Paid Statistics Research Opportunities for CSU Undergrads!

During Summer 2023, the College of Natural Sciences and the Department of Statistics will be sponsoring several paid research opportunities for undergraduates.

- Students will work with Statistics faculty and graduate students on a variety of research projects (see full list attached).
- **Students will be paid $20 per hour up to $1500.**
- The schedule is flexible, but it is expected that students will meet with faculty advisors at least once a week, attend weekly group meeting and spend up to 10 hours per week working on the project between **05/16/23 – 06/30/23**.
- Participating students should expect to work independently, learn new methods, debug code, find and read relevant literature.
- At the end of the project, students will be required to summarize and present their work.
- Students are strongly encouraged to submit their work at CSU CURC (Celebrate Undergraduate Research and Creativity) in Spring 2024.

Requirements:

- Be a CSU undergraduate student in Fall 2023.
- Be available to work and meet with faculty between 05/16 – 06/30.
- **Apply by Friday 04/21/23**. Your application should include a personal statement about why you are interested in this research opportunity. Please feel free to describe your interest and/or qualifications for specific projects listed. Students will be evaluated on their qualifications, enthusiasm, interest level, and potential to benefit from the research experience.

Preference will be given to students:

- Majoring or minoring in Statistics or Data Science.
- Planning the graduate in Spring 2024 or later.
- Available to attend group meetings in-person on Tuesdays at 1pm.

Interested? Apply here:

https://forms.gle/JDnbjpKjFXYQUPeW6

For full consideration, apply no later than Friday April 21st

Summaries of potential projects and detailed descriptions start on the next page.

Summary of project titles and advisors:

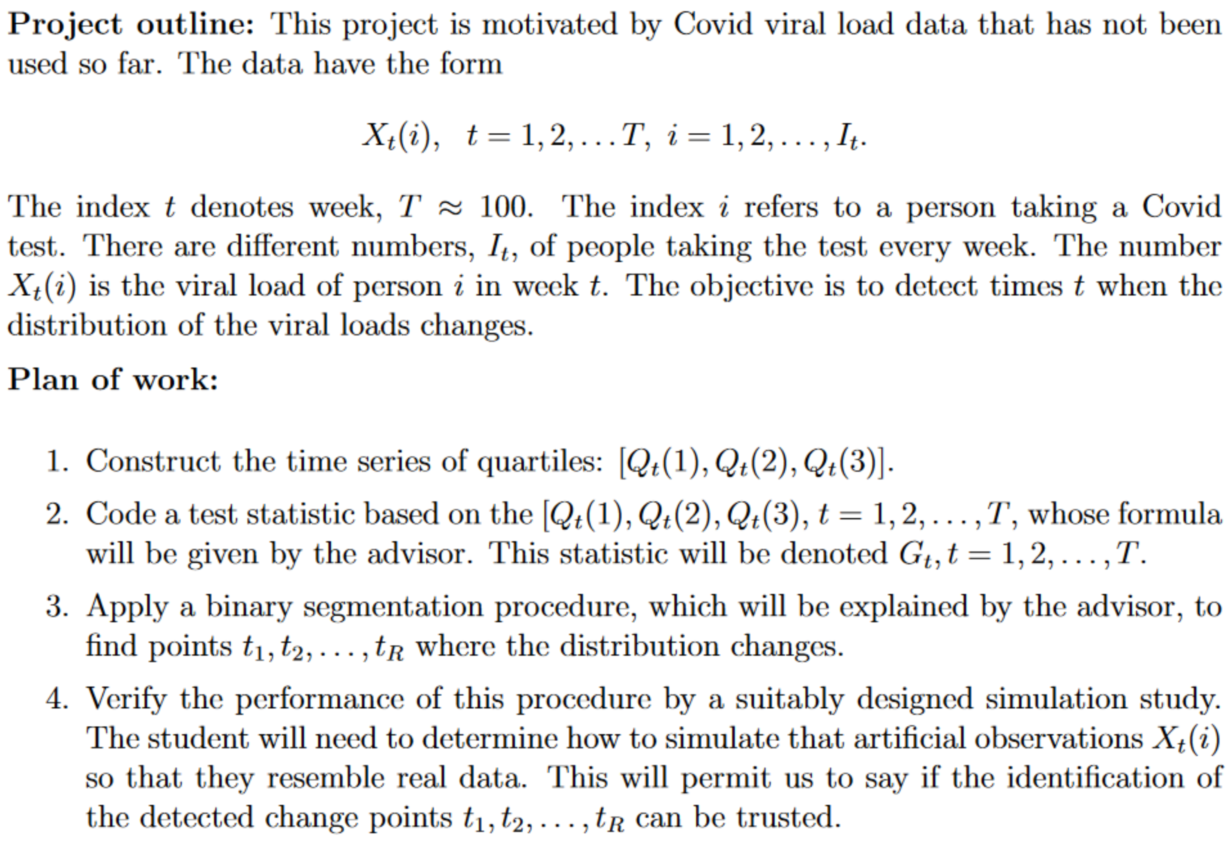
1. The Fast and the Furious: Tracking the Effect of the Tomoa Skip on Speed Climbing (Andee Kaplan)
2. Change point detection in a time series of distributions (Piotr Kokoszka)
3. What Do Philosophers Really Believe: A Statistical Analysis of Opinions of Current Philosophers (Aaron Nielsen)
4. Adventures in Softball Analytics (Aaron Nielsen)
5. Bayesian variable selection methods for criminal justice research (Matt Koslovsky)
6. Effect of air quality index on dog's eyes (Ann Hess)
7. How well do models of publication bias in meta-analysis reflect reality? (Ben Prytherch)
8. Assessing regression to the mean effects in Colorado CMAS testing results (Ben Prytherch)
9. Optimizing refresh speed in R-Shiny simulation apps. (Ben Prytherch)
10. Unmeasured Spatial Confounding with Count Data (Kayleigh Keller)
11. Barn Swallow Characteristics (Kayleigh Keller)
12. Exercise Program for Cancer Survivors (Ann Hess)

