STA 712 Semester Project

Overview: STA 711/712 provide an overview of many core concepts in statistics. Inevitably, however, there will be more interesting and useful topics than we can fit in the two courses. The good news is that after taking 711 and 712, you should have the background to learn many of these topics on your own!

The purpose of this project is for you to learn a new statistical tool, method, or concept, to apply it to real data, and then to teach it to your classmates.

- You will work in small groups (2–3 members) for this project. Working in groups will allow you to dive deeper into the material, and collaboratively learning a new skill is valuable for both academic and industry careers.
- Learning a new topic will also give you good practice reading and searching the statistical literature. With new statistical research published every day, knowing how to navigate the literature is crucial for keeping up-to-date in the field.
- Teaching a new concept to your peers is both valuable for learning (you can't teach something well unless you *really* understand it), and is also important practice with statistical communication. Academic jobs involve regular conference presentations; industry jobs involve presentations to clients and colleagues, and may include conferences too.

This project will be completed over the full semester. Throughout the semester, there will be smaller check-in assignments to help guide your work. At the end of the semester, your group will submit a written report that summarizes your topic/tool of choice and applies it to real data. Your group will also teach a 50-minute lesson during one of the final class periods.

Timeline and tentative due dates

Here is a rough timeline for the project (exact dates are tentative and subjective to change). Full requirements for each component are described below.

- Week 1: Form groups and choose a topic (**due:** Monday, September 4)
- Week 4: Meet with Dr. Evans to discuss the scope of your project and what you plan to cover (**due:** your meeting with Dr. Evans must take place by Friday, September 22)
- Week 8: Finalized topic outline, literature summary, and applied data example. Time check-in. (**due:** Friday, October 20)
- Week 12: Meet with Dr. Evans to discuss your lesson plan. (**due:** your meeting with Dr. Evans must take place by Friday, November 17)
- Week 14: Written report and full lesson plan (slides, activities, etc.). (**due**: Friday, Dec. 1)
- Week 14/15: Presentations!

Forming groups

You will work in a group of (approximately) three students. You may choose your own groups, or I would be happy to assign you to a group. The best groups will have a team with similar interest in a topic, but differing abilities and interests in terms of statistical theory and applications (e.g., a group member who is excited about proving theoretical properties, a group member who is excited to dig into the background literature, a group member who is interested in identifying interesting applications, etc.).

Project roles

Here are some suggested project roles for group members. In each case, the person assuming that role is responsible for that aspect of the project. It doesn't mean that they will do all that part of the project by themselves; it means that they are responsible for dividing that work up among the members of the group and ensuring that it is done and recorded correctly.

- **Research director:** Responsible for defining the scope of the project, and the literature searches. Identifies what needs to be searched for in the literature, divvies up the literature searches to be performed among the group members, and coordinates changes in the searches based on information gathered and changes in direction. They are also responsible for making sure that the citations in the project are complete and accurate.
- **Applied data leader:** Responsible for identifying appropriate motivation for the chosen method/tool/concept (i.e., why would a statistician use this in real practice?), and finding a real dataset to which the method can be applied. Coordinates analysis of the data with the chosen method/tool/concept; they are responsible for the actual implementation.
- **Communication director:** Responsible for the final written report and the 50-minute lesson. This involves coordinating contributions to the report, developing a lesson plan, and creating the materials needed for the lesson.

Grading with groups

- (Generally) All members of the group get the same grade for the project.
- If all members of a group think that one member isn't pulling their weight then you can come and talk to me about it. I can fire that member and have them do the project on their own.
- Students should keep a record of all the times that they worked on the project and the work that they did. Every time that you spend more than 15 minutes on the project you should write down the start and stop time and what they did. Group members should be spending roughly the same amount of time on the project (I will ask you to report on time spent on the project so far at the Week 8 check-in).
- Attendance: you should keep track of who is or isn't showing up to group project meetings.

Assignment

Week 1: groups and topic

Form groups of 2–3 students. Each group emails Dr. Evans the following information:

- Group members and their tentative roles
- Proposed topic: a short (one-paragraph) summary of the topic you plan to investigate, with a couple references. You should explain why you want to explore this topic: why would a statistician be interested in learning this new concept?

Your proposed topic may change, with agreement of all group members, until the Week 4 check-in. If your group does change the topic before Week 4, all group members must email Dr. Evans with notice of the change, describing the new project topic.

Week 4: initial scope check-in

Each group meets with Dr. Evans to discuss the scope of the project, including:

- Topic: A description of the topic you will be exploring. You should be more detailed than the Week 1 email. E.g., if your project is on permutation tests, which specific permutation tests will you cover? Will you contrast permutation tests with other parametric and nonparametric tests? etc. What are the main theoretical results you will cover?
- Motivation: You should make an argument motivating your work. Why should someone be interested in what you are doing? What do you hope people will learn from your project?
- Data: As best you can, describe where you will find your data, and what kind of data it is. Be as specific as you can, listing URLs and references. **Note:** You may *not* copy an existing data analysis from the literature. You are responsible for finding new data for your analysis.
- References: A preliminary list of references you will use. You should have read, and be prepared to discuss, each reference in the list.

