in Statistics and Data Science

Speaker Bios

PROF. NANCY FLOURNOY

University of Missouri

https://stat.missouri.edu/people/flournoy

Nancy Flournoy is Curators Distinguished Professor Emerita at the University of Missouri System and former department chair at the University of Missouri and American University. She was founding Director of Clinical Statistics at the Fred Hutchinson Cancer Research Center. Her unit merged into the newly created Division of Public Health when she left to become the first female Statistics Program Director at the NSF. Nancy is Fellow of the ASA, IMS, AAAS, WAAS, and Elected Member of ISI. Her honors include the Founders Award from ASA, Elizabeth Scott and F.N. David Awards from the Committee of Presidents of Statistical Societies, a Distinguished Service Award from the NISS, an Outstanding Performance Award from the NSF and the highest distinguished research awards from the University of Missouri and American University. Recent work focuses on the effects of informative design adaptations on inference including induced bias and distributional alternations.

DR. JUNGNAM JOO

NHLBI/NIH

Dr. Jungnam Joo is a mathematical statistician (biomedical) at the Office of Biostatistics Research of the National Heart, Lung and Blood Institute(NHLBI), National Institutes of Health. Prior to joining NHLBI, she was a chief investigator and head of the Division of Population Science and Epidemiology at the National Cancer Center, Korea.

Dr. Joo received her B.S. in statistics from Seoul National University, Korea and her M.S. and Ph.D. in applied mathematics and statistics from State University of New York.

Her current research interests include clinical trial methodology, machine learning methodology for risk prediction and statistical genetics.

DR. ERIC LEIFER

NHLBI/NIH

Dr. Eric Leifer is a Mathematical Statistician at the National Heart, Lung, and Blood Institute (NHLBI), where he has worked since 2000. He provides statistical leadership for numerous clinical trials, focusing on heart failure, blood and marrow transplant, and COVID-19.
He holds a Ph.D. in Mathematical Statistics from the University of Maryland, College Park, where he also earned two M.A. degrees in Mathematical Statistics and Mathematics. Dr. Leifer has contributed to several professional committees, including the Blood and Marrow Transplant Clinical Trial Network and the Heart Failure Collaboratory.
His interests include monitoring clinical trials, hierarchical endpoint analysis, meta-analysis, heart failure, and blood and marrow transplantation.

Bayesian Alternatives for Group Sequential Designs

Nancy Flournoy, University of Missouri

Prominent Bayesian scholars (e.g., Berry and Ho, 1988) argue that Bayesian philosophy permits multiple interim decisions to stop or continue a trial without adjustment or penalty. In contrast, frequentist practice is to adjust stopping boundaries to control Type 1 error, recognizing that without adjustment it converges to one as the number of interim tests increases.

It is standard Bayesian practice, to update the posterior distribution with new information that becomes available. However, for interim decisions in group sequential designs, it is standard NOT to condition the sampling density on decisions that have been made. The consequence is that the likelihood is invariant to the decision and information in the decision is lost.

Because it is also standard Bayesian philosophy that an analysis be performed on the experiment that was actually run, and not on experiments that might have been run, we investigate incorporating information about the interim decision into the post-interim-decision posterior. We explore the consequences of conditioning the sampling density on the interim decision for subsequent posterior analyses in the context of a two-stage design with an early stopping option.

Comparison of Bayesian and Frequentist Monitoring Boundaries Motivated by the Multiplatform Randomized Clinical Trial

Jungnam Joo, NHLBI/NIH

The coronavirus disease (COVID) 2019 pandemic highlighted the need to conduct efficient randomized clinical trials with interim monitoring guidelines for efficacy and futility. Several randomized COVID 2019 trials, including the Multiplatform Randomized Clinical Trial (mpRCT), used Bayesian guidelines with the belief that they would lead to quicker efficacy or futility decisions than traditional “frequentist” guidelines, such as spending functions and conditional power.

We explore this belief using an intuitive interpretation of Bayesian methods as translating prior opinion about the treatment effect into imaginary prior data. These imaginary observations are then combined with actual observations from the trial to make conclusions.

Using this approach, we show that the Bayesian efficacy boundary used in mpRCT is actually quite similar to the frequentist Pocock boundary. In a pandemic where quickly weeding out ineffective treatments and identifying effective treatments is paramount, aggressive monitoring may be preferred to conservative approaches, such as the O’Brien-Fleming boundary. This can be accomplished with either Bayesian or frequentist methods.

IDWSDS 2024

Session Info

S54

TECHNICAL SESSION

October 8th, 20:00 - 21:00 UTC

Statistics Helping the Growth of Business

CLICK HERE TO ACCESS THE ZOOM ROOM

Together with new, more efficient and powerful computing capabilities, within the context of Data Science, Statistics is playing a fast-growing role in business and society. This session intends to show how this may be done, through the presentation of some interesting case studies.

ORGANIZER: Suhwon Lee, University of Missouri

CHAIR: Suhwon Lee, University of Missouri

SPONSOR: Caucus for Women in Statistics and Data Science

IDWSDS 2024

Speaker Bios

S54

HUGO PEREIRA

University of Lisbon

Hugo Miguel Pereira is a master's student in Applied Mathematics in Economics and Management at the University of Lisbon. He graduated with a degree in Applied Mathematics from the same institution in 2022, where he developed a strong interest in statistics and economics. Currently, as part of his master's program, Hugo is conducting a thesis on economic analysis of a new medical prognostic device at Ophiomics, the company responsible for its development.

HELENA MOURIÑO

University of Lisbon

Helena Mouriño is an Associate Professor at the Faculty of Sciences, University of Lisbon. She holds a PhD in Statistics and Operational Research, with a specialisation in Probability and Statistics from the University of Lisbon. Currently, she is the Coordinator of the MSc in Biostatistics at the Faculty of Sciences, University of Lisbon. She is also a member of the interdisciplinary thematic network RedeSAÚDE (Health Network) at the University of Lisbon, which promotes collaborative efforts to address national and international challenges in the health sector. Helena's research interests include data analytics, with a focus on regression modelling, time series analysis, and AI primarily applied to health sciences. She is also interested in quantitative causal inference. Helena has authored several articles in these areas and has supervised numerous MSc students in Biostatistics.

