Department of Statistical Sciences (DoSS)
2021 Summer Undergraduate Research

The Department of Statistical Sciences is seeking applicants interested in conducting a summer research project under faculty supervision. Projects will take place for a 16 week period from May-Aug, 2021. In addition to NSERC USRAs, the Department has also launched the Summer Undergraduate Research Award (SURA) prize that comes with a $500 cheque and may be given to more than one project.

All students who are currently enrolled in a Department of Statistical Sciences (DoSS) Specialist or Major program and who have completed STA302 with at least a B+ by the time of application are eligible to be considered for one of these awards. You will be asked to rank your top 3 projects. Students will be considered based on their academic record, experience and their statement of research interest. Students who are shortlisted for specific projects may be invited for an interview with the prospective supervisor.

Projects

Developing a Shiny App for the R package SWIM

Supervised by: Silvana Pesenti

In this research project we develop a Shiny App for the R package SWIM. SWIM is an R package which provides a novel methodology to perform sensitivity analysis for models used in finance and insurance in a numerically efficient way. Shiny (also an R package) allows to build interactive web applications using R. The goal of this research project is to build a Shiny app called ShinySWIM, that allows a visualisation of the scope and the functionalities of the R package SWIM. Shiny apps are an ideal way to visualise and provide an interactive understanding of features of SWIM.

Applying students should have a good background in statistics and probability; Advanced programming skills in the language R; Shiny app development preferrable.

Multi-scale Modeling of Time Series Data for Classification of Shark Behavior

Supervised by: Vianey Leos Barajas

Multiple sensors attached to sharks allow for collection of fine-scaled movement data over long stretches of time. Different sensors record data at different temporal resolutions, e.g. position data is collected at a coarser temporal scale than accelerometer data, but the main biological research goal remains the same: “what behaviors is the shark exhibiting?”. From a statistical standpoint, the research goal becomes, “what classification algorithms are appropriate to identify shark movement patterns from multi-scale time series data?" For this project, the student will make use of multi-scale time series data, i.e. position and accelerometer data, collected from horn sharks and implement classification algorithms to construct classifiers for horn shark behavior.

Applying students should have knowledge in Bayesian inference and time series as well as the software: Stan.

Generalizing Convex Analysis and Statistical Applications

Supervised by: Leonard Wong

Convex duality plays a fundamental role in probability and statistics. Given a convex function, we can define its conjugate which is another convex function. Particularly nice are *convex functions of Legendre type*, where the Legendre transformation behaves like an isomorphism. This concept was used in the study of exponential family (parametric densities which generalize e.g. the normal distribution) to establish a duality with Bregman divergence (a useful class of loss functions). We are interested in the q-exponential family which generalizes the exponential family and includes many power-law distributions. The q-exponential family leads to a different convex duality: while a convex function is given as the maximum of a family of affine functions, we consider instead a family of "logarithmic functions". The main aim of this project is to extend the notion of Legendre functions to our new setting. This will provide a rigorous justification of the duality between the q-exponential family and a logarithmic divergence, and pave the way for potential statistical applications.

This is a theoretical project. Applying students should have strong background in real analysis (measure theory is useful but not strictly necessary), as well as previous exposure to convex analysis (e.g., conjugate and subdifferential). Our main reference is the book *Convex Analysis* by Rockafellar and we will study part of it in the initial phase of the project.

Evaluating the Effect of Teaching Fundamental Computing Skills on Subsequent Research Careers

Supervised by: Rohan Alexander

There are many assumed skills that are critical in modern sciences but rarely taught such as fundamental research computing skills, including Unix shell, GitHub, and Python/R. We evaluate whether explicitly teaching this assumed knowledge contributes to the likelihood of an undergraduate student continuing on a research path. We identify groups of undergraduates conducting research projects over summer and provide training in fundamental computing skills before they begin. We use surveys immediately before and after the training to evaluate the short-term effectiveness of the training. This work will also establish the basis for a longitudinal study to evaluate the effect of training in such skills on entering research-intensive career paths, such as a PhD program.

Applying ttudents should be comfortable using R within the Tidyverse ecosystem. They must also have taken statistics courses such that they are comfortable creating surveys, designing experiments, and evaluating them. Strong applications would provide evidence of this by linking to relevant GitHub repos. Students should also be comfortable writing papers, and again strong applications would link to examples.

Bayesian Computing for Air Pollution and Mortality

Supervised by: Patrick Brown

Dr. Brown’s team at the Centre for Global Health Research is studying the relationship between daily variations in air pollution in Canadian cities and mortality due to circulatory and respiratory conditions. Bayesian inference for case/crossover models is a core part of this project, which is a computational challenge as the partial likelihood function for these models is a non-linear combination of latent variables and is therefore not compatible with INLA.

This summer research project will involve creating an R package for case/crossover models based on the R code developed by the research group. Specific tasks include

- becoming familiar with case/crossover models and their use in air pollution studies
- writing functions to provide a user-friendly interface to the existing ‘back-end’ functions
- creating worked examples (vignettes) and function help files
- assisting with drafting a manuscript on the methodology for a statistical software journal

Applying students should have knowledge in Bayesian inference, ideally INLA, R programming skills. Some knowledge of survival analysis and environmental epidemiology would also be an asset.

Variable Selection for Longitudinal Data

Supervised by: Tharshanna Nadarajah

High-dimensional longitudinal data with a large number of covariates have become increasingly common in many biomedical applications. The identification of a sub-model that adequately represents the data is necessary for easy interpretation. Also, the inclusion of redundant variables may hinder the accuracy and efficiency of estimation and inference. The joint likelihood function for longitudinal data is challenging, particularly in correlated discrete data. A class of methods via regularization can be developed to target such applications, and they are lauded for computational efficiency and stability.

Applying students should have applied statistics and R programming language knowledge.

A Systematic Evaluation of Reproducibility in Economics

Rohan Alexander

In this research project we will examine the reproducibility of journal articles published in 2020 in the top 5 economics journals – American Economic Review, Econometrica, the Journal of Political Economy, the Quarterly Journal of Economics, and the Review of Economic Studies. We will first download the papers and supporting materials such as code and data. We will create a public-facing Shiny application that provides information about each article as well as summary statistics. We will then attempt to use the provided materials to reproduce and then replicate the articles. We will write-up our systematic findings in a paper that could be itself published in one of these journals. We will also write comment-style papers about papers of particular interest. Economics is only just beginning to understand the extent of the replication-crisis in the discipline and this work will be a significant contribution to this effort.

Applying students should be comfortable using R within the Tidyverse ecosystem. They must also have taken statistics courses such that they are comfortable using and evaluating generalized linear models. Strong applications would provide evidence of this by linking to relevant GitHub repos. Students should also be comfortable writing papers, and again strong applications would link to examples.

How to Apply

Student Eligibility:

1. Must be an undergraduate student currently enrolled in a Department of Statistical Sciences Specialist or Major program.
2. Currently registered full-time or part-time student at the time of application.
3. Must have completed STA302 with at least a B+ grade.

How to submit your application:

Please fill out and submit the following application form:
Application for DoSS Summer Undergraduate Research Awards 2021 (office.com)

If you have any questions regarding these awards, please contact ug.statistics@utstat.utoronto.ca.

Completed applications are due by 11:59PM ET Monday, March 1, 2021