All group members must be present at the meeting.

Week 8: finalized proposal

Each group will submit a document finalizing and expanding their project proposal from Week 4. The document should have the following sections and content:

- **Motivation and background:** This section should be approximately 2–3 pages (single-spaced).
  - Provide a brief overview of the project topic.
  - Explain why the topic of your project is important, and summarize (with citations) how it has been used in real research and applications.
  - Summarize relevant prior literature.
- **Content outline:** Provide a bullet-point outline of the content you will cover, and any key results. For example, suppose your project topic was logistic regression. Your outline could include:

  - Definition of the logistic regression model, and connection to exponential dispersion models
  - Fitting the model (Fisher scoring)
  - Inference with logistic regression
    * Key results: consistency and asymptotic normality
    * Wald tests and LRTs
  - Model diagnostics
    * Empirical logit plots
    * Quantile residual plots

  Be as detailed as possible.

- **Simulations plan:** The simulations section investigates properties of your project topic with simulated data. For example, in a logistic regression project, you might use simulations to assess:

  - Sample size required for the distribution of $\widehat{\boldsymbol{\beta}}$ to appear approximately normal
  - Comparison of estimates between logistic and linear regression (to motivate using logistic regression for binary outcomes)
  - Robustness of logistic regression to violations of the assumptions

  Your simulations section should address at least one question about your proposed topic (e.g., robustness to violations of the assumptions), and should simulate data in at least three different scenarios. For the Week 8 project proposal, describe how you might use simulations to explore your topic; contact Dr. Evans ahead of time if you are unsure.

- **Case study plan:** This section should be approximately 1–2 pages (single-spaced), and should describe the data you will use to demonstrate your topic, and the analysis you will perform. Your case study plan must address the following:

  - What is the source of the data, and how were they collected?
  - What is the size of the data, and what does a row in the data represent?
  - What information (e.g., variables) is available in the data?
  - What is the overall research question you wish to address, and how will you address it using your project topic?
  - What software will you use?

- **References:** A bibliography for all the references cited in your proposal document. There is no specific number of references required, but I have found that a good literature summary and background (including motivation) usually includes around 15 or more references.

Time check-in: When you submit your finalized proposal, each group member should also report how much time they have spent on the project so far. Remember, group members should be spending roughly the same amount of time on the project.

Week 12: discuss lesson plan

Each group meets with Dr. Evans to discuss their plan for a 50-minute lesson on their project. This lesson should be taught at the level of STA 712. After the lesson, students should be able to answer:

- What is your project about?
- Why do statisticians care about the project topic, and how is it used in practice?
- What are the key results for your topic?

You are welcome to use whichever teaching style you find most effective, and you can choose how you want to organize the lesson. The Eberly Center at Carnegie Mellon University has some good advice on lesson planning:

`https://www.cmu.edu/teaching/designteach/teach/instructionalstrategies/lectures.html`

During this meeting, Dr. Evans will give suggestions on ways to improve the lesson. All group members must be present at the meeting.

Week 14: full written report

Your full written report should be organized similar to an academic paper, including the following sections:

- **Introduction:** The introduction provides a brief overview of the topic (without going into technical details), and gives motivation. Summarize why the topic is important. Introductions are generally 1–2 pages (single spaced).
- **Methods:** Describe, in detail, the topic you have chosen. This section should roughly follow your content outline from the Week 8 proposal, and should provide (not necessarily in this order):
  - Formal definitions for your topic
  - Any assumptions required
  - Any key theoretical results (e.g., asymptotic normality for logistic regression)
  - Details on implementation, diagnostics, etc. (when relevant)
  - A literature summary of important research on your topic
- **Simulations:** Implement the simulations described in your Week 8 project proposal. Remember that your simulations should address at least one question about your proposed topic (e.g., robustness to violations of the assumptions), and should simulate data in at least three different scenarios. This section should:
  - Describe in detail how the data were simulated (there needs to be sufficient detail that a reader could reproduce your results without the code), including the number of simulations, the generating distributions, sample sizes, etc.
  - Describe the outcome being assessed (e.g., the type I error rate, coverage of confidence intervals, the distribution of $\widehat{\boldsymbol{\beta}}$, etc.)
  - Summarize your results with appropriate figures and/or tables

  - Discuss and interpret the simulation results: what did you learn from the simulations?

