Statistical Inference

Project Designers: *Ali Mikaeili, Kamand Mesbah, Reza Salamat*
Instructor: Mohammadreza A. Dehaqani

Fall 2024

Final Project

Introduction

Picture yourself at the entrance of a grand, ancient library. Towering shelves filled with leather-bound volumes stretch high above, their spines glinting in the dim lamplight. These books hold centuries of carefully preserved wisdom, yet they also contain overlooked details and fragile assumptions that can hide beneath polished prose. Research is a vast kingdom, and peer-reviewed journals are among its grandest halls. However, even studies that appear strong at first glance may contain hidden gaps in logic or untested assumptions waiting to be discovered by a perceptive eye.

In this project, you become the scholar-knight who does more than simply stroll through the halls of knowledge. Armed with your critical thinking and statistical tools, you will embark on a journey unfolding in two steps. First, you will immerse yourself in a chosen dataset, learning to navigate its variables, noticing its peculiarities, and forming a meaningful research question. Once you have honed your skills in data exploration, you will shift your focus to a published claim and examine its foundations. In doing so, you will learn the value of gathering and testing evidence with careful scrutiny. This project, including its bonus component, is worth 4 points out of 20. It's probably less time-intensive than one typical homework assignment, but it's a great chance to revisit what you've learned and show off your analytical skills.

Proposal (Due Dey 26th)

At the beginning of your journey, you will select a dataset that genuinely piques your curiosity. It may come from a familiar domain or introduce you to a new field. Your task is to probe this data with fresh eyes, study its structure, identify potential outliers, and address any missing values hindering clarity. As you delve deeper, you will understand the patterns and relationships waiting to be uncovered.

You will gradually refine a specific question about the dataset through your explorations. Perhaps you will discover an unusual correlation between variables, an intriguing gap in the data, or an unexpected trend that calls for further explanation. By the time you have made these initial discoveries, you should be able to articulate a few research questions that capture the core of your curiosity.

Once your research question takes shape, you will submit a proposal outlining how to investigate it. This proposal is worth 10 percent of your total project score, making it a critical step. Your plan should describe the nature of your chosen dataset (obviously including an active link to the dataset) and any early observations or peculiarities you have noticed. It should also explain which statistical methods you believe will be most helpful in your exploration. For example, you may propose to test whether a particular variable follows a normal distribution or to compare two groups using a parametric or non-parametric method. You might even hint at a regression analysis if you suspect specific predictors influence an outcome variable meaningfully. In setting forth these ideas, you chart a course for your analysis and show why your question matters.

To strengthen your proposal, you can choose a paper that uses this dataset. This paper can help you implement the bonus part of the project, providing additional insights and direction. The bonus part is an optional component that can enhance your overall score. Remember to include two extra paragraphs about the paper if you choose to attempt the bonus part. However, choosing a paper or attempting the bonus part

does not guarantee approval of your proposal. It may still be rejected if the proposal lacks clarity, coherence, or depth.

Approval of this proposal ensures that your exploration has direction and substance before you move on. Much like a traveler consulting a map before venturing into unknown lands, you will benefit from clarifying the steps you intend to take. Once your proposal is confirmed, you will be ready to bring your refined analytical skills to the project's next phase, where you will confront a published study's claim and test its strength.

Important: you should formulate three distinct questions related to your dataset. Provide at least two preliminary analyses for each question to explore potential insights. Please, aim to keep your analysis concise and focused, ensuring the entire proposal (excluding the bonus part) fits within two pages.

Exploring Your Dataset

After refining your analytical approach, you will focus on a specific claim you wish to explore as part of your proposal. This claim will form the backbone of your investigation, serving as the focal point of your statistical inquiry. You will approach it like an investigator questioning a witness, seeking either confirmation or contradiction. Perhaps your initial data exploration reveals assumptions about data distributions that deserve further scrutiny, or your observations suggest that the claim's conclusion extends beyond what the evidence truly supports.

Your proposal outlined how you plan to re-examine this claim using the dataset you gathered or prepared. Your analysis will incorporate the methods detailed in the TODO section, allowing you to draw upon various statistical techniques to investigate the claim. If your findings affirm your proposed claim, you will reinforce its validity. If they do not, you will identify weaknesses or gaps in the claim and suggest thoughtful directions for further exploration.

Recognizing that your proposal is a crucial checkpoint in this process is essential. Approval of your proposal ensures that your approach to statistical inquiry is grounded in rigorous and well-prepared work. This structured process allows you to proceed with the confidence and precision of a well-prepared scholar, ensuring that your investigation is thorough and meaningful. Remember that this part is worth 45% of your total score.

Project TODO

You can choose up to four items from the visualizations section, four from each of estimation and testing sections, and up to three from the