# URPS 21: Undergraduate Research Program in Statistics for Winter 2021

**Director of Undergraduate Programs: Prof. Edward Ionides**

**Undergraduate Program Coordinator: Gina Cornacchia**

- Is this program for me?
- The application process.
- The projects.
- What to expect if I join a project.
- Do I get course credit? Can I use the project for an honors thesis? For the Data Science major capstone requirement?
- Other opportunities for undergraduate research in statistics.

# 1. Some statistical problems based on an online ranking

We want to study different problems in which a (partial) on-line ranking of different "players" is needed. As an example, we may consider modifications of the Elo's ranking method for soccer teams during a season (e.g., the English Premier League) and investigate whether these improve the prediction of future standings of the teams/players or the ability to detect the change in performance, e.g., after a manager is replaced. Another possible scenario to consider is the selection of k candidates from a pool of n candidates interviewed one after the other when the information about each candidate has a cost proportional to its quality, and the latter can be controlled sequentially by the selector sequentially. The project can have aspects that are mostly mathematical or mostly applied depending on the team.

Supervisors: Prof. Ya'acov Ritov and Debarghya Mukherjee

# 2. Detection of clutch players

This project concerns clutch and anti-clutch (chocking) players in professional sports, which is the phenomenon whereby a player excels during times of higher pressure, such as at the end of a game or during championships (Compare Reggie Miller, Kobe Bryant and Christiano Ronaldo to the Charles Barkley and Lionel Messi of the world). By analyzing modern data sets and utilizing computer simulations, we will investigate the existence of the clutch performance as well as identify potential clutch players.

Supervisors: Prof. Ya'acov Ritov and Michael Law

# 3. Modeling and data analysis to understand spatiotemporal epidemiology of dengue virus

Dengue fever is an emerging infectious disease which is now widespread through Africa, Asia, South America and Central America. Spatiotemporal patterns of dengue incidence are hard to explain using existing epidemiological models. This suggests there are gaps in our scientific understanding. New models will be proposed and computationally intensive statistical inference methods will be used to examine their success. The research project involves participating in a team of scientists and statisticians.

Supervisors: Prof. Edward Ionides and Kidus Asfaw

# 4. Deep reinforcement learning for conformer generation

Chemical bonds are often easily rotatable, allowing molecules to rapidly interconvert between many different geometric configurations known as conformers. The total number of possible conformers grows exponentially with the size of the molecule. However, only a certain subset of low-energy conformers are needed to describe typical behavior of the molecule. The goal of this project is to explore how deep reinforcement learning (RL) can be used to generate a set of conformers that is representative of all likely low-energy conformers. We currently have an RL model called TorsionNet that has performed well on the conformer generation problem for several classes of molecules. One part of this project is to explore how we can improve TorsionNet to generate conformers for larger, more complex molecules. Since most large molecules are made up of smaller, often repeated local groups, we will explore ways to train an agent that utilizes knowledge of the local parts of the molecule to generate conformers for the overall molecule, such as through hierarchical reinforcement learning and curriculum/transfer learning. The second part of this project is to improve the current codebase into a more robust hypothesis-testing framework so that we can more easily implement and test custom agents and environments. We work mainly in Python and PyTorch and require prior experience with the language.

Supervisors: Prof. Ambuj Tewari and Dr. Joshua Kammeraad

# 5. Accurate, usable causal discovery package

Subject-matter experts typically think of their datasets as variables linked by cause/effect relationships, forming a large, complex causal system. Diagrams (directed acyclic graphs, also called Bayesian networks) provide a natural way to conceptualize these systems. In order to be practical for applied researchers, methods for estimating the causal structure underlying a dataset must have usable code and estimates must be accurate. For this project, undergraduate researchers will systematically apply an existing causal discovery package to real-world and simulated data with the goal of assessing the package's usability and accuracy. Undergraduate researchers will be responsible for two deliverables: (1) active collaboration to improve the package's usability/documentation on different computing environments and (2) creating a reproducible python (jupyter notebook) program that generates a report on accuracy for simulated data and results for several publicly available, real-world datasets. A working knowledge of python is necessary.