Detailed project descriptions:

1. **The Fast and the Furious: Tracking the Effect of the Tomoa Skip on Speed Climbing**
Advisor: Andee Kaplan
Requirements: Proficiency in (and enthusiasm for) R or python, including tidyverse, web scraping, and visualization.
Preferred: Experience with climbing a plus. Interest in time series also a plus.

Sport climbing is an athletic discipline comprised of three sub-disciplines – lead climbing, bouldering, and speed climbing. These three sub-disciplines have distinct goals, traditionally resulting in specialization of athletes into one of the three events. The year 2020 marked the first inclusion of sport climbing in the Olympic Games. While this decision was met with much excitement from the climbing community, it was not without controversy. The International Olympic Committee had only allocated one set of medals for the entire sport, necessitating the combination of the three sub-disciplines into one competition. Due to this format decision, many athletes who specialized in lead climbing and bouldering were forced to train and compete in speed climbing for the first time in their careers. One such athlete was Tomoa Narasaki, a World Champion boulderer from Japan, who introduced a new method of approaching the speed event that had never been attempted before. This approach, deemed the "Tomoa Skip," was subsequently adopted by many of the top speed climbers. Concurrently, many speed records began to fall at a seemingly more rapid rate (from 5.48 seconds in 2017 to the current record of 5.009 seconds in 2022). Speed climbing involves both timed and head-to-head competitions where climbers must ascend a 15 meter wall with 5 degrees of overhang that contains the same pattern of obstacles every time. Due to this format, records can be compared across time, dating back to the first introduction of the standard speed route in 2007. In this project we will investigate the effect of the Tomoa Skip on the sport of speed climbing. We will first collect data via web scraping and manually cataloguing speed climbers' techniques before performing exploratory data analysis and incorporating methods from change point detection, time series, and bootstrapping to answer two questions: (1) Did the Tomoa Skip result in a decrease in speed times? and (2) Do climbers who utilize the Tomoa Skip have a higher risk of falling (and being disqualified)?