- **Case study:** This section shows how your topic can be used to address real statistical questions. You should provide the following:

  - Describe the source of the data, and how they were collected
  - Summarize the data: the size, what a row in the data represents, the variables available, any missing data, etc.
  - Present the research question you will answer, and explain why you are using your project topic to answer that research question
  - Show any exploratory data analysis (EDA) needed to address the research question
  - Use your topic to analyze the data and address the research question (e.g., fitting a model, testing a hypothesis, making predictions, etc.)
  - Summarize your results

- **Code and data availability:** No code should be included in your written report, but all the code and data needed to reproduce your analysis should be available. Create a GitHub repository to host your code and data, and link the repository in your report with the following statement:

  > All code and data needed to reproduce the simulations and analyses in this report are available at [the GitHub repository].

- **References:** A bibliography for all the references cited in your written report. All references must be properly cited; you may use any formatting for the citations as long as the citations are consistent, standard usage, and contain all relevant information for your reader to find the appropriate documentation. I recommend using BibTeX to cite sources in LaTeX documents.

Week 14: full lesson plan

Each group will send Dr. Evans their full lesson plan and materials; these will be due in advance of the actual presentations. You should submit final versions of any slides, lecture notes, and class activities which you plan to use.

Week 14/15: presentations!

During one of the last few class periods, your group will give a 50-minute presentation, following the lesson plan you discussed in Week 12 and finalized in Week 14.

Possible topics

Your project topic should be related to the general themes of STA 711 and 712 (estimation, inference, modeling, prediction, robustness, etc.), and must include new material that we have not covered in class. Here are a few suggestions to consider, with a couple references to get you started. You may need to refine the scope of these topic suggestions somewhat. You may also come up with your own topic, with Dr. Evans' approval.

- Permutation tests for regression problems
  - DiCiccio, C. J., & Romano, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association*, 112(519), 1211-1220.
  - Anderson, M. J., & Robinson, J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1), 75-88.
  - Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses.* Springer Science & Business Media.
- Nonparametric tests based on ranks, depths, or graphs
  - Liu, R. Y., & Singh, K. (1993). A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421), 252-260.
  - Bhattacharya, B. B. (2019). A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(3), 575-602.
  - Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395), 799-806.
- Conformal prediction
  - Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094-1111.
  - Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. New York: Springer.
  - Vovk, V., Nouretdinov, I., & Gammerman, A. (2009). On-line predictive linear regression. *The Annals of Statistics*, 1566-1590.
  - Shafer, G., & Vovk, V. (2008). A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(3).
- Testing random effects in mixed effects models
  - Crainiceanu, C. M., & Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1), 165-185.
  - Self, S. G., & Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398), 605-610.
- Multiple testing and false discovery rate

  - Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289-300.
  - Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3), 479-498.
  - Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16), 9440-9445.

- Penalized spline smoothing
  - Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89-121.
  - Claeskens, G., Krivobokova, T., & Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3), 529-544.
  - Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4), 735-757.
  - Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press.

Resources and advice

Finding data

There are lots of great places to find data. Jo Hardin at Pomona College has a great list here if you need help getting started.

Planning simulations

The following papers give useful advice for planning and designing simulation studies:

- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074-2102.
- Boulesteix, A. L., Groenwold, R. H., Abrahamowicz, M., Binder, H., Briel, M., Hornung, R., ... & Sauerbrei, W. (2020). Introduction to statistical simulations in health research. *BMJ Open*, 10(12), e039921.

Planning and delivering a lesson

See the following links for lesson planning guidance:

- The Eberly Center, Carnegie Mellon University
- Washington University in St. Louis, Center for Teaching and Learning
- Vanderbilt University, Center for Teaching

Advice for presentations:

- Budget your time. You have 50 minutes, but it is amazing how quickly that time goes. You will not be able to get into all the details of your project; focus on the key material which you want other students to learn about the topic.

- If you use slides, don't make the slides too cluttered. You can also consider annotating slides with a tablet, or adding supplementary details (like proofs) on the board.
- Begin with motivation: your audience should understand why they want to learn about this new material.
- Speak loudly and clearly. Remember that you know more about your topic that anyone else in the room, so speak and act with confidence!

Technically Speaking (`http://techspeaking.denison.edu/Technically_Speaking/Home.html`) also has some good advice for presentations in general.

Written reports: format and style

- Your final report should be written like an article or research paper: in full sentences and paragraphs, with headings for each section. You should not write your report with question numbers or as a list of bullet points. Scientific articles are generally written in third person, though "we" can also be acceptable ("we can see from Figure 2.1 ...") in some disciplines.
- Reports should be written using LaTeX. I recommend using Overleaf to collaborate on a shared document. There is no required template for this project.
- No code should be included in the report. Put all code needed to reproduce your simulations and data analysis in a GitHub repository, and reference this repository in your written report.
- Figures should have labeled axes, and should be clear and easy to read. Figures should also be captioned and numbered. Captions should provide enough information to understand what is being plotted, but interpretation can be left to the main text. Refer to figures by their number in the text. Make sure that any figures you include are discussed in the text.
- Tables should be nicely formatted, and have a number and caption. Refer to tables by their number in the text.

Using LaTeX

Overleaf has a set of great tutorials to help you with LaTeXissues that arise. Dr. Evans can also help if you get stuck.

Grading