**Paid Statistics Research Opportunities for CSU Undergrads!**

During Summer 2024, 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 $1250.
- The schedule is flexible, but it is expected that students will meet with faculty advisors at least once a week and spend up to 9 hours per week working on the project between **05/14/24 – 06/28/24**.
- 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 summarize and present their work.
- Students are strongly encouraged to submit their work at CSU CURC (Celebrate Undergraduate Research and Creativity) and/or MURALS (Multicultural Undergraduate Research Art and Leadership Symposium) in Spring 2025.

**Requirements:**
1. Be a CSU undergraduate student in Fall 2024.
2. Be available to work and meet with faculty between 5/14 – 6/28.
3. Apply by Sun 4/21. Your application should include a personal statement about why you are interested in this research opportunity. Students will be evaluated on their enthusiasm as well as if they meet the required and preferred qualifications.

**Preference will be given to students:**
1. Majoring or minoring in Statistics or Data Science.
2. Planned graduation of Spring 2025 or later.

**Apply here:**
[https://forms.microsoft.com/Pages/DesignPageV2.aspx?subpage=design&token=e56d632a27c048559259a02062b5ccac&id=Aoi1r3r_sUurITZ_8uz8iv3jgZE89e5PIOhd2ks6bhUMTBO1EYVQj1PTEpKMIkFZUUVZUzM5MklWTS4u](https://forms.microsoft.com/Pages/DesignPageV2.aspx?subpage=design&token=e56d632a27c048559259a02062b5ccac&id=Aoi1r3r_sUurITZ_8uz8iv3jgZE89e5PIOhd2ks6bhUMTBO1EYVQj1PTEpKMIkFZUUVZUzM5MklWTS4u)

Summaries of potential projects and detailed descriptions start on the next page.
Summary of project titles and advisors:

- Mixed models for ordinal outcomes with application to equine pre-purchase surveys (Tianjian Zhou)
- Extreme Fire Danger Conditions in the Western US (Ben Shaby)
- College Football Analytics (Aaron Nielsen)
- Integrating multi-source datasets with unignorable missingness (Tianying Wang)
- Statistical Learning in Cancer Genomic Studies (Tianying Wang)
- Quantum Computing: First Exposure (Yunpeng Zhao)
- Simulation of point processes (Dongzhou Huang)
- Tracking sea ice with remote sensing data (Karissa Palmer and Yawen Guan)
- Network inference from grouped observations using hub models (Yunpeng Zhao)
- Visual processing of information (Ann Hess)
- Probiotic Gut Health (Ann Hess)
- Timescale Effects for Environmental Exposures (Kayleigh Keller)
- Early assessment of FCPD's Homeless Outreach and Proactive Engagement (HOPE) program (Ben Prytherch)
- When do journalists report effect sizes? (Ben Prytherch)

Detailed descriptions appear below.

**Mixed models for ordinal outcomes with application to equine pre-purchase surveys**

**Advisor:** Tianjian Zhou  
**Requirements:** STAT341, proficiency in R, proficiency in linear regression  
**Preferred:** STAT342, STAT460

In standard linear mixed models, the assumption is that the error terms and random effects follow a normal distribution, making them ideal for analyzing continuous outcome data. However, when dealing with ordinal outcomes, such as those commonly encountered in surveys with Likert scale responses, the validity of conclusions from standard linear mixed model analysis is debated. To address this, various modifications and extensions to the standard linear mixed model have been proposed, including proportional odds, adjacent category, stereotype logit, and continuation ratio models, tailored specifically for ordinal outcomes within the context of mixed models.

In this project, the student will gain proficiency in analyzing data using standard linear mixed models and fitting such models in R. They will then explore alternative options for ordinal outcomes and identify suitable R packages for model fitting. Finally, the student will apply both a standard linear mixed model and an ordinal outcome mixed model to a dataset of survey radiographs for equine pre-purchase examinations, comparing the results obtained from each approach.
Extreme Fire Danger Conditions in the Western US  
**Advisor:** Ben Shaby  
**Required Courses:** STAT 341  
**Preferred Courses:** STAT 430, 440

It is well known that in the Western US, a small percentage of wildfires cause a huge percentage of the damage; that is, the extreme events drive most of the risk. But what drives the extreme events? We've found that the Fire Weather Index (FWI), which summarizes fire potential due to weather conditions, is predictive of the sizes of extreme wildfires. So is Energy Release Component (ERC), which summarizes the potential heat from available fuel. FWI and ERC capture different aspects of fire danger conditions. A reasonable conjecture might be that the most dangerous conditions occur during contemporaneous very high values of both FWI and ERC. We will try to figure out whether the probability of very large FWI and ERC have changed over time, as well as the more interesting (and more challenging!) question of how likely they are to be very large at the same time.