2. **Change point detection in a time series of distributions**
   Advisor: Piotr Kokoszka
   Requirements: Proficiency in R, understanding of basic concepts of probability and statistics including distributions, quantiles, test statistic, significance level of a test and the sampling distribution of an estimator.



Project outline: This project is motivated by Covid viral load data that has not been used so far. The data have the form

$$X_t(i), \quad t = 1, 2, \ldots T, \; i = 1, 2, \ldots, I_t.$$

The index t denotes week, $T \approx 100$. The index i refers to a person taking a Covid test. There are different numbers, I_t, of people taking the test every week. The number $X_t(i)$ is the viral load of person i in week t. The objective is to detect times t when the distribution of the viral loads changes.

Plan of work:

1. Construct the time series of quartiles: $[Q_t(1), Q_t(2), Q_t(3)]$.
2. Code a test statistic based on the $[Q_t(1), Q_t(2), Q_t(3), t = 1, 2, \ldots, T$, whose formula will be given by the advisor. This statistic will be denoted $G_t, t = 1, 2, \ldots, T$.
3. Apply a binary segmentation procedure, which will be explained by the advisor, to find points $t_1, t_2, \ldots, t_R$ where the distribution changes.
4. Verify the performance of this procedure by a suitably designed simulation study. The student will need to determine how to simulate that artificial observations $X_t(i)$ so that they resemble real data. This will permit us to say if the identification of the detected change points $t_1, t_2, \ldots, t_R$ can be trusted.

3. **What Do Philosophers Really Believe: A Statistical Analysis of Opinions of Current Philosophers**
   Advisor: Aaron Nielsen
   Requirement: STAT341

Starting in 2009, PhilPapers, an online database of philosophy research, began surveying professional philosophers on their philosophical views. Philosophers David Bourget and David Chalmers have a forthcoming paper analyzing the results of the 2020 survey. For more details, see https://philarchive.org/archive/BOUPOP-3

This project will require a student to build an R Shiny app to visualize and explore the PhilPapers survey data. The student's final presentation will include sharing the app with an audience as well as summarizing the results of the survey. The student will also have the option of co-presenting the results to CSU's Philosophy Department in the fall.

4. **Adventures in Softball Analytics**
   Advisor: Aaron Nielsen
   Requirements: Sports Stats 1 or 2 or STAT342

   Description: Baseball analytics have been rapidly developing over the past decade. High speed cameras in baseball stadiums collect a plethora of data including pitch movement, pitch spin, batted ball exit velocity, and more. This technology is now becoming more common in collegiate softball programs. The CSU softball program is excited to collaborate with our department on statistical analyses using this camera captured data.

   The topics of the project will be decided by the student in consultation with their advisor and the softball team. Some possible projects include:

   (1) analysis of pitch effectiveness using spin rate, pitch movement, etc.
   (2) analysis of pitch sequencing (what order should pitches be thrown)
   (3) optimizing fielding positions (CSU utilizes a 5-person infield for some opposing hitters)
   (4) analysis of historical softball games using ELO rating to measure a program's strength
   (5) calculating "linear weights" for collegiate softball over time and using it to evaluate hitters with wOBA (weighed on-base average)

5. **Bayesian variable selection methods for criminal justice research**
   Advisors: Matt Koslovsky and Victoria Terranova
   Requirements: STAT 440, R coding experience

   What happens early in the correctional process in the pretrial phase impacts later sentencing outcomes. Many features of the pretrial phase carry financial obligation to the court. This means that pretrial defendants who lack financial means may experience a more severe sentence.

   The goal of this project is to identify features of the pretrial phase that are associated with sentencing outcomes (e.g., sentence length, fees) using advanced Bayesian methods for variable selection. The student working on this project will gain experience in criminal justice research, wrangling data, building hierarchical Bayesian models in R, and interpreting/reporting results.

