Undergraduate Research Opportunities in Statistics for Winter 2019

- [ ] Changes in electrophysiology across a season of high school football
- [ ] Parent preferences in child sport participation
- [ ] Political attitudes and prediction of election results
- [ ] Cancer drug screening
- [ ] Statistical inference on viral genomes
- [ ] Neuroscience of addiction
- [ ] Modeling academic achievement in Michigan K-12 public schools
- [ ] Interactive statistical computing & visualization of functional data
- [ ] Active learning in the streaming setting with purely random trees
- [ ] Optimizing learning in an online platform

Changes in electrophysiology across a season of high school football

This study will evaluate changes in Event Related Potentials (ERPs; ie changes in brain electrophysiology) across a season of high school football participation. Each athlete was equipped with a helmet sensor and evaluated prior to the competitive season and again at mid- and post-season timepoints. Non-contact sport athletes were evaluated at the same time points. This project will evaluate: 1) the overall change in ERPs from pre- to mid- to post-season, 2) the change relative to control athletes, and 3) the change related to head impact exposure as recorded by the helmet sensors

Faculty supervisor: Xuming He

Parent preferences in child sport participation.

This study will evaluate the willingness of parents with and without a medical background to allow their children to participate in contact and non-contact sports. Survey data have been collected from approximately 6000 respondents with assorted demographic information. The primary question will be understanding the willingness of sport participation based on sex, medical training, and personal concussion history.

Faculty supervisor: Xuming He

Political attitudes and prediction of election results

The goals of this project are to understand changing political attitudes and predict election results. We will do this by implementing a new method for estimating a response parameter when the covariates for the entire population are known. We will be using data from previous elections, social media, and various surveys. The undergraduate researcher will help with gathering data from various online sources, data cleaning and organization, and writing code for method implementation.

Faculty supervisor: Johann Gagnon-Bartsch

Cancer Drug Screening

The student researcher will be given a large, complex dataset from a cancer drug screening experiment. The dataset will include information on the effectiveness of hundreds of drugs on hundreds of different cell lines, in addition to genomic information on the cell lines. The goal of the project will be to build prediction algorithms that are able to determine which drugs are most effective against which types of cancers, with an ultimate goal of customizing drug treatments to individual patients. The student researcher will learn to work with several kinds of data (e.g. gene expression), methods to integrate different types of complex data, and various machine learning algorithms.

Faculty supervisor: Johann Gagnon-Bartsch

Statistical inference on viral genomes

Many statistical methods and models have been developed for study of human genetics, but less so for viral genetic sequences. This project involves adapting some of these methods (e.g., principal component analysis, hidden Markov models, expectation-maximization algorithms) to viral genetic data with the overall goal of inferring recombination rate and detecting structure in the data. With the guidance of a graduate student, the undergrad will learn about and implement various modeling and machine learning techniques used in genetics, and carry out interpretation and visualization of the results.

Faculty supervisor: Ed Ionides

Neuroscience of addiction

The data are from a neuroscience pilot experiment in the lab of Prof. Shelly Flagel (Psychiatry), investigating the links between brain chemistry and addiction in rats. The aim of the project is to predict, based on levels of different chemicals in the brain, whether or not a given rat will display different types of behaviors associated with addiction after several days of conditioning experiments. The pilot dataset consists of approximately 60 rats, with about five behavioral measures and ten brain chemistry measurements for each rat; the full scale experiment with involve many more rats and neurochemical measurements, with data becoming available over the next year.

Faculty supervisor: Liza Levina

Modeling academic achievement in Michigan K-12 public schools

The Hansen-Fredrickson Research Group is seeking qualified undergraduate students to work on a project analyzing patterns in academic achievement among K-12 students in the Michigan public schools system. Our undergraduate research assistant(s) will replicate a set of existing models fit to longitudinal student data in order to perform model diagnostics and extend the models. The developed models will play a supporting role in evaluating impacts of certain statewide education initiatives. The research assistant will develop skills using the R programming language for modeling of moderately large datasets, with use of multi-level regression techniques and with graphical diagnostics of regression models. Applicants should have some exposure to programming statistical analyses and familiarity with regression modeling.

Faculty supervisor: Ben Hansen

Interactive Statistical Computing and Visualization of Functional Data

The Argo data set involves a network of about 4000 floats dispersed throughout the Atlantic, Indian and Pacific oceans: `http://www.argo.ucsd.edu/`. These floats operate autonomously and they continuously measure the ocean temperature and salinity as a function depth, latitude and longitude. Thus, these data provide snapshots of the state of the open oceans in unprecedented spatial and temporal detail. The project involves building a series of R Shiny Apps: web-apps coded in R that will help visualize and understand the complex Argo data set. Each Argo float produces temperature profiles indexed by depth ranging from 0 to 2000 meters. These data can be viewed as samples from a curve (a function of depth) and the goal is to estimate the mean of the curves as well as to characterize the covariance structure of the temperature as a function of depth.

Faculty supervisor: Tailen Hsing and Stilian Stoev

Active learning in the streaming setting with purely random trees

Imagine you are in a situation where acquiring unlabelled data is cheap, but labelling them is expensive, for example scraping images, text or audio from the internet (cheap) and requiring humans to label them (expensive). As a result you can only request a small amount of the data to be labelled, and you want to build the best statistical model you can with that finite budget of labels. This is the goal in active learning: to develop algorithms which automatically decide what data should be labelled. We currently have a new algorithm for active learning for regression in the pool setting, where we can see all our unlabelled data and select any point to label at any time. This project entails theoretical and computational work extending existing implementations of our algorithm for the streaming setting, testing multiple variations with simulations and experiments on real data.

Faculty supervisor: Ambuj Tewari

Optimizing learning in an online platform

Intelligent tutoring systems aim to sequentially present questions to learners so as to maximize learning. One method of question selection is to estimate a learners current ability and then present a question that is on the edge of that ability. Intuitively, questions that are too simple will not increase understanding and questions that are too difficult can cause demotivation/frustration. It other words, the goal is to estimate that range of question difficulties that will promote understanding and increase concept mastery. With data from industry collaborators, we aim to develop a sequential decision making algorithm to maximize student mastery in this way. The main goals for the research assistant are to: (1) measure question difficulty by looking at proportion of correct answers and look at the stability of question difficulty over time; (2) identify the optimal range of question difficulty for a given student.

Faculty supervisor: Ambuj Tewari