AMERICAN UNIVERSITY
WASHINGTON, DC
DEPARTMENT OF MATHEMATICS AND STATISTICS

Graduate Students Capstone Presentations

9:30am – 12:00pm, Wednesday, April 19th, 2023

Don Myers 114
And via Zoom (link)

- **Jaehee Lee**, M.S. in Data Science
  Topic: *Topic Detection using Dowker Complex*
  Advisor: Dr. Michael Robinson

- **Aarushi Sahejpal**, M.S. in Data Science
  Topic: *Scrum Methodology in Journalism*
  Advisor: Dr. Michael Robinson

- ***Olivia Freides,*** M.S. in Data Science
  Topic: *Developing a Sheaf Library in R*
  Advisor: Dr. Michael Robinson

- **Rachel Carnes**, M.S. in Quantitative Analysis
  Topic: *Model Risk Analysis Internship Experience at Charles Schwab*
  Advisor: Dr. Jun Lu

- **Lucille Bonin,** M.S. in Statistics: Theory and Methods
  Topic: *Research Internship at AARP*
  Advisor: Dr. Jun Lu

- **Pape Theodore Seye**, M.S. in Statistics: Theory and Methods
  Topic: *My experience in statistical consulting*
  Advisor: Dr. Aleka Kapatou

- **Karene Matoka**, M.S. in Statistics: Theory and Methods
  Topic: *Exact Tests for Random Mating in Autotetraploids (MS Thesis Defense)*
  Advisor: Dr. David Gerard

AMERICAN UNIVERSITY
WASHINGTON, DC
DEPARTMENT OF MATHEMATICS AND STATISTICS

Students and faculty are cordially invited!

Time/Date: 9:30 am, Wednesday, April 19th, 2023.
Location: Don Myers 114
And live via Zoom:
https://american.zoom.us/j/97954072785?pwd=Z1p5bWo2a1IxYkVPeUF3TUJGN2dCZz09
Or call +1 301 715 8592
Meeting ID: 979 5407 2785 Passcode: 733650
Contact: Dr. Jun Lu, lu@american.edu

Abstract

- ***Topic Detection using Dowker Complex***. *Jaehee Lee*

  This project aims to classify documents into specific topic categories based on relevant/common terms. Term Frequency-Inverse Document Frequency (TF-IDF), is being widely used to determine how relevant terms are to a given document. However, in some cases TF-IDF is not effective, so we used the Dowker Complex to explore how relevant the terms are between the documents. The Dowker Complex is an abstract simplicial complex based on word usage among documents. In this research, its application is used in a matrix format to represent the relationship between terms and documents using statistical software in R. We are still in the research process, but the current results show that the Dowker Complex separates the documents by their topics, as measured by the topic's probability, more efficiently than TF-IDF. This research can be applied to help librarians and researchers.

- ***Scrum Methodology in Journalism***. *Aarushi Sahejpal*

  Scrum methodology, a common agile framework, is centered around sprints, customer feedback, and lifecycle design. While commonly reserved for product management and software development, its use-cases can be applied to more interdisciplinary fields such as journalism and media to promote collaboration between engineers and writers.

- ***Developing a Sheaf Library in R***. *Olivia Freides*

Sheaf theory is a mathematical framework used to study the properties of complex systems through the local structures of global spaces. It has applications in various fields, including signal processing, data modeling and analysis, and network analysis. Broadly, this project aims to present sheaves as optimal data structures and broaden their accessibility to the R community. The talk will introduce *sheafr*, a package for sheaf manipulation in R, and look at its API, production process and progression, and functionality. After this term, we look to publish *sheafr* on CRAN to contribute to growing the accessibility of Sheaf theory in data science.

- ***Model Risk Analysis Internship Experience at Charles Schwab.*** *Rachel Carnes*

  During the Spring Semester of 2023, I have been given the opportunity to work as a full-time intern at Charles Schwab. I work with the Technical Validation group of the Model Risk Oversight team. The overall goal of my group is to identify any risk associated with the technical aspects of a model, such as model implementation design, data management, model change management, access management, and model execution. I have the unique opportunity to get an in-depth look and understanding for models in production with the firm. Through the internship, I have been able to apply my background in statistics, coding languages, as well as my basic understanding of databases. This internship also provides me with many learning opportunities in areas such as technology and software along with mathematical methodology and research.

- ***Research Internship at AARP***. *Lucille Bonin*

  From last June through the middle of April, I had an internship at the American Association of Retired Persons (AARP) for approximately 20 hours per week. I was a research intern for AARP Library, a team housed under the AARP Research Department (Policy, Research and International). The main purpose of the AARP Library is to field internal research requests from other employees across the organization, including literature reviews, personal vetting and secondary research analysis; along with providing a physical library space for employees. My work primarily included extracting insights from large consumer survey datasets from our vendors and organizing those insights into a report for the client. I also created data visualizations to be displayed on the organization's internal library website.

- ***My experience in statistical consulting.*** *Pape Theodore Seye*

  During the Spring 2023 semester I had a great experience with the Statistical Consulting course at American University. The concept of the course is to assist departments' students

and researchers with the statistical challenges they encounter while conducting their research. I (and the teams I was working with) had a wide range of subjects going from clinical psychology research on suicide treatments to research on students' resilience evolution throughout their stay at American University.

A wide range of statistical tools were used throughout the semester. Some projects only required 'basic' tools such as: ANOVA, simple linear regression, logistic regression, providing summary statistic and visualization etc. Other projects required more advanced tools, and a lot of additional research and readings to familiarize ourselves with the new tools that we would present and explain to the client later on. The new tools included GPower, for power and sample size calculations, unusual models and SPSS. We also learned concepts of experimental design. Finally, we obtained IRB training on ethical considerations when the research involves human subjects (IRB stands for Institutional Review Board). We also received an IRB Certificate to include in our resumes.

My presentation will go over the statistical concepts that I discovered thanks to this experience and how it has informed me about the many challenges that come with working with real life data and ongoing-research projects.

- ***Exact Tests for Random Mating in Autotetraploids.*** *Karene Matoka.*

  In statistical genetics, a common task is to evaluate the assumption of random mating, which is a baseline for population genetic models and can be used as a quality control procedure. In diploids, which are organisms with two copies of their genome, this is done through tests for Hardy-Weinberg equilibrium (HWE). However, for autotetraploids, which are organisms with four copies of their genome, the assumption of random mating and HWE are different and there has been less work on evaluating random mating in these organisms. The standard approach for HWE tests in diploids is to use exact tests, which control for Type I error for finite sample sizes. However, there are currently no exact tests available for autotetraploids. In this study, we provide two approaches for exact tests for random mating in autotetraploids. The first approach conditions inference on sufficient statistics, while the second uses a split likelihood ratio. Although these approaches exactly control for Type I error, simulations show that they are too conservative for use. Therefore, we recommend using a standard likelihood ratio test or a standard chi-squared test, which we implement. We demonstrate all of these approaches using a dataset of autotetraploid white sturgeon.