Statistics Department
Senior Project Ideas
2017-18
September 13, 2017

Supervisor: Prince Afriyie

1. **Multiple Hypothesis Testing.** When we reject the null hypothesis in an ANOVA F-test, knowing that at least one of the population means differ leads to the question of which pair of groups have different means. The chances of committing Type I error increases as the number of tests increases. Methods that deal with this issue are called multiple comparisons procedures (MCP). Failure to compensate for multiple hypotheses testing can have important real-world consequences. The project entails literature review of several multiple testing procedures, simulation in R and application to real life data (high dimensional genomic data - microarray or RNA-sequencing).
   Necessary Background: STAT 305 and STAT 331. (STAT 425, 426, 427 would be helpful but is not required).

2. This project is a corollary to the first. We will develop an R package for various multiple testing procedures (there is already an R package for MCP, but it does not have the recent multiple testing procedures in literature).
   Necessary Background: STAT 331.

3. **Characterization of soil microbial communities and scion development as influenced by sustainable vineyard practices.** California's Central Coast viticulture and enology industry has been expanding rapidly and has resulted in significant economic growth, especially within the Edna Valley American Viticulture Area (AVA). The sudden expansion of grape acreage, the prolong drought conditions and the close proximity to other farmland and urban areas, makes environmentally conscientious agricultural practices paramount to the long-term viability of the wine and viticulture industry. Sustainable vineyard practices have been defined by the California Sustainable Winegrowing Alliance as being environmentally sound, economically feasible and socially equitable. There is a public demand and genuine industry interest in adapting more sustainable practices such as cover cropping as a natural means to reduce vine vigor, build organic matter, attract beneficial insects and impact soil water availability. Other practices considered environmentally sensitive including the use of cultivation in place of herbicide applications, and the application of organic fertilizer, including grape pomace (discarded skins and seeds).
   Goal: The goal is to understand how vineyard soil health, vine development and wine quality are influenced by environmentally sensitive farming practices.
   Role: Dr. Jean Dodson Paterson (Viticulture) and Dr. Katherine Watts (Chemistry) have enlisted some undergraduate students to collect soil samples and RNA-sequencing data from this experiment. Your role is to help make sense of the data (data analysis!).
   Necessary Background: STAT 301, STAT 302 and STAT 323.

Supervisor: Matt Carlton

1. How do you create simultaneous 95% CIs for multinomial proportions? How do you determine the sample size required to estimate all the proportions to within a prescribed margin of error at 95% simultaneous confidence? And what if you're sampling from a relatively small finite population?

   The project would begin with a literature search, involve a heavy amount of simulation, and then (possibly) some theoretical results.

   Necessary Background: Basic computing, familiarity with simulation. Completion of STAT 418 preferred but not required.

Supervisor: Beth Chance

1. Two recent articles outlined a set of lessons that could be used to explore randomness and statistical inference with sixth grade students and using "design-based research" to develop statistical knowledge for teaching. In this project we would build on these articles to develop/modify lessons to use with 4-6th graders, deliver the lessons with students at Teach Elementary, and evaluate their effectiveness.
   Recommended Background: Completion of Stat 410 is preferred but not required.

2. Research assistant to help with NSF grant evaluating student learning and attitudes with a randomization-based curriculum using hierarchical linear models. The project would involve learning about HLMs, applying such models to the data, and comparing student performance across institutions. Data can be cross-classified with students' prior knowledge and attitudes about statistics, and other demographic information. The data analysis will be the basis of a journal article in statistics education. Ideally would also help with development of Stat 414 on HLMs to be offered for the first time in Winter.
   Necessary Background: Stat 324 and facility with R are required.

3. An investigation of the bootstrap methodology and its appropriateness for the introductory curriculum. Doing some reading on bootstrapping (I can point you to some classic and new textbooks explaining the idea) including some of the different implementations for bootstrapping (e.g., percentile intervals vs. fancier stuff). The project will involve simulations involving bootstrapping (does it do what they claim, comparing methods); investigation of some ways to incorporate bootstrapping into an intro course (trying out some different applets, technologies that are out there now, see what might be intuitive for an intro student) including the debate of whether to use bootstrapping for intervals and other techniques for testing or to use bootstrapping for sampling and other techniques for random assignment; making some final recommendations as to whether, and if so how, bootstrapping should be taught in the intro course, possibly instead of t-intervals. This could involve testing out some activities in a class like Stat 217 during the year.