6. **Effect of air quality index on dog's eyes**
   Advisor: Ann Hess
   Requirements: STAT341, 342

   This is a collaborative project with faculty from Clinical Sciences. The goal of the study is to understand the effect of air quality index (AQI) on dog's eyes. This study follows 15 dogs over 4 time points to see if environmental factors have an impact on their normal ocular surface parameters. The student working on this project will gain practical data analysis experience including summary statistics, visualization, mixed model analysis and interpreting results.

7. **How well do the assumptions underlying models of publication bias reflect reality?**
Advisor: Ben Prytherch
Required courses: STAT 341 and 342
Preferred course: DSCI 335

"Publication bias" refers to systematic differences between outcomes in the subset of scientific research that gets published and the broader set of scientific research that is actually performed. One form of publication bias is selection for statistical significance, by which primary research outcomes are more likely to be made public if they produce "p < 0.05" than if they produce "p > 0.05". This, in turn, leads to upward bias in published effect size estimates.

The theory behind how publication bias affects published outcomes is straightforward, but measuring this effect is not. A 2020 study attempted to do so by comparing meta-analyses (whose results may be influenced by publication bias) to pre-registered replication studies (whose results shouldn't be influenced by publication) found dramatic differences in effect size estimates of similar phenomena between the two approaches. Pre-registered replication studies were found to produce smaller estimated effect sizes than their respective meta-analyses that attempted to estimate the same effects. The authors proposed that this difference was due to publication bias. In 2022, a critique of this paper was published, claiming that publication bias could not fully account for the observed differences in effect sizes estimated via replication studies vs. via meta-analysis. This claim was supported by a statistical re-analysis under a "worst case scenario" model of publication bias in which only the set of non-significant results in the meta-analyses were analyzed, some of which still produced effect size estimates larger than those in the replication studies. This, the critique says, demonstrates that it is not possible for publication bias to fully account for the discrepancy, since the non-significant outcomes could not have been the product of publication bias.

This "worst case scenario" model makes the seemingly reasonable assumption that non-significant results in published papers cannot have gotten there via selection for statistical significance. The goal of this research project is to investigate how reasonable this assumption is. The student researcher will perform a systematic review of the original studies which went into the meta-analyses in question, and use existing meta-analytic tools for diagnosing publication bias. The student will:

- Track down and organize papers used in meta-analyses.
- Identify which results from original papers were used in the meta-analyses.
- Learn to use some meta-analysis tools for diagnosing publication bias

Preferred qualifications:

- Willingness to independently learn how a variety of meta-analysis models work
- Willingness to independently learn how to implement R packages for meta-analysis
- Willingness to independently research current controversies and open questions in the use of meta-analysis
- An interest in reading through results sections of published papers to uncover details about how the input values for meta-analyses are selected

8. **Assessing regression to the mean effects in Colorado CMAS testing results**
Advisor: Ben Prytherch
Required course: STAT 341
Preferred courses: STAT 342, DSCI 335, STAT 400

The Colorado Measures of Academic Successs (CMAS) assessment is a statewide standardized test given to students in the 3rd through 12th grades. Test results are used to:

- Give parents and guardians information about their child's educational progress
- Help schools evaluate teaching effectiveness
- Help the Colorado Department of Education assess schools and school districts. This includes using scores to identify under-performing schools and school districts, and potentially intervene.

Colorado uses a statistical model that quantifies student growth year-to-year relative to other students who have had similar scores in the past. This is meant to produce fairer interpretations than would be produced using raw test scores. There is existing literature on how regression to the mean can bias the interpretation of changes in test scores over time, and there are tools for statistically reducing this bias, but these are not used in Colorado.

For this project a student (or students) will use publicly available CMAS data and simulation tools to investigate whether or to what extent regression to the mean effects in CMAS results create bias in assessments of teacher, school, and district performance.