This project will require:

- Learning the basics of extreme value analysis
- Learning cutting-edge aspects of extreme value analysis
- Finding, downloading, and wrangling weather and ERC data from public repositories
- Producing exploratory plots of the data to inform subsequent modeling choices
- Coming up with plausible models for time trends in the extremes
- Writing R code to fit the models
- Summarizing and visualizing the results

Preferred qualifications:

- Familiarity with Bayesian data analysis
- Experience writing well-organized R code

College Football Analytics (2 students)  
**Advisor:** Aaron Nielsen  
**Required course:** STAT 342  
**Preferred courses:** STAT 351/451

This project requires two students to work with the football teams at Colorado State University and/or University of Nebraska. Students will work with these football teams’ analytics specialists for a variety of tasks. The topics of the project will be decided by the student in consultation with their advisor and the football team. Some of the tasks may include: analyzing and classifying offensive formations using Pro Football Focus (PFF) software, producing age curves to help players project their size in advance of the NFL combine, assessing play calling strategies using an expected points added framework, and updating weekly scouting reports to summarize strengths and weaknesses of opponents. Students that are successful with this project may have
the opportunity to continue in their role as interns with the football team and/or continue their analyses as an independent study during the 2024 fall season.

---

**Integrating multi-source datasets with unignorable missingness**

**Advisor:** Tianying Wang  
**Requirements:** Proficiency in SQL or Python, understanding of basic concepts of statistics and database management.

Project background: The informative missing (or unignorable missing) data is a well-recognized issue in various disciplines, including epidemiology, bioinformatics, engineering, climate science, etc. In recent years, increasing literature has acknowledged the necessity of recognizing such missingness and modeling the mechanism of it. Many researchers in domain sciences have also been aware that ignoring such an issue or treating such missingness as random missing will lead to bias estimation and invalid statistical inference, which will severely impact the decision-making process.

Project objectives: this project aims to build a database to link and match different sources of data, identify the missing pattern, and provide a data dashboard for integrating multiple data sources. This project will provide students with hands-on experience in data processing and analysis. Students are strongly encouraged to submit their work at CSU CURC (Celebrate Undergraduate Research and Creativity) and MURALS (Multicultural Undergraduate Research Art and Leadership Symposium) in Spring 2025. Depending on students’ availability, follow-up research aims to include modeling informative missingness and developing a statistical framework for inference with applications in public health.

---

**Statistical Learning in Cancer Genomic Studies**

**Advisor:** Tianying Wang  
**Required Courses:** STAT 341  
**Preferred Courses:** STAT 430, 440  
**Requirements:** Proficiency in R

Project background: With tremendous advances in biotechnologies, vast amounts and various types of data are generated in genomics, transcriptomics, metabolomics, etc. This provides excellent opportunities for researchers to learn more comprehensive underlying disease progressions and their associations with genetic and genomic information. However, statistical learning in large-scale genetic data raises many questions and challenges due to computation and heterogeneity.

Project objectives: This project aims to learn genomic associations and structures using Genomic Data Commons. By harmonizing genomic and clinical data, including demographics, diagnosis, and treatment information, students will get hands-on experience in learning and applying statistical models in bioinformatics. Students are strongly encouraged to submit their work at
CSU CURC (Celebrate Undergraduate Research and Creativity) and MURALS (Multicultural Undergraduate Research Art and Leadership Symposium) in Spring 2025. Depending on students’ availability, follow-up research aims to include modeling longitudinal or functional data with applications in genomics and epidemiology.

Quantum Computing: First Exposure
Advisor: Yunpeng Zhao
Requirements: Linear algebra (DSCI/MATH 369), experience in programming

Quantum computers have the potential to perform high-speed computations based on a fundamentally different manner of storing and processing data – quantum superpositions and unitary transformations. Although reliable quantum computing hardware is still in development, the mathematical framework for quantum algorithms has been well-established, and a number of highly impactful algorithms under this framework have been proposed, such as Shor’s algorithm for integer factorization and Grover’s algorithm for search.

This project will guide the students in implementing two milestone quantum computing algorithms: Shor’s algorithm and Grover’s algorithm. The students will have the opportunity to build quantum circuits and algorithms using Qiskit, a Python SDK for quantum computing developed by IBM Research.

Simulation of point processes
Advisor: Dongzhou Huang
Course requirement: STAT 341/STAT 342. Taking STAT 420 would be advantageous.

Point processes have garnered considerable attention recently, owing to their extensive applications in queueing systems, neural networks, and image analysis. For instance, in queueing systems, point processes are utilized to model customer arrivals, while in neural networks, neuronal activity can also be conceptualized as a point process. Before delving into theoretical examinations of their statistical properties, it is prudent to pause and employ computational methods to verify them. To facilitate this, simulating point processes becomes essential. Therefore, the main task of this project is to propose methods to simulate point processes including inhomogeneous Poisson processes.

The first goal of this project is to simulate Poisson process, employing its fundamental property: the interarrival times follow independent and identically distributed exponential random variables. Subsequently, we aim to progress to simulating inhomogeneous Poisson process with a given rate function. If time permits, we will also study the Hawkes process, as well as its simulations.

Tracking sea ice with remote sensing data
Advisors: Karissa Palmer and Yawen Guan
Required: STAT 341; proficiency in R, be able to independently learn and problem-solve through utilization of R package documentation.