Supervisor: Jimmy Doi

1. There are many methods that have been proposed for generating confidence intervals for the binomial proportion π. We typically assess the performance of a confidence interval method by determining coverage probability and expected or average length properties. One popular method was developed by Clopper and Pearson (1934) however their method yields confidence intervals that can be very long and have very high coverage. In a recent paper by Thulin (2014) the author investigates modified versions of the Clopper Pearson interval that yield better length and coverage properties. We will go through this paper and develop R code to reproduce the paper's results. Finally, using your R code you will develop a corresponding app in Shiny.
   Thulin (2014), "Coverage-adjusted Confidence Intervals for a Binomial Proportion", Scandinavian Journal of Statistics, Vol. 41: 291—300.
   Recommended Background: Completion of STAT 331 (with a B or better). STAT 418 would be helpful but is not required.

2. When using log-linear/Poisson regression to model count data we assume the underlying reference distribution is Poisson. However there are times when data exhibit greater variability than expected for a generalized linear model. This phenomenon is known as overdispersion. To address this issue, instead of using Poisson regression we can turn to negative binomial regression. This topic is at most briefly discussed in our categorical data analysis course. To better understand this rich topic we will do selected readings/investigations from Negative Binomial Regression (2011) by Joseph Hilbe (2nd Edition, Wiley). All coding investigations will be done in R. Finally, using your R code you will develop corresponding apps in Shiny.
   Necessary Background: Completion of STAT 331 (with a B or better) and STAT 418 (with a B or better).

Supervisors: Samuel Frame

1. **Loan-Level Loss Forecasting Models**. Since the financial crises and Great Recession, banks are required to more rigorously evaluate the riskiness of their portfolios and assets. The stress tests are usually based on loan-level statistical models that are designed to forecast defaults and losses. The Frannie Mae Single-Familty Loan Data is publicly available historical loan acquisition and (now) performance data that is very similar to most of the data banks use (economic data can be obtained elsewhere). These data are fairly large and may be of interest to data science students, and the past two students who have worked on this utilized distributed computing. The goal of this project is to investigate loss forecasting methods and develop a loss forecasting model.
   Necessary Background: CSC 101, CSC 202, STAT 324, STAT 331, STAT 418 (and a whole pile of data science classes if you want to go that route)

2. **Starmine Predicted Earnings Surprises.** StarMine is a subsidiary of Thomson Reuters specializing in quantitative analytics, and one of their metrics is the StarMine SmartEstimate which is StarMine's predicted earnings per share. The predicted earnings surprise is the difference between the Starmine

SmartEstimate and the consensus average (at the date of publication). After the NYSE and NASDAQ exchanges close every day, Thomson Reuters publishes *The Day Ahead* newsletter (TDA). On the first trading day each week, the TDA StarMine's SEPS for selected companies.

I have been collecting this data for several years. Myself and several students have investigated the relationship between the surprise earnings and short-term returns, and also forms trading algorithms based on the surprises. I have two years of data that need to be processed, analyzed, and the results compared to the current work (I believe there was an intervention in the how the surprises are selected for inclusion in the TDA).

I am also very curious about how accurate these predictions are. Their sales materials indicate they are 80% accurate (accurate defined as the prediction is in the direction of the actual earnings). It is possible to get historical earnings data from various places including COMPUSTAT, and then compare the published predictions to the actual predictions (including trends in revisions of several consecutive releases).
Necessary Background: STAT 324, STAT 331

Supervisor: Hunter Glanz

1. **It's the Time of the Season for Modelling**: What do seasons look like to a satellite? This project will involve collecting and organizing data from a Remote Sensing instrument such as MODIS or Landsat, and investigating the times at which vegetation begins to grow (greening) and fade (browning) each year. And have these times changed over time? How do these times vary by vegetation type and latitude? Is there any more to be gained by including information about land surface temperature?
   Necessary Background: STAT 330/331, STAT 324, STAT 419

2. **(Social) Network Comparison Investigation:** Suppose we have two social networks, constructed from co-occurrence data. That is, two individuals in a network are connected if they have been observed together. These two networks can be described and compared in a number of ways, but can we evaluate the significance of any differences observed? How? This project will investigate this via simulation and other methods.
   Necessary Background: STAT 330/331, STAT 305

3. **Participatory Mapping - Democratizing City Planning:** Dr. Greg Brown in the Environmental Science Department is interested in investigating numerous questions including those surrounding the theory of NIMBY-ism (Not In My BackYard). Data have been collected and continue to be collected from residents of several areas pertaining to their perceptions, experiences, and preferences about their neighborhoods and city resources. Join this project to help investigate some very unique and interesting geographic survey data.
   Necessary Background: STAT 330/331, STAT 324