Supervisor: Dr. Octavio Mesner

# 6. Sample complexity of phylogeny estimators under gene duplication and loss

Phylogenomic methods use a large amount of genomic data to estimate phylogenetic trees. Quartet-based estimators have strong guarantees for the number of data (gene trees) needed to estimate a phylogeny with high probability. However, the presence of duplicated genes sometimes requires entire gene trees to be disregarded. In this project, we investigate further the relationship between gene duplication (and loss) rates and the amount of data sufficient for accurate estimation of a phylogeny.

Supervisor: Dr. Brandon Legried

# 7. Nutrition, life history, and adult health

Our goal is to better understand the relationships between nutrition, life history, and adult health outcomes using data from a longitudinal study conducted in Mali (Africa) that is based at UM. Specifically, this project aims to use Gini regression and analysis of concomitants to understand the extent to which childhood undernutrition and adult overnutrition may conspire to increase the risk for elevated blood pressure in adulthood. While this question has been studied extensively using conventional statistical regression methods, conclusions remain elusive due to the particularly complex interactions among these risk factors. This project will explore the applicability of some seldom-used methods for regression analysis that may complement and reveal weaknesses in the prevailing approaches. The student should have basic Python progamming skills.

Supervisors: Prof. Kerby Shedden and Sanjana Gupta

# 8. Matching and randomized experiments

In pair randomized experiments, participants are organized into pairs, and then one participant in each pair is randomly assigned to treatment. Ideally the participants in each pair will be as similar as possible, and they would therefore vary only on whether or not they were assigned to treatment or control. In observational studies, matching can often be used to emulate this approach. While a researcher does not have control over the treatment assignment mechanism, they can attempt to match participants based on observed characteristics to form pairs of similar participants (one which has received treatment and the other control). In this project, the undergraduate researcher will explore the interaction between matching and randomized experiments (e.g., the use of matching techniques within the analysis of randomized experiments). This project will require a solid foundation in R.

Supervisors: Prof. Johann Gagnon-Bartsch and Ed Wu

# 9. Cancer drug screening

The student researcher will be given large and complex data from cancer drug screening experiments. The data will include information on the effectiveness of hundreds of drugs tested on hundreds of cell lines, in addition to genomic information on the cell lines. The drug screening data contains widespread measurement error, which causes problems during analysis. With the ultimate goal of improving personalized cancer treatment, the student researcher will explore the drug screening data, adapt methods of measurement error detection, and build prediction algorithms to determine which drugs are most effective against which types of cancers. The student researcher may also develop simulations of such drug screening data to improve experimental design methods. The student researcher will learn to work with a variety of real-world, messy data (e.g. gene expression), methods to integrate different types of complex data, and various machine learning algorithms.

Supervisors: Prof. Johann Gagnon-Bartsch and Zoe Rehnberg

# 10. Covariate adjustment in small randomized experiments

Common statistical techniques to estimate causal effects from experimental data often incorporate adjustments using covariates in order to ensure that the treatment groups are balanced in terms of characteristics related to the outcome. There are a variety of design based methods for estimating treatments effects using covariance adjustment. In this project, we are interested in determining how to perform inference with each of these causal estimates (i.e. compute a p-value) with a finite sample size and compare the methods empirically in terms of the type I error rate control. We will also consider the strengths and weaknesses of each method under different data settings. The student researcher will conduct simulations in R to empirically compare these treatment effect estimation methods and determine the best ways to visualize their results. Basic knowledge of R is required, and the student will improve their R skills over the course of the project.

Supervisors: Prof. Johann Gagnon-Bartsch and Charlotte Mann

# 11. Post-selection inference for multi-task lasso

In the classical setting, statistical models are fixed before any data are examined, with only model parameters left to estimate. With complex modern data, it is common to select the model itself based on an exploratory analysis of the data, which makes any subsequent statistical inference unreliable (the reproducibility crisis). Post-selection inference accounts for model selection in subsequent inference. This project develops post-selection inference for multi-task regression, where a set of related regression problems share a common set of covariates. This arises, for example, in predicting treatment outcomes for different tumor types from genetic data, or performance on cognitive assessment tasks from fMRI brain images. We have developed a method that assumes there are three types of covariates and learns them from data: those that act as common signals affecting all the response variables, those that affect only some responses but not all, and those that are irrelevant altogether. The undergraduate team will run existing Python code on simulated and real data and produce numerical and visual summaries to evaluate validity of proposed post-selection inference (e.g. whether confidence intervals achieve the desired coverage probability), as well as compare to other benchmark methods.