RAQUEL FONSECA

University of Lisbon

Raquel João Fonseca is an Assistant Professor at the Faculty of Sciences, University of Lisbon. She completed her PhD in robust optimization of international portfolios at Imperial College in 2011, having prior worked in financial consulting at PricewaterhouseCoopers in Lisbon. She is a member of the coordination team of the MSc in Mathematics Applied to Economics and Management and has supervised several students in the finance, actuarial and banking fields. Her main research interests are robust optimization, operations research applied to finance and economics, and more recently health economics.

RICARDO GALANTE

SAS Portugal and University of Lisbon

Ricardo Galante is a Principal Analytics & AI Advisor at SAS Institute – Iberia, where he spearheads advanced analytics and AI initiatives within the region. He leverages over 20 years of experience in the analytics market, including deep expertise in machine learning, artificial intelligence and advanced analysis, to provide invaluable strategic support to clients across retail, consumer goods, business services, finance, energy, and manufacturing. He is also an international speaker, sharing his insights at several industry events. Additionally, he is an Invited Professor in Data Science at the University of Lisbon, and IPAM (Instituto Português de Administração de Marketing – Marketing Business School), sharing his knowledge in Statistics, Data Science, and AI. He is a PhD student in Statistics and Machine Learning at the University of Lisbon and has a master's degree in Bayesian Inference, along with a degree in Statistics from the Federal University of São Carlos (Brazil).

MARIANA FRANCO

Novobanco

Mariana Franco completed in 2021 the bachelor's degree in Applied Mathematics from the Faculty of Sciences of the University of Lisbon. In 2023, finished the master's degree in Mathematics Applied to Economics and Business at the Faculty of Sciences of the University of Lisbon, presenting a final project with the theme "Default Prediction Models for Medium-Sized Enterprises", which goal was to develop a model to predict the probability of default (PD parameter) when granting credit to a medium size enterprise. Currently at the position of Risk Analyst in the Model Validation Unit at Novobanco, she works on the validation of various risk models, such as IRB models.

TERESA ALPUIM

University of Lisbon

Teresa Alpuim is a Professor of Statistics in the Faculty of Sciences at the University of Lisbon. She graduated in Mathematics, concentrating on Statistics, Operations Research and Computing, and attained a Ph.D. In Probability Theory at University of Lisbon. After beginning her academic career, she changed her focus to research in Data Analysis, Statistical Modeling and Forecasting in Meteorology, Environmental Sciences and Actuarial Science. More recently she has a growing interest in Data Science and how to use it in finding good solutions to manage financial risk. Recently she has promoted a tight collaboration between universities and companies and was involved in several projects with banks and insurance companies. She has held diverse academic administrative positions, served several mandates in the board of the Higher Education National Union and presently is the coordinator of the MS in applied mathematics to Economics and Business, offered by Faculty of Sciences of Lisbon.

Abstracts

Kaplan-Meier estimates and Markov models in health economic analysis: a statistical approach to business decisions

Hugo Pereira, University of Lisbon
Helena Mouriño, University of Lisbon
Raquel Fonseca, University of Lisbon

Liver transplantation (LT) is primary curative option for patients with hepatocellular carcinoma (HCC). Due to the scarcity of cadaveric donor livers, selection criteria have been established, but they are very restrictive. This study compares a new criterion, HepatoPredict (ClassI and ClassII), against existing ones (Milan Criteria (MC), UCSF, Up-to-7, AFP Model, and MetroTicket 2.0) using a cost-effectiveness analysis from the U.S. healthcare system perspective to determine which criteria is better. A Markov model was used to simulate the health status of patients with HCC who underwent LT over five years. Transition probabilities, costs, and utility were obtained from published data. Recurrence probabilities, calculated using Kaplan-Meier estimators, were based on a cohort of 149 patients from Portugal and Spain. We analysed the recurrence-free survival, life years gained, quality of life and the incremental cost-effectiveness ratio (ICER) relative to the MC. HepatoPredict offers the best benefit but has a higher cost. The ICER of HepatoPredict-ClassI and HepatoPredict-ClassII relative to the MC was $16 085.43/QALY and $39 407.58/QALY, respectively, both below the cost-effectiveness threshold (U.S. GDP per capita, $81 632.25/QALY), which means that HepatoPredict is acceptable in the U.S. healthcare system. It is the most cost-effective criterion and optimized organ allocation although deceased donor liver scarcity, with significant advantages for healthcare system.

Integrating Statistics and Machine Learning to Forecast New Products Sales: the Case of a Portuguese Brewery

Ricardo Galante, SAS Portugal and University of Lisbon

The introduction of new products is an important point of growth for any brewery, yet the inherent uncertainty of consumer preferences poses a significant challenge. Traditional forecasting methods may struggle to accurately predict demand in this dynamic market. This research explores the potential for machine learning (ML) systems to enhance demand forecasting for new products within the Portuguese brewery industry. We propose a framework utilizing historical sales data, market trends, and relevant external factors such as seasonality and economic indicators. A suite of ML algorithms, including cluster analysis, regression models, decision trees, and potentially neural networks, will be evaluated for their predictive performance. The study aims to: • Identify the most important predictors of demand for new brewery products. • Compare the accuracy of various ML algorithms in this forecasting context. • Develop a practical ML-based forecasting system tailored to the Portuguese brewery sector. This research provides breweries with data-driven insights into demand for new products, aiding in decision-making, production planning and, ultimately, improving resource allocation and profitability. Keywords: New Product Forecasting, Cluster Analysis, Gradient Boosting, Demand Forecasting, Machine Learning.