Preferred: STAT 342 and STAT 400

Arctic sea ice regulates the heat transfer between the ocean and the atmosphere, playing a significant role in the global climate. When narrow linear cracks, known as leads, form within the sea ice, the heat from the ocean is transferred to the atmosphere. Data for studying the Arctic are primarily collected from satellite imagery, which are processed to provide sea ice motion, enabling the tracking of sea ice motion and the monitoring of its trajectories over time. This project involves visualization and modeling of sea ice trajectories to infer the presence and characteristics of leads. The student will gain experience in data visualization and analysis, and learn how to use simulation studies and metrics for evaluating statistical models.

Network inference from grouped observations using hub models
Advisor: Yunpeng Zhao

Requirements: Experience programming in R

Social network analysis presupposes that observed social behavior is influenced by an unobserved network. Such analysis has also been applied to research in animal behavior and recommender systems. Traditional approaches to inferring the latent network use pairwise descriptive statistics that rely on a variety of measures of co-occurrence. While these techniques have proven useful in a wide range of applications, the literature does not describe the generating mechanism of the observed data from the network.

The mentor and his collaborators have proposed a model-based approach – called hub models – to infer implicit network structure from temporal-independent/dependent grouped data. The methods were originally implemented in Matlab. R is a free programming language that has been widely used in statistical computing. The research assistant will assist the mentor in developing an R package, including hub models and other traditional non-model-based approaches from social sciences. This will improve the visibility and dissemination of this research for practitioners.

Visual processing of information
Advisor: Ann Hess

Requirements: STAT341, 342

This is a collaborative project with faculty from Psychology. The goal of the study is to understand factors that affect visual processing of information. Participants are given the task of identifying the center location on a graph with varying features (color, shape, spread, etc). The student working on this project will gain practical data analysis experience including visualization, mixed logistic regression analysis and interpreting results.

Probiotic Gut Health
**Advisor:** Ann Hess  
**Requirements:** STAT341, 342

This is a collaborative project with faculty from Food Science and Human Nutrition. This is a double blind, randomized parallel arm (3 arms) clinical intervention study with measurements taken at baseline and after 6 weeks of a probiotic or placebo intervention. The primary outcomes include daily stool records and gastrointestinal health questionnaires. Secondary outcomes include inflammatory markers in blood and stool, stool short chain fatty acids, and gut microbiota. The student working on this project will gain practical data analysis experience including visualization, mixed model analysis and interpreting results.

---

**Timescale Effects for Environmental Exposures**  
**Advisor:** Kayleigh Keller  
**Required Course:** STAT 341  
**Preferred courses:** STAT 342, STAT 400

Long-term exposure to environmental pollutants has been linked to many different adverse health outcomes. However, most studies use annual or multi-year averages that ignore information about the temporal structure of exposure. Is continuous exposure to moderate levels better or worse than a series of short-term exposures to higher levels? This project will use simulations to evaluate the differing impacts of exposure patterns on estimated health effects. Experience with R at the level of STAT 400 and familiarity with concepts of epidemiology are preferable.

---

**Early assessment of FCPD's Homeless Outreach and Proactive Engagement (HOPE) program**  
**Advisor:** Ben Prytherch  
**Required Course:** STAT 341  
**Preferred courses:** STAT 342, DSCI 235 or proficiency with dplyr package or other data wrangling tools

In April 2023, the Fort Collins Police Department established the Homeless Outreach and Proactive Engagement (HOPE) team, whose priorities are “providing individuals with services/resources, promoting safety, and addressing environmental concerns” with respect to people in Fort Collins experiencing homelessness. The HOPE team has collected data from multiple sources, and would like to use this data to assess the effectiveness of the program.

This project will involve a lot of data exploration and wrangling (combining data from multiple sources and re-formatting various ways, depending on the desired analysis). The project will also involve working with members of the HOPE team to decide what kinds of analyses are best suited for answering their questions of interest.
For this project to be a success, the student researcher will need good communication skills, the ability to work independently, and an openness to trying out ideas without knowing whether they will work. There is no existing statistical analysis plan; putting one together will be a major element of this project. Students with an interest in public policy, community policing, and homelessness are encouraged to apply.

---

**When do journalists report effect sizes?**

**Advisor:** Ben Prytherch  
**Required Course:** At least one statistics course  
**Preferred courses:** STAT 341, DSCI 335

Journalists and book authors who write about scientific research regularly share the results of quantitative studies with their readers. Often, these results are reported in a categorical or directional manner only (e.g. "people exposed to _____ were more likely to report ____."). Sometimes though, effect size magnitudes are reported (e.g. "people exposed to _____ were 20% more likely to report ____."). In this project, the student researcher will look through articles and books aimed at general audiences, recording when and how results from quantitative research studies are reported. The student will then look up the original studies, comparing how the results were originally reported to how they were subsequently reported. Are there some fields in which effect size magnitudes are more likely to be reported? Are effect sizes more likely to be reported when they are larger? To what extent does the manner in which published results are originally described (such as in the abstract or conclusion) influence the way they are described in popular media?

This project will involve first creating a data collection and analysis strategy to be pre-registered, then collecting data by reading through popular articles and books and then looking up the original published papers, and then implementing the pre-registered strategy. Students interested in statistical communication and open science practices are encouraged to apply.