Supervisors: Prof. Liza Levina, Prof. Snigdha Panigrahi and Natasha Stewart

# 12. Human brain connectivity

In neuroscience, networks are used to represent brain connectivity patterns (pairwise connections between multiple locations in the brain obtained from neuroimaging data are the network edge weights). Additional data on the same locations in the brain is often available as well (network node features). We are interested in understanding the relationship between these brain measurements and phenotypic measurements such as a psychiatric diagnosis or performance on a cognitive task. We are developing several methods that aim to both predict the phenotype accurately and provide interpretable insights into which parts of the brain play a role. The undergraduate team on this project will work with the PhD student to improve and apply R code for these methods to human brain imaging data from our neuroscience collaborators.

Supervisors: Prof. Liza Levina and Dan Kessler

# 13. Learning to embed ICD codes

As an important component of electronic health records (EHR), ICD codes represent the diagnosis and treatments given by clinicians. Interpretable and efficient embedding of ICD codes can help analyze and process EHR data for further downstream machine learning tasks. In this project, we treat the co-occurrence of ICD codes as a hypergraph and aim to map ICD codes to lower-dimensional numerical vectors while maintaining their main dependence structures. The undergraduate researcher will be responsible for implementing an embedding learning algorithm that we have developed, and compare its performance for downstream prediction tasks with existing approaches.

Supervisors: Prof. Ji Zhu and Weijing Tang

# 14. Classroom size and academic success

In the mid-1980's, the STAR Project in Tennessee randomly assigned students to three different classroom sizes to understand the effect of student-teacher ratios on academic success. This study has been one of the most influential pieces of evidence about classroom size policy, but some analyses have drawn criticism over student attrition. Critics charge that over time, as students drop out of the study, the remaining students no longer represent proper random samples necessary for justifying causal claims. Similar criticisms are often leveled at observational data where researchers would like to move beyond association to make causal inferences. Balance tests give a formal method of answering these criticisms by seeing how different the student groups are on observed variables and quantifying if the difference is large relative to what might be expected in a proper randomized controlled trial. Currently, existing balance testing methods are confined to binary treatment-control comparisons. The undergraduate researchers will help to develop balance tests for multinomial and ordinal treatments. The work will involve theoretical development, software implementation, and applied data analysis of the STAR Project and other data sets.

Supervisors: Dr. Mark Fredrickson and Prof. Ben Hansen

# 15. The Kalamazoo Promise

In 2005, anonymous donors launched the Kalamazoo Promise, a guarantee to pay 100% of tuition for public school students in Kalamazoo who proceed on to college. Our research group seeks qualified students to work on a project utilizing publicly available school-level data on Michigan high schools to assess the impact of that program on Kalamazoo students and their propensity to attain higher education. In particular, participant(s) will match Kalamazoo high schools with similar high schools around the state and then analyze the effect of the Kalamazoo Promise on college enrollment and graduation rates. The current project will culminate in two interrelated research products: a reproducible program in R that matches schools and performs outcome analysis and a report explaining modeling decisions.

Supervisors: Prof. Ben Hansen and Tim Lycurgus

# 16. Analysis and Visualization of Darknet Internet Traffic

The project involves designing and developing an interactive web-application via the R Shiny platform for the analysis of Darknet internet traffic. The Darknet is traffic routed to the space of unassigned (dark) part of the network, not to be confused with the "Darkweb", a collection of sites for illicit activity. Darknet traffic originates from 1) misconfigured or malicious hosts who are scanning for cybersecurity vulnerabilities in the network, 2) from randomly spoofed IP packets aiming to attack specific victims (the so-termed 'backscatter' traffic) and 3) from networking misconfigurations. The dashboard will access a Big Query Google data-base, which contains real-time Darknet traffic measurements. It will compute real-time summary statistics and cybersecurity threat indices, which quantify, classify, and potentially localize cybersecurity threats to the network. The project will train the student in statistical methods such as exploratory data analysis, principal component analysis, extreme value theory, clustering, and classification.

Supervisors: Prof. Stilian Stoev and Dr. Michalis Kallitsis (Merit Network)