4. **Data Science Specialization:** Johns Hopkins University offers a specialization in data science comprised of nine free online courses through Coursera. The series of month-long courses gives an introduction to the basic tools used by data scientists, training in good practices, and an overview of data science. The courses cover reproducible research, version control, R programming, data extraction and cleaning, graphics, machine learning, generalized linear models, and developing data products using Shiny. In this project, you would complete the nine courses and analyze a large data set of your choosing in a project that combines at least three new concepts you learned in the courses.
   Necessary Background**:** STAT 324, 331

Supervisors: Hunter Glanz and John Walker

1. **The Sabermetrics of eSports**: The world of competitive gaming has been growing steadily for some time, but in the last few years there have been some significant steps forward. While some of the biggest games in professional eSports are different game types, they are all ripe for analysis. Methods for valuing players and evaluating performance are needed! If you're interested collecting, organizing, and analyzing eSports data then join us.
   Necessary Background: STAT 330/331, STAT 324.
   Recommended Background: STAT 419, Any DATA courses

2. **Computer Gaming Related Senior Projects**: The results of many computer games can be simulated outside of the game itself. The student would choose a game, then either write code to simulate results from the game or use an existing simulator and then analyze the results. The goals of the project would depend on the game selected. For example, in the game World of Warcraft, there are two common programs to simulate a player's damage done. Are these two simulators equivalent? If not, under what conditions do the simulated results differ? What model can we use to predict a player's damage done based on the character's attributes? Some familiarity with the selected game is required, but actual game play is not since all results would be simulated outside of the game.
   Necessary Background: STAT 323, STAT 324, STAT 330, STAT 331
   Recommended Background: Other courses depending on the specific idea (e.g. STAT 425

Supervisors: Karen McGaughey and Soma Roy

1. This project is open to anyone with an interest in computer experiments.

What are computer experiments? Computer simulators are often used when it is either impossible or infeasible to observe the actual systems or processes. Examples of such processes include climate change, spread of infectious diseases, and car crashes. Often the simulators or codes are very complex,

requiring many hours or days for a single simulation, and thus the number of times we may implement the code to collect data needs to be as small as possible. The running of such a code at a few chosen input settings comprises a computer experiment. We will use our statistical knowledge to write code to replace or emulate the time-consuming computer simulators. Your senior project will involve, among other things, reading papers about design and analyses of computer experiments, writing algorithms and code for one or more topics of interest to you.
Necessary Background: Computing skills - R; SAS
Recommended Background: STAT 323, 324, 305, 425

Supervisor: Shannon Pileggi

1. Study design with appropriate sample size recommendations is essential for grant funding, yet the power calculations can be quite sensitive to the assumed parameter values. This is inspired by a researcher in the kinesiology department who wants to examine the effectiveness of bottle feeding an infant with a clear bottle (control group) versus an opaque bottle (treatment group). The primary aim is to determine if weight gain is lower in the treatment group, as assessed by weight for length z-scores (WLZ). In both the treatment and control group, the infants' WLZ is to be assessed at baseline and 4 week follow up, so at the end of the study we can compare the difference in the WLZ scores between the two groups. In this project we will do simulation to examine how the correlation in WLZ scores between the baseline and 4 week follow up affects the sample size and power calculations.
   Recommended Background: Completion of STAT 323 and either STAT 330 or 331

2. One of the reasons R rapidly became an indispensable part of statistical analysis is because of the ability for users to easily share and contribute content through packages. In this senior project, we will create an R package for functions that accompany Chance and Rossman's ISCAM textbook, as well as write a vignette to demonstrate usage. We will also investigate best programming practices in R, learn what is meant by reproducibility and how to achieve it, and compare using R workspaces versus packages.
   Necessary Background: STAT 331

3. How does active versus passive homework assignments affect student learning outcomes in STAT 217? In this senior project, you will design a research study from start to finish to compare an older, passive version of lab preparation assignments (watching videos) versus a newer, active version of lab preparation assignments (DataCamp). In fall 2017 we will seek IRB approval and design learning outcomes and assessment methods, in winter of 2018 we will execute the experiment with two sections of STAT 217, and in the spring of 2018 we will analyze the data.
   Recommended Background: STAT 410