Credit Risk as a Tool for the Leverage of Investment in Business: Evaluating the Probability of Default for Medium Size Companies

Mariana Franco, Novobanco
Teresa Alpuim, University of Lisbon

Granting credit is one of the main banking activities and an essential factor for economic growth. However, a poor or careless risk assessment can have serious negative consequences, the most recent case being the so-called subprime crisis of 2008. Since then, the banking authorities have been tightening the regulations for granting credit, especially regarding the solvability ratio which has to be greater than 8%. The contribution of credit risk to this ratio is very important and is calculated on the basis of three parameters: the Probability of Default (PD), the Exposure at Default (EAD) and the Loss Given Default (LGD). We propose an evaluation method for the Probability of Default for medium-size companies, on the basis of which the enterprise will be classified as compliant or non-compliant. As default is a binary variable, we use logistic regression with explanatory variables extracted from a wide range of information about the situation of the enterprises, related to its balance sheet and other qualitative information. After a carefull treatment of the explanatory variables, including the elimination of highly correlated variables and the grouping of different levels of some categorical variables, several logistic regression models were constructed, using different statistical approaches. All the models showed good prediction capabilities but we selected the approach that, while keeping a good predictive capability, produces the most parsimonious model.

IDWSDS 2024

Session Info

S55

CAREER OR PROFESSIONAL DEVELOPMENT SESSION

October 8th, 21:00 - 22:00 UTC

Navigating Scientific Challenges and Personal Milestones: Stories from Woodroofe Awardees, leading statisticians and data scientists

CLICK HERE TO ACCESS THE ZOOM ROOM

This session highlights the career journeys of two distinguished Woodroofe Awardees from the Caucus for Women in Statistics and Data Science (CWS), offering a unique look into the intersection of professional achievements and personal transitions. The speakers will explore the challenges of working across disciplines, the excitement and unpredictability of collaborative research, and how key life events—such as the COVID-19 pandemic and maternity leave—have shaped their paths. Through reflections on statistical theory, such as change-point detection, and real-life experiences, they will share valuable insights on adapting to change, the importance of curiosity and passion in research, and the critical role of mentorship and support systems in their success. This session promises to inspire and inform, showing how statisticians navigate both scientific and life transitions.

ORGANIZER: Dong-Yun Kim, NHLIB/NIH

CHAIR: Dong-Yun Kim, NHLIB/NIH

SPONSOR: Caucus for Women in Statistics and Data Science

Speaker Bios

PROF. YANG CHEN

U of Michigan

https://yangchenfunstatistics.github.io/yangchen.github.io//

Yang Chen is an assistant professor at the Department of Statistics at the University of Michigan and a research assistant professor at the Michigan Institute for Data and AI in Society (MIDAS). She got her Ph. D. in Statistics at Harvard University in 2017 and B. Sc. in Mathematics at the University of Science and Technology of China in 2011.
Her research is focused on statistical methodology and computational algorithms motivated by scientific applications, especially in astronomy, space, and climate sciences.
Yang Chen is an Elected Fellow of the International Statistical Institute (ISI) and the winner of the Caucus for Women in Statistics and Data Science (CWS) Woodroofe Award in 2024. Yang Chen is an associate editor for the Journal of the American Statistical Association (JASA) Review, The American Statistician, Statistical Science, and The New England Journal of Statistics in Data Science.

PROF. YAO XIE

Georgia Institute of Technology

https://www2.isye.gatech.edu/~yxie77/

Yao Xie is the Coca-Cola Foundation Chair and Professor at Georgia Tech in Industrial and Systems Engineering and Associate Director of the Machine Learning Center. She holds a Ph.D. in Electrical Engineering from Stanford University and was previously a Research Scientist at Duke University.
Her research lies at the intersection of statistics, machine learning, and optimization in providing theoretical guarantees and developing computationally efficient and statistically powerful methods for problems motivated by real-world applications.
She has received several prestigious awards, including the NSF CAREER Award, INFORMS Gaver Early Career Award, and the CWS Woodroofe Award. Xie serves as an Associate Editor for several top journals and is actively involved in prominent AI and statistics conferences.

DR. DONG-YUN KIM

NHLBI/NIH

https://www.kimdongyun.com/

Dr. Kim is a mathematical statistician at the Office of Biostatistics Research within National Heart, Lung, and Blood Institute (NHLBI), Bethesda, Maryland and adjunct professor at the department of statistics, George Mason University. She received a PhD in Statistics from the University of Michigan, Ann Arbor in 2003.
Her research interests include fully sequential monitoring in clinical trials, change-point inference, and statistical genetics. Currently she is involved in large NHLBI-sponsored clinical trials and intramural projects in MRI imaging, pulmonary disease and cancer research.
Dr. Kim served as the President of the Caucus for Women in Statistics and Data Science (CWS) in 2023. Currently, she is serving as a board member for the Korean International Statistics Society (KISS) and a co-Chair of International Day of Women in Statistics and Data Science (IDWSDS).

A Statistician's Fun and Challenges Playing in Physicists' Backyard

Yang Chen, U of Michigan

Driven largely by scientific applications, my statistical work has led me on a collaborative journey with researchers in the physical sciences. This journey began with challenges, experienced interruptions due to COVID-19 and maternity leaves, but eventually gained rapid momentum.

Throughout this process, several key elements have sustained me: a passion for my work, intellectual curiosity, effective mentoring, teamwork dynamics, and unwavering family support.

A Life Pursuit of Change-Points

Yao Xie, Georgia Institute of Technology

"Change-points" can be abrupt jumps in sequential data, which can be observable, or indirectly observable. Life is full of change-points, and mine not an exception. In work, I study change-point detection: the statistical theory and method for quickly detecting changes from noisy data, which is a foundational topic for statistics.

In life and career, I try to face and adapt to change-points with a positive upspring. I would like to share my life journey thus far along this line.

IDWSDS 2024

Session Info

S56

TECHNICAL SESSION

October 8th, 21:00 - 22:00 UTC

Innovative Health Data Science Applications

CLICK HERE TO ACCESS THE ZOOM ROOM

In this session, students will present their innovative applications of statistics and data science to address impactful health topics. First, Sonya Eason will present an analysis of the "Survey of Prison Inmates, United States, 2016" dataset funded from research by Bureau of Justice Statistics to better understand the factors that are involved in prisoner access to healthcare providers. Yule Fu will discuss a comparative analysis of various natural language processing models for detecting emotions in student responses related to mental health during the COVID-19 pandemic. Caitrin Murphy will present an extension of functional principal component analysis to better classify eating disorders for type 1 diabetic patients. Carol Wang's project focuses on building parametric and nonparametric generative models to synthesize microbiome compositional data. Finally, Luopeiwen Yi will discuss the impact of image preprocessing on enhancing Diabetic Retinopathy detection models across various camera sources.