This project will require:

- Locating, organizing, and analyzing publicly accessible CMAS data.
- Reading documentation for the R package used to analyze CMAS data; learning to use it and reading code to see how its functions work.
- Reading papers on what is currently known about regression to the mean effects in testing data, and methods for reducing their biasing effects.
- Writing code that simulates test score data under various assumptions, then running analyses and identifying how assumptions about the data generation process affect statistical outcomes.
- Investigating state of Colorado policies and procedures on the use of CMAS results in accountability review for teachers, schools, and school districts.

9. **Optimizing refresh speed in R-Shiny simulation apps**
Advisor: Ben Prytherch
Required courses: STAT 341 and STAT 342
Preferred courses: DSCI 336, any CS or DSCI courses involving benchmarking and optimizing code

"Shiny" is an R package used to create interactive apps using R. On an app, the user can change values of input variables (using, for example, a slider on a number line) and the app will refresh the output that it displays (e.g. plots and tables of summary statistics). Shiny has a set of internal rules for observing changes in input values and determining which parts of underlying code need to be re-run in order to refresh the output shown in the user interface. There are additional packages which allow Shiny app developers to monitor an app as its inputs change and create a log showing time elapsed between each step in the refresh process.

In this project, a student will initially identify inefficiencies in apps currently used in CSU STAT classes and improve their refresh time. From there, the student will devise test apps of varying complexity and use these to create a set of concise and accessible guidelines to assist future app developers. To succeed in this project, the student will need to:

- Learn the basics of using the Shiny package to create apps.
- Independently learn the underlying structure of how a Shiny app responds to changes in input variables to refresh itself.
- Research existing literature on Shiny app optimization.
- Devise tests to identify how various elements in a Shiny app affect refresh time.
- Create a report that provides guidance for Shiny app developers who are concerned with refresh speed.

Preferred qualifications:

- Sufficient background in statistics and coding (R or otherwise) to be able to read and understand documentation for R packages
- Completion of computer science courses beyond the introductory level
- Experience optimizing code to reduce processing time
- A strong understanding of the statistical methods and models taught in STAT 341 and 342
- Experience (or at least interest in) writing documentation to convey to other app developers what you have learned.

10. **Unmeasured Spatial Confounding with Count Data**
    Advisor: Kayleigh Keller
    Required Course: STAT 341
    Preferred courses: STAT 342, STAT 400

    Unmeasured spatial confounding is a challenge in estimating the health effects of environmental exposures. Any unmeasured factors that vary across space and are related to health outcome can bias the estimated exposure-response relationships. Flexible semi-parametric methods exist for accounting for this bias, but have been primarily developed for continuous outcome data. This project will investigate the performance of these methods with Poisson-distributed outcomes. This will be primarily done via simulation, although analysis of large cohort studies may be possible.

    Experience with R at the level of STAT 400 and familiarity with concepts of epidemiology are preferrable.

11. **Barn Swallow Characteristics**
    Advisor: Kayleigh Keller
    Required course: STAT 341
    Preferred courses: STAT 342, STAT 460, DSCI 235

    Barn swallows are social birds and their characteristics can vary across different sub-populations. Using field data collected from multiple sites across multiple years, this project will investigate differences in bird characteristics (e.g., morphology, plumage). Additional lines of inquiry may include relationships between (i) bird characteristics and social behavior and (ii) nesting locations and predation. This project will involve data cleaning and harmonization, extensive descriptive statistics, and some inferential comparisons. Familiarity with R is required; experience with biology and network statistics is preferrable.

12. **Exercise Program for Cancer Survivors**
    Advisor: Ann Hess
    Required course: STAT341

    This is a collaborative project with faculty from Health and Exercise Science. A group of cancer survivors participated in an 8 week exercise program. Investigators are interested in who benefits the most in terms of increased physical activity. Hence, the response variable is change in physical activity (after vs before the program). Several predictors will be considered including pre-program physical activity, cancer treatment status, demographics, etc. The student working on this project will gain practical data analysis experience including summary statistics, visualization, regression analysis, model selection and interpreting results.