4. A standard component of first class in survival analysis is to use parametric distributions to identify models for your data. This involves computing maximum likelihood estimates for parameters, using graphical techniques to evaluate fit, and using numeric methods to evaluate fit. While this

functionality is readily available in Minitab, it is not readily available in R. In this senior project we will build tools (R package with functions or Shiny app) to facilitate this process in R.
Recommended Background: STAT 417

Supervisor: Kevin Ross

1. Investigate an interesting topic in probability, by reading relevant research, doing some math, and running simulation studies. Here's just one example: The standard version of the "collector's problem" assumes that all the prizes are equally likely. But what happens if some prizes are harder to collect than others?
   Necessary Background: Stat 305, 331
   Recommended Background: Stat 425, 426

Supervisors: Kevin Ross and Dennis Sun

1. Contribute to the development of Symbulate, a Python package which provides a user friendly framework for conducting simulations involving probability models like those covered in Stat 305 and 405.
   Necessary Background: Stat 305, fluency in Python, strong software engineering skills.

Supervisor: Soma Roy

1. **Teaching statistics in K-12.** This project should be especially beneficial to someone who wants to go into teaching statistics in K-12.
   You will be involved in:
   - Conducting a review of where things stand at present with teaching statistics in K-12
   - Researching the topics taught in one of the following: Elementary school, Junior High, or High school
   - Researching the Common Core State Standards for Statistics
   - Writing learning tasks/activities that improve student understanding of statistical concepts.
   - Writing assessment tasks/activities.

   Recommended Background: STAT 410

2. **Using simulation to carry out Bayesian Analysis.** The premise of Bayesian statistics is to use prior knowledge in conjunction with data to make decisions and arrive at appropriate conclusions. This

project will involve learning and exploring concepts of Bayesian statistics, and building teaching tools that can help students understand and learn such concepts. Topics explored will include Bayesian inference (hypothesis testing and intervals) and how Bayesian inference methods differ from frequentist inference methods (those that you have seen in classes like STAT 301 and 302), and Bayesian experimental design. For example, we will explore the multi-armed bandit problem (https://en.m.wikipedia.org/wiki/Multi-armed_bandit), and create interactive visualization tools (most likely in *R*) that will help show what factors affect the expected winnings from a "multi-armed bandit," and how. We will also learn how the multi-armed bandit problem relates to design of experiments. Necessary Background: STAT 305, 323, 324

3. Research questions of interest based on survey data related to health studies. You will be required to, among other things:
   - Conduct a literature review of existing work related to your questions of interest.
   - Analyze the data to answer questions of interest.

   Recommended Background: STAT 323, 324; a thorough knowledge of regression analysis will prove to be useful.

Supervisor: Anelise Sabbag

1. Opportunity to get experience teaching some introductory statistics topics to non-majors and become familiar with the statistics education research literature.
   - Choose 2-3 statistical topics to focus on and explore what statistics education research has been done on these topics.
   - Develop a lesson plan based on recommendations from statistics educators and scholars.
   - Supervised teaching of these topics (STAT 130 or STAT 217)
   - Develop and administer assessment to students.
   - Evaluate students' performance on topics taught.
   - Recommendations of changes in lesson plan and assessment.

   Necessary Background: Completion of STAT 410 preferred but not required.

Supervisor: Andrew Schaffner

1. **Mother-child interactions in breastfeeding and infant growth trajectories.** Working with Dr. Alison Ventura (KINE) we will test a conceptual model that depicts feeding interactions as dyadic processes wherein risk for rapid weight gain is predicted by the interaction between infant communication of satiation and maternal responsiveness to infant cues, as well as the dyad's ability to develop patterns of interaction that become more sensitive and effective over time. The data is from coded videos of a longitudinal study of 325 mother-infant dyads assessed when infants were fed via bottle or solid foods

at 2 weeks and 2, 4, 6, 9, and 12 months. We will first attempt to identify clusters of dyad behavior, and then examine whether those clusters are predictive of rapid weight gain between birth and 12 months. Further, we will attempt to identify ay maternal or infant correlates of cluster membership.

Methods: We will use hidden markov models to identify latent states to characterize the dyad behaviors and probabilities of changing states during the course of feedings. A large (if not the entire part of this senior project) will be an independent study of the use and application of hidden markov models.
Necessary Background: Helpful (STAT elective) courses: 419

2. **Agricultural data analysis.** Working with Dr. Lauren Garner (AEPS) we will examine how different pruning and flower removal methods impact pomegranate harvests in terms of yield, chemical composition, and marketability. This project consists of several years (continuing this year) of harvest. Methods: We primarily will used mixed models to analyze these data as there is substantial nesting of random effects in the experimental design.
Necessary Background: Helpful (STAT elective) courses: 423

3. **Your interest**…I would be very excited to support your pursue in an independent topic of your choosing, Come talk.