ORGANIZER: Andrea Lane, Duke University

CHAIR: Andrea Lane, Duke University

Speaker Bios

SONYA EASON
Duke University

Sonya Eason is a third-year undergraduate at Duke University, majoring in statistics. She is passionate about utilizing statistics in biology, healthcare, and social justice causes.

YULE FU
Duke University

Yule Fu is a sophomore undergraduate at Duke University, majoring in math and computer science with minors in bioinformatics and philosophy. Her passion for interdisciplinary exploration has led to engagement in multiple research experiences, centered around the intersection of mathematics, machine learning, and their impactful applications across various fields.

CAITRIN MURPHY
Duke University

Caitrin Murphy is a 3rd year PhD student in the Statistical Science department at Duke University. Her research interests are in functional data analysis, with a particular focus on applications in healthcare.

CAROL WANG
Duke University

As a Data Scientist at Meta, Carol Wang applies her expertise in statistical modeling and machine learning to optimize compute resource efficiency. She is also a PhD new graduate in Statistics at Duke, where her research focuses on modeling microbiome compositional data and developing tree-based nonparametric generative models for learning conditional distributions.

LUOPEIWEN (TINA) YI
Duke University

Luopeiwen (Tina) Yi is a data scientist with a strong foundation in economics, business, and legal analysis. She is currently pursuing a Master's in Interdisciplinary Data Science (MIDS) at Duke University. With a Bachelor of Arts in Economics and International Relations from New York University, she combines technical expertise with a passion for social impact, dedicated to using data-driven insights to drive meaningful change.

Ethical Modeling of Healthcare Provider Access in Prisons

Sonya Eason, Duke University

Building parsimonious models with large population datasets poses several technical and substantive challenges. To name a few, there's missing data, the need for interpretable and ethically selected variables, and imbalance bias. This analysis aims to provide information on the odds of prison inmates accessing healthcare providers, primarily on the basis of medical history and demographics, while also developing a model to predict whether a healthcare provider was seen since prison admission. Model evaluation was completed utilizing k-fold cross validation and ROC-AUC methods.

A Comparative Analysis and Uncertainty Quantification of Machine Learning Models for Sentiment Analysis on Mental Health Data

Yule Fu, Duke University

This research presents a comparative analysis of various natural language processing (NLP) models for detecting emotions in student responses related to mental health during the COVID-19 pandemic. The models evaluated include a lexicon-based approach, Bag of Words, TF-IDF, MentalBERT, and GPT-3.5. The goal was to determine the effectiveness of these models in classifying emotions into categories such as anxiety, depression, and stress. The models were assessed based on accuracy, precision, recall, and F1 score, with results varying from different models. The study underscores the potential of advanced NLP techniques in analyzing emotional responses in textual data, providing valuable insights into the psychological impacts of the pandemic on students.

Functional Principal Component Analysis for Truncated CGM Data

Caitrin Murphy, Duke University

Functional principal component analysis (FPCA) is a key tool in the study of functional data, however, existing methods do not apply when functional observations are truncated, e.g., the measurement instrument only supports recordings within a pre-specified interval. We extend the FPCA framework to accommodate truncated noisy functional data and demonstrate the use of the resulting FPC score predictor in the generalized functional linear model. Interest in this setting is motivated by blood glucose data measured on a continuous glucose monitor (CGM) that only supports readings in the interval 40 – 400 mg/dL. We illustrate the practical value of the proposed method in a clinical application regarding the classification of eating disorders in type 1 diabetic individuals using truncated CGM trajectories.

Tree-Based Generative Models for Microbiome Compositional Data

Carol Wang, Duke University

Analyzing the human microbiome compositional data provides insights into various aspects of human health. Generating synthetic datasets of microbiome compositions under certain conditions is usually of interest to practitioners. In this work, we present two classes of generative models for microbiome compositional data: (1) a parametric model based on a single phylogenetic tree, and (2) nonparametric models based on tree ensembles, and discuss their pros and cons.

Examining the Impact of Image Preprocessing on Diabetic Retinopathy Detection: A Cross-Dataset Analysis with Varied Camera Sources

Luopeiwen (Tina) Yi, Duke University

Diabetic retinopathy (DR) is a significant global health issue, especially in low- and middle-income countries. Machine learning models offer promise in DR diagnosis, but their performance can vary due to disparities in image acquisition. This research examines the impact of image preprocessing techniques, such as cropping and color normalization, on model generalization across different datasets. Using EfficientNetB0, we found that cropping improved performance by focusing on relevant features, while combining cropping with color normalization reduced dataset disparities. For the Messidor dataset, preprocessing increased accuracy from 58.3% to 70.8%. These findings highlight the importance of preprocessing in enhancing model accuracy and generalization for DR detection.

IDWSDS 2024

Session Info

S57

TECHNICAL SESSION

October 8th, 21:00 - 22:00 UTC

Methods for Diverse Types of Outcomes, Mediators, and Confounders in Causal Mediation Analysis

CLICK HERE TO ACCESS THE ZOOM ROOM

A causal mediation analysis allows researchers to understand mechanistic pathways underlying an exposure-outcome relationship by decomposing this relationship into a direct effect and an indirect effect through one or more mediators. While traditional methods for mediation analysis have estimated these effects using a set of linear regression models by taking a product or difference in regression coefficients, causal methods based on the counterfactual outcomes framework have provided researchers with increased flexibility in the types of variables that can be considered. This session will highlight new statistical methods in causal mediation analysis for diverse types of outcomes, mediators, and confounders. The first talk will focus on conducting mediation analyses when either the outcome or mediator is zero-inflated by integrating marginalized zero-inflated models into the analysis. The second talk will introduce a new method for conducting mediation analyses with a time-to-event outcome and a time-varying latent mediator. The third talk will illustrate a comparison of causal methods for mediation analysis when multiple correlated mediators of interest, providing practical recommendations for selecting an appropriate method. Finally, the fourth talk will discuss methods for confounder control and sensitivity analysis, especially when there are either too many confounders to include or when key confounders may be missing or unavailable.

ORGANIZER: Melissa Smith, University of Alabama at Birmingham

CHAIR: Samantha Seals, University of West Florida

Speaker Bios

DR. LEANN LONG

Wake Forest University School of Medicine

https://school.wakehealth.edu/faculty/l/d-leann-long

Dr. Leann Long is an Associate Professor of Biostatistics and Data Science at Wake Forest University School of Medicine. Her methodological research focuses around zero-inflated count data. She leads statistical activities for several clinical trials and has expertise in coordinating centers for large cohort studies.

DR. XIAOXIAO ZHOU

University of Alabama at Birmingham

https://scholars.uab.edu/19406-xiaoxiao-zhou

Dr. Xiaoxiao Zhou is an Assistant Professor in the Biostatistics Department at the University of Alabama at Birmingham. Before joining UAB, she completed her postdoctoral fellowship in the Statistics Department at Duke University. Dr. Zhou's research interests include causal inference, Bayesian methods, longitudinal data analysis, survival analysis, and latent variable models. Her work has been published in diverse journals, including Statistics in Medicine and Structural Equation Modeling: A Multidisciplinary Journal.

DR. MELISSA SMITH

University of Alabama at Birmingham

https://scholars.uab.edu/17587-melissa-smith

Dr. Melissa Smith is an Assistant Professor in the Department of Biostatistics at the University of Alabama at Birmingham. She completed her BA in Mathematics at Colorado College and her PhD in Biostatistics at the University of Iowa, where she was a National Science Foundation Graduate Research Fellow. Melissa's methodology research is motivated by the statistical challenges she encounters in her collaborative work, particularly in observational studies. She is currently developing statistical methods in the areas of causal mediation analysis, causal decomposition analysis, and environmental mixture modeling.

DR. MILICA MIOČEVIĆ

McGill University

https://www.milicamiocevic.com/

Dr. Milica Miočević is an Associate Professor of Quantitative Psychology at McGill University. Her research focuses on statistical mediation analysis, methods for data synthesis, Bayesian statistics, and mediation analysis in single case designs with broad applications in the social, health, and behavioral sciences. She was awarded the Marie Curie Individual Fellowship and she is currently a Dawson Scholar. In her free time she enjoys traveling and reading.

Using Marginalized Zero-Inflated Models in Causal Mediation Analysis

Leann Long, Wake Forest University School of Medicine

Zero-inflated count data are common in healthcare research, but these data can be challenging to incorporate into mediation analyses. Traditional zero-inflated regression techniques provide latent class interpretations, which are often not directly mappable to specific research questions. Marginalized zero-inflated models jointly account for the excess zeroes and directly model the overall mean count, which makes their application to mediation analyses valuable. This talk will discuss causal mediation methods for data with either zero-inflated mediators or outcomes and discuss currently available software implementation. These methods are then illustrated through real applications. The zero-inflated number of alcohol beverages per week is evaluated as a potential mediator to explain sex differences in cholesterol levels. Diabetes status is examined as possible mediator in the relationship between body mass index and the zero-inflated number of inpatient visits.

Causal Mediation Analysis for Multivariate Longitudinal Data and Survival Outcomes

Xiaoxiao Zhou, University of Alabama at Birmingham

This study proposes a joint modeling approach to conduct causal mediation analysis that accommodates multivariate longitudinal data, dynamic latent mediator, and survival outcome. First, we introduce a confirmatory factor analysis model to characterize a time-varying latent mediator through multivariate longitudinal observable variables. Then, we establish a growth curve model to describe the linear trajectory of the dynamic latent mediator and simultaneously explore the relationship between the exposure and the mediating process. Finally, we link the mediating process to the survival outcome through a proportional hazards model. In addition, we use the mediation formula approach to assess the natural direct and indirect effects and prove the identifiability of the causal effects under sequential ignorability assumptions. A Bayesian approach incorporating the Markov chain Monte Carlo algorithm is developed to estimate the causal effects efficiently. Simulation studies are conducted to evaluate the empirical performance of the proposed method. An application to the Alzheimer's Disease Neuroimaging Initiative study further confirms the utility of the proposed method.

Comparison of Methods for Mediation Analysis With Multiple Correlated Mediator Variables

Melissa Smith, University of Alabama at Birmingham

Various methods have emerged for conducting mediation analyses with multiple correlated mediators, each with distinct strengths and limitations. However, a comparative evaluation of these methods is lacking, providing the motivation for this project. In this talk, we will introduce six mediation analysis methods for multiple correlated mediators that provide insights to the contributors of health disparities. We assess the performance of each method in identifying joint and/or path-specific mediation effects in the context of binary outcome variables varying mediator types and levels of residual correlation between mediators. We will share the results of a comprehensive simulation study and an application of the methods to the REasons for Geographic And Racial Differences in Stroke (REGARDS) study. This talk will focus on providing valuable guidance for researchers grappling with complex multi-mediator scenarios, enabling them to select an optimal mediation method for their research question and dataset.

Dealing With Confounders in Statistical Mediation Analysis

Milica Miočević, McGill University

Statistical mediation analysis is used to estimate the indirect effect of an independent variable on an outcome through a mediator. Prior research has shown that including confounders and pure predictors of the outcome leads to unbiased estimates of both indirect and direct effects. Conversely, failing to account for confounders can result in biased estimates of the indirect effect. Ideally, all relevant confounders should be measured and incorporated into the statistical model. However, this is not always feasible, either because the confounders are not included in the dataset or there are too many potential confounders.

This talk will be divided into two parts. The first part will discuss sensitivity analyses designed to quantify the impact of unmeasured confounders on estimates of the indirect effect. The second part will explore optimal methods for modeling large numbers of measured potential confounders in mediation analysis.

IDWSDS 2024

Session Info

S58

HISTORY OF WOMEN STATISTICIANS SESSION

October 8th, 22:00 - 23:00 UTC

Legacies of Women in Data Science and Statistics

CLICK HERE TO ACCESS THE ZOOM ROOM

This session will explore the lasting impacts and ongoing legacies of women in leadership in data science and statistics. Each speaker will discuss legacies of women and their impacts on the speaker's careers in statistics and data science. Among others, Kimiko Bowman, Annie Randall, Gertrude Cox, Florence Nightingale, Grace Hopper, and Mollie Orshansky made substantial, meaningful, and impactful contributions to our field and beyond. We will describe their work and how we, and other women in statistics, are impacted by their work today.

We will celebrate the diverse contributions of women to different subfields of statistics and data science, including biostatistics, statistics education, demography, economics and machine learning. Prominent women in statistics and data science have long-term and far-reaching impact on our field.

ORGANIZER: Emily Griffith, NC State University

CHAIR:

Speaker Bios

DR. JULIA SHARP

National Institute of Standards and Technology

https://sites.google.com/site/julialsharp/home

Julia L. Sharp is a Mathematical Statistician at the National Institute of Standards and Technology. She was previously a Professor and Director of the Graybill Statistics & Data Science Laboratory in the Department of Statistics at Colorado State University. Julia is a widely recognized expert in applied statistics and statistical collaboration and was selected as a Fellow of the American Statistical Association (ASA) in 2022 and received the Outstanding Mentor Award from ASA's Section on Statistical Consulting in 2021. She earned her PhD in Statistics from Montana State University.

DR. EMILY HECTOR

NC State University

https://www.emilyhector.com

Emily C. Hector is an assistant professor in the Department of Statistics at NC State. Emily's work focuses on distributed estimation and inference with big data. She received a National Science Foundation CAREER award from the Division of Mathematical Sciences to develop new statistical methods for data integration. She earned her PhD in Biostatistics from the University of Michigan.

DR. EMILY GRIFFITH

NC State University

https://sites.google.com/ncsu.edu/emilyhgriffith/

Emily H. Griffith is a professor of the practice in the Department of Statistics and the Director of Data Science Consulting in the NC State Data Science Academy (DSA). Emily is a recognized expert in statistical consulting and collaboration, received the Outstanding Mentor Award from ASA's Section on Statistical Consulting in 2022 and was elected a Fellow of the American Statistical Association (ASA) in 2023. She earned her PhD in Statistics from NC State University.

IDWSDS 2024

Session Info

S59

TECHNICAL SESSION

October 8th, 22:00 - 23:00 UTC

HER Impact on EHR: Women Innovators in Electronic Health Records Research

CLICK HERE TO ACCESS THE ZOOM ROOM

As the volume of data from electronic health records (EHRs) continues to rise, so does the accessibility to clinically meaningful variables. This growth has led to increased use of EHR data for analysis, research, and policy-making, attracting biomedical researchers with its low cost of data collection. The application of EHR data has expanded across various clinical domains, including HIV/AIDS, genetics, emergency medicine, and therapeutic effectiveness. However, challenges such as data quality, selection bias, and data sharing need to be addressed for accurate and effective analyses. This session will cover approaches to these challenges: developing risk prediction models using biased EHR data with external predictors and a semiparametric method; evaluating generative large language models (LLMs) in healthcare with a framework that integrates qualitative and quantitative methods; operationalizing a whole-person health score in EHRs despite data quality issues using targeted validation and a measurement error framework; and advancing EHR-based research through contextual thinking, with examples from oncology to emphasize the importance of aligning statistical analyses with medical practice patterns and scientific questions.

ORGANIZER: Ashley Mullan, Vanderbilt University

CHAIR: Ashley Mullan, Vanderbilt University

Speaker Bios

DR. SARAH LOTSPEICH

Wake Forest University

https://www.sarahlotspeich.com

Sarah Lotspeich is an Assistant Professor in Statistical Sciences at Wake Forest University, with a secondary appointment in Biostatistics and Data Science. She co-leads the Spatial and Environmental Statistics in Health (SESH) Lab and helps organize Florence Nightingale Day, engaging local students in statistics and data science. Sarah completed a Postdoctoral Fellowship in Biostatistics at UNC Chapel Hill and earned her Ph.D. in Biostatistics from Vanderbilt University. Her research tackles challenges in analyzing error-prone observational data, focusing on international HIV cohorts, electronic health records, and health disparities. She also develops methods for statistical modeling with censored covariates, applicable to Huntington's disease. Sarah has published in journals such as Biometrics and Statistics in Medicine and received the 2023 David P. Byar Early Career Award. When she's not writing code, you can find Sarah cross-stitching, crocheting, or rewatching The Mindy Project.

DR. LE WANG

Loyola Marymount University

https://cse.lmu.edu/department/math/faculty/?expert=le.wang1986

Le Wang is an assistant professor in the Department of Mathematics, Statistics and Data Science at Loyola Marymount University. She obtained her Ph.D. in Biostatistics from the University of Pennsylvania. Her research focuses on two-phase sampling designs and analytical methods to improve statistical efficiency in biomedical studies. Additionally, she works on development and evaluation of risk prediction models using electronic health record data. Le Wang also holds a Master's degree in Biology, which fuels her interest in interdisciplinary research, particularly in the field of global change ecology. Her diverse academic background enables her to bridge the gap between statistical methodologies and biological applications, contributing to a deeper understanding of the effects of global environmental changes on ecosystems.

DR. CHUAN HONG

Duke University

https://biostat.duke.edu/profile/chuan-hong

Chuan Hong is an Assistant Professor of Biostatistics and Bioinformatics at Duke. Before joining Duke in 2021, she was a postdoc fellow and an instructor of biomedical informatics in Harvard Medical School. Dr. Hong's area of excellence is data science with a unique combination of expertise in both biostatistics and biomedical informatics. Her research is centered on the development of novel statistical and machine learning methods for predictive analytics and precision medicine using large-scale biomedical data with four primary focuses: (1) learning with complex and imperfect outcomes from EHR; (2) promoting fairness in federated and transfer learning methods; (3) automating data harmonization across diverse data sources; and (4) evaluating AI models in healthcare settings.

DR. REBECCA HUBBARD

Brown University

https://vivo.brown.edu/display/rhubbar1

Dr. Hubbard is the Carl Kawaja and Wendy Holcombe Professor of Public Health and Professor of Biostatistics and Data Science at the Brown University School of Public Health. Her research focuses on development and application of statistical methodology for studies using data from electronic health records (EHR) and medical claims. This work encompasses evaluation of screening and diagnostic tests, methods for comparative-effectiveness studies, clinical risk prediction, and health services research. Dr. Hubbard's methodological research emphasizes development and deployment of statistical tools to support valid inference for EHR-based analyses, accounting for complex data availability and data quality issues. She is an elected Fellow of the American Statistical Association, Co-Editor of the journal Biostatistics, a statistical editor for the New England Journal of Medicine, and has published over 200 peer-reviewed papers in the statistical and medical literature.

Operationalizing a Whole-Hospital, Whole-Person Health Score in the EHR Despite Data Quality Issues

Sarah Lotspeich, Wake Forest University

The allostatic load index (ALI) is an informative summary of whole-person health, drawing upon biomarkers to measure lifetime strain. Borrowing data from electronic health records (EHR) is a natural way to estimate whole-person health and identify at-risk patients on a large scale. However, these routinely collected data contain missingness and errors, and ignoring these data quality issues can lead to biased statistical results and incorrect clinical decisions. Validation of EHR data (e.g., through chart reviews) can provide better-quality data, but realistically, only a subset of patients' data can be validated. Thus, we consider strategic ways to harness the error-prone ALI from the EHR to target the most informative patient records for validation. Specifically, the validation study is designed to achieve the best statistical precision to quantify the association between ALI and healthcare utilization in a logistic regression model. Further, we propose a semiparametric maximum likelihood estimator for this model, which robustly corrects data quality issues in unvalidated records while preserving the power of the full cohort. Through simulations and an application to the EHR of an extensive academic learning health system, targeted partial validation and the semiparametric estimator are shown to be effective and efficient ways to correct data quality issues in EHR data before using them in research.

Developing Well-Calibrated Risk Prediction Models Using Biased EHR Data

Le Wang, Loyola Marymount University

To enhance risk assessment using Electronic Health Record (EHR) data, it is desirable to enrich predictors from external sources. To assess the added value of the external risk predictors, we compare the performance of models using only EHR predictors to those incorporating both EHR and external predictors. However, biased evaluation may occur if the study sample does not represent the target population. To address this discrepancy, a semiparametric method was developed assuming the availability of a base model that generates unbiased risk estimates in the target population using only EHR predictors. It enforced calibration of the enriched model by using risk estimates from the base model as constraints in the maximization of the likelihood function for model fitting. The predictive accuracy of the resultant model depends on the extent to which the sample data deviate from the target population. To this end, we evaluate the effectiveness of a propensity score matching approach to improving predictive accuracy via simulation studies and application to Penn EHR data. We then develop inference procedures for unbiased evaluation of the improvement in predictive accuracy.

Evaluating Generative Large Language Models in Healthcare

Chuan Hong, Duke University

The rapid evolution of large language models (LLMs) has ushered in a new era of computational linguistics, yet a systematic approach to their evaluation, particularly in sensitive domains such as healthcare, remains nascent. This work bridges these gaps by offering a detailed and integrated review of qualitative evaluation, quantitative evaluation, and meta-evaluation with examples. We propose an integrated crosswalk between qualitative and quantitative assessment methods. For quantitative evaluation, our review enhances a taxonomy of evaluation metrics, categorizing them based on essential dimensions such as unit of analysis (granularity), human reference, contextuality, supervision, content veracity and trustworthiness. In addition to generic settings, our work distinctively emphasizes additional considerations vital in the healthcare sector. The proposed framework harmonizes qualitative insights, such as user-focused evaluations, with objective quantitative metrics. We present a detailed “go-to menu” of evaluation criteria, tailored to address specific healthcare applications, and emphasize distinct aspects in both pre-deployment and post-deployment phases. Our findings underscore the essential need for a comprehensive evaluation framework for LLMs in healthcare, illustrating the limitations of existing methodologies in meeting the sector's distinct demands.

Context is Queen: Advancing EHR-Based Research Through Our Lived Experience With Medicine, Science and Data

Rebecca Hubbard, Brown University

The modern data science era has upended the traditional statistical notion that data are expensive. In an era of ready availability of vast quantities of data generated as a by-product of digital interactions, obtaining data is easy but deriving valid conclusions from them is hard. Data derived from electronic health records (EHR) have many strengths including large sample size, timely availability, and representation of clinical care and outcomes as they occur in routine medical practice. However, the complex processes giving rise to these data, which hinge on patient interactions with the healthcare system, medical provider practice patterns, and health system coding conventions, must be accounted for in analysis to avoid biased inference. In this setting, grounding statistical analyses in the context of the relevant medical practice patterns and the scientific question of interest is key to good data science. In this presentation I will discuss the use of statistical methods for selection bias, missing data and informative observation processes to encode contextual knowledge in statistical analyses. Through examples in oncology I will illustrate the paramount role of context in EHR-based analyses and the erroneous conclusions that result from context-agnostic analyses. My overarching objective is to highlight the role of contextual thinking in EHR-based research and the importance of bridging science and statistics in analyses of modern data sources.

IDWSDS 2024

Session Info

S60

TECHNICAL SESSION

October 8th, 22:00 - 23:00 UTC

The Thriving Neurodivergent Statistician and Data Scientist: Success and Failures (Learning Opportunities) Within the Field

CLICK HERE TO ACCESS THE ZOOM ROOM

Neurodivergence includes individuals with various neurological or developmental conditions, such as ADHD, autism spectrum disorder, dyslexia, and dyspraxia, among others. The National Institutes of Health estimates that about 15-20% of the U.S. population is neurodivergent, making it crucial to understand their presence in the workforce, including fields like statistics and data science.

In an upcoming panel discussion, panelists will share their personal experiences with neurodivergence and explore how they leveraged their unique talents in statistics. They will discuss successes and failures, illustrating how setbacks paved the way for future achievements. The conversation will highlight how some panelists developed supportive relationships with coworkers who advocated for them.

This informal format will encourage audience participation, resembling a relaxed coffee talk. Panelists will provide specific examples of real situations and outcomes, emphasizing the importance of clear communication about working styles, strengths, and reasonable accommodations. By openly discussing expectations, neurodivergent individuals can educate their peers and shift focus toward productivity rather than personal quirks. Overall, the panel aims to foster understanding and collaboration in diverse work environments, showcasing the value of neurodivergent perspectives in the field.

ORGANIZER: Frank Rojas, NORC at the University of Chicago

CHAIR: Jessica Kohlschmidt, The Ohio State University

IDWSDS 2024

Speaker Bios

S60

FRANK ROJAS

NORC at the University of Chicago

https://www.linkedin.com/in/frank-alexander-r-08956523

Frank Alexander Rojas is a Statistician II in the Statistics and Data Science department at NORC at the University of Chicago, where he has worked since July 2022. He writes and implements SAS and R programs for data extraction, manipulation, and analysis, focusing on statistical tasks like sampling, weighting, analytical modeling, and causal inference. Prior to NORC, Frank spent six and a half years as a Research and Assessment Analyst at the University of Maryland College Park, where he employed various research designs and analytical tools to gather data on students, faculty, and staff. His passion lies in causal inference, particularly using structural causal models to elucidate subject matter and data-generating processes. Frank holds a Bachelor's degree in Psychology from Florida State University and two Master's degrees—one in Higher Education Administration from Florida International University and another in Measurement, Statistics, and Evaluations from the University of Maryland College Park.

DR. SAMANTHA ROBINSON

UA, UADA, and UAMS

https://www.linkedin.com/in/samantha-robinson-02623567/

Samantha Robinson is an Associate Professor of Statistics and Data Science at the University of Arkansas (UA) where she is also the Director of the Center for Statistical Research and Consulting. She also currently serves as a senior applications system analyst for the University of Arkansas for Medical Sciences. Prior to her current roles at the UA, she served as the Director of Data Science Initiatives as well as an Endowed Professor in Mathematics, Clinical Associate Professor, and Department Vice Chair in the Department of Mathematical Sciences.

DR. CHRISTINE CHAI

Internal Revenue Service

https://sites.google.com/site/christinepeijinnchai/

Christine Chai is a Statistician at the Internal Revenue Service (Seattle office). After completing a PhD in statistical science at Duke University in 2017, she started her career at the U.S. Census Bureau and then the U.S. Department of Housing and Urban Development (HUD) in Washington DC. Then she moved to Seattle to work at Microsoft for five years, and rejoined the federal government in May 2024.

DR. DAVID NICHOLS

Northwestern University

https://www.it.northwestern.edu/departments/it-services-support/research/staff/nichols.html

David leads statistical consultations, training, and project support for researchers at Northwestern University. With decades of experience in statistical software support and development, he has consulted on diverse projects across various fields. Previously, David worked at SPSS Inc. and IBM, focusing on SPSS software support. He has addressed thousands of statistical issues, authored technical notes and articles, and created statistical macros and Python-based commands. As Lead Statistician for SPSS, he contributed to developing new procedures and integrating Python Scikit-learn libraries. As Lead Statistician/Data Scientist for IBM Watson Machine Learning Visualization, he earned two IBM Outstanding Technical Achievement Awards and co-authored a U.S. patent. David has a PhD in Research Methodology and Quantitative Psychology from the University of Chicago, an AM in Philosophy, and a BA in Psychology from Michigan State University. He enjoys family trips to national parks and has done the Chicago Marathon three times.

DR. CATHERINE STARNES

Knowesis, Inc.

catherinep.starnes@gmail.com

Catherine Starnes is a biostatistician on a DoD human performance contract through Knowesis Inc.Before entering work in the tactical space, she was a statistics professor and statistics program director at Belmont University, and an adjunct statistics instructor for the Fire and Safety programs at Eastern Kentucky University. Additionally she has worked as a statistical consultant at multiple colleges within the University of Kentucky in different academic statistical consulting laboratories. She holds a PhD in Epidemiology and Biostatistics from the University of Kentucky.

DR. JANA ASHER

Independent Statistician Consultant

https://researchers.mq.edu.au/en/persons/ayse-bilgin

Dr. Jana Asher, PStat®, has published over 65 book chapters, peer-reviewed articles, and manuscripts related to topics ranging from the Human Development Index to human rights measurement to statistical methods such as record linkage. She is an internationally recognized expert on statistics related to human rights and gender-based violence measurement. She is a Fellow of the American Statistical Association, an Elected Member of the International Statistical Institute, and was honored in 2022 by the Caucus of Women in Statistics with their Societal Impact Award.

IDWSDS 2024

Session Info

S61

CLOSING SESSION

October 9th, 00:00 - 00:30 UTC

Closing

CLICK HERE TO ACCESS THE ZOOM ROOM

Join Jessica Kohlschmidt, CWS Executive Director and Cynthia Bland, CWS President, and members of the 2024 IDWSDS organizing committee as we close out this year's conference. We welcome any final questions or discussions from the audience and will briefly touch on the opportunity to publish in our Journal of Statistical Theory and Practice.
Special volume.

Be on the lookout for next year's date!

CYNTHIA BLAND
CWS President

JESSICA KOHLSCHMIDT
CWS Executive Director

IDWSDS 2024

Sponsors

The 2024 IDWSDS Organizing Committee gratefully acknowledges our sponsors and thanks them for their support of our program! We would not be able to have such an amazing program without all these wonderful sponsors. An additional thanks to all the 2024 Friends of IDWSDS for their help in making this years' conference a success!

Gold Sponsors

Sinfonica

WISK
Women In Statistics in Korea

Silver Sponsors

UKRI MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

Bronze Sponsors

CRC Press
Taylor & Francis Group

Minitab

RTI INTERNATIONAL

Springer

Wake Forest University
School of Medicine
Biostatistics and Data Science

연세대학교 보건대학원
GRADUATE SCHOOL OF PUBLIC HEALTH YONSEI UNIVERSITY

IDWSDS 2024

Thank you for joining us at this year's conference! We truly appreciate everyone's contributions in making this a day filled with inspiration and encouragement. While today is a celebration of the incredible achievements of women in statistics and data science, we hope the energy and momentum carry on throughout the year. We look forward to having you with us again next year—mark your calendars for October 14th!