Supervisor: Jeff Sklar

1. **Research Methods to Explore Educational Data**. Use various statistical methods (for example, linear regression, logistic regression, or discrete-time survival analysis methods) to examine various educational outcome variables such as persistence, drop-out, GPA, or graduation for Cal Poly students.
Recommended Background: STAT 324

2. **NFL Analytics** (with Dr. Andy Guyader from ARCE). Football games produce large amounts of data. Using NFL drive-level data from 2000-2016, a metric to evaluate the performance of each individual drive in a game has been developed taking into account starting field position and drive result (e.g. punt, field goal, touchdown), ultimately resulting in the "expected points added (EPA)" per drive. Variants of the EPA can also be computed to produce a team's Average Halftime EPA, Average Game EPA, or Season EPA. The current project may involve examining the ability of the EPA to predict game outcomes and/or creating predictive models that include EPA and additional explanatory variables.
Recommended Background: STAT 324

Supervisor: Heather Smith

1. **Survey Research in Industry.** I work with a local survey research firm, Opinion Studies, providing statistical consultation on a variety of survey research projects. Opinion Studies works primarily in the fields of litigation and consumer research. Your role on this senior project would involve working collaboratively with me, the principal of Opinion Studies, and with selected clients to provide support for survey research efforts. This support could span the full range of survey research tasks: questionnaire design, training of interviewers, sample selection, data management and cleaning, data summary and analysis, report writing, and meeting with clients. This senior project would start in the fall term.
   Necessary background:
   - An interest in working collaboratively with a senior survey researcher
   - An interest in working with several survey research clients
   - Strong skills in:
     - Oral communication,
     - Written communication, especially the creation of graphs, tables, and reports,
     - EXCEL, JMP and/or Minitab, SAS and/or R,
     - Data management and statistical analysis in the programs listed above, and
     - Selection and implementation of statistical methods, especially in a survey research setting.
   - Satisfactory completion of STAT 421 is preferred.

2. **Survey Research at Cal Poly.** At Cal Poly there are always staff members who are interested in obtaining assistance with the many aspects of survey design and analysis. This senior project would be similar to Project #1, except it would involve helping a Cal Poly client instead of an industry client.
   Necessary background: Same as Project #1.

Supervisor: Dennis Sun

1. The Evolution of Chant. Gregorian chant is the oldest tradition in Western music. Musical notation had not been invented, so chant was transmitted orally from one monk to another. You can imagine that each time a chant was transmitted, the learner might change the chant slightly--either because they misheard or misremembered it (like in a game of telephone) or because they thought it might sound better a different way. As a result, by the time a chant reached a far-off place like England, it sounded quite different from the original chant from Italy.

   In this project, you will use the differences between these different versions to trace the spread of Gregorian chant throughout Europe, much like how biologists use differences in DNA to create evolutionary trees. This project will require some amount of data cleaning and R programming and will potentially lead to a publication.
   Necessary Background: STAT 331 and the ability to read music is required.

2. Karl Pearson and the Degrees of Freedom Controversy. In 1900, Karl Pearson introduced the chi-squared test. However, he got the formula for the degrees of freedom wrong. The first person to suspect that something was off was Udny Yule (of "Yule-Walker" fame), who built a mechanical contraption to carry out hundreds of simulations. Finally, in 1922, R. A. Fisher came up with a proof that Pearson was wrong, although Pearson never accepted his error. In this project, you will investigate the history of this controversy and develop a lesson plan that uses this history to enliven chi-square tests and simulations.
   Necessary Background: STAT 427 is required. A strong interest in history and writing is recommended.

3. Resampling Methods in Python. Resampling methods, like the bootstrap and permutation tests, are not straightforward to implement in Python. In this project, you will learn about various resampling methods (including the bootstrap, jackknife, subsampling, and permutation tests) and develop a Python library that makes implementing these methods simple for practitioners. Such a package would be a major contribution to the statistics ecosystem in Python.
   Necessary Background: DATA 301, CSC 102, and STAT 331 are required. STAT 427 is recommended.

4. Topic Modeling and Authorship Analysis. Topic modeling allows us to automatically find "topics" in large corpuses of text documents and determine which topics are represented in any given document. You will learn a few common topic modeling methods (such as matrix factorization) and apply them to analyze documents whose authorship is unknown or disputed, like the Federalist Papers. This project will potentially lead to a publication.
   Necessary Background: DATA 401 is required.

5. Most Powerful Nonrandomized Tests. The Neyman-Pearson lemma says that the likelihood ratio test is the most powerful test of simple hypotheses. However, when the probability distributions are discrete, it is usually impossible to construct tests that are exactly level alpha, unless the test is allowed to be randomized (i.e., for some observed values, we flip a coin to decide whether or not to reject the null hypothesis). Randomized tests are considered by most practitioners to be unacceptable.

   In this project, you will use ideas from computer science (specifically, algorithms) to design optimal nonrandomized tests that are more powerful than the likelihood ratio test for discrete probability distributions. This project will potentially lead to a publication.
   Necessary Background: STAT 427 and CSC 349 are required.

6. Finite-Sample Properties of Permutation Tests. In classes that teach simulation-based inference, the two-sample t-test is presented as an approximation to the permutation test. But is the t-distribution really a more accurate approximation to the permutation distribution than the normal distribution? And should we pool or unpool the variances? Also, how does the power of the permutation test compare with its parametric counterparts? In this project, you will conduct a large-scale simulation study to make recommendations.
   Necessary Background: 331 is required. STAT 427 is recommended.

7. Sample Splitting for Estimating Standard Errors. The bootstrap and jackknife are two well-known methods for estimating the standard error of a statistic. In this project, you will investigate another approach based on sample splitting, by conducting a large-scale simulation study and perhaps also doing some theoretical calculations. This project will potentially lead to a publication.
   Necessary Background: STAT 331 and STAT 427 are required.

8. Two-Sided Fisher's Exact Test and Confidence Intervals for the Odds Ratio. Fisher's exact test is used to test the hypothesis that two proportions are equal. Unlike the two-sample z-test for proportions or the chi-square test, it does not require large sample sizes and is valid even for small samples. However, there is some controversy about how to define a two-sided Fisher's exact test. The two-sided Fisher's exact test is important because it is the test that is inverted to obtain confidence intervals for the odds ratio. In this project, you will read about the different approaches, as well as investigate an alternative approach that I have in mind. You will implement them and compare their properties (e.g., Type I error and power). This project will likely lead to a publication.
   Necessary Background: STAT 331 and STAT 427 are required.

Supervisor: John Walker

1. **SHINY Applets for Regression and/or ANOVA Concepts.** I have a number of ideas for applets to demonstrate advanced ANOVA and regression concepts. I'd like to work with a student to develop some SHINY applets in R to demonstrate these concepts. Examples: How is multicollinearity affected by different levels of correlation between predictors? How do differing levels of dependence in the regression errors affect the results of a regression and how effective are different tests at detecting dependent errors? The applets would be used by students in future statistics classes.
   Necessary Background: STAT 323, STAT 324, STAT 331

2. **Power Analysis for Advanced Experimental Designs.** Most statistics software programs have power analysis commands for simple experimental designs; however, often consulting client want to know the power for more advanced designs. Both SAS and R have some commands for power analysis in complicated designs. The first part of this project is to identify which programs support power for which designs. Once we know what SAS and R cannot do, we can develop software tools (most likely in R) to cover a few designs that are not covered by SAS or R. Ideally, we would write functions to compute the power exactly and check those calculations with simulations. In cases where the exact power calculation is too difficult, we can rely on simulation alone for the answer. The student will need good knowledge of some advanced designs from STAT 423 (split plot, repeated measures, etc.). Knowledge of SAS and R is also important to do the simulations, although the student doesn't need to be equally good at both programs. For ease of use, these results could be displayed in a SHINY applet.
   Necessary Background: STAT 330, STAT 331. STAT 423

Supervisor: Consultants Vary by Quarter

Supervisors:
Fall: Jeff Sklar
Winter: Heather Smith and TBD
Spring: TBD

Project Description: Consulting Intern

Each quarter one (or two) faculty members serve as the statistics department consultants. The consultant helps students and faculty across the campus with the statistical aspects of their research. Some examples of the type of work the consultant does include helping design surveys, sampling and experimental designs, sample size calculations, and data analysis. As a senior project for a statistics major, you would shadow each of the consultants over the year, attending meetings with clients, conducting analyses, and writing summary reports for the client and for a personal journal.

Necessary background**:**

- An interest in working collaboratively with other researchers
- Strong skills in:
  - Communication, both oral and written
  - Applied statistics
  - JMP, Minitab, SAS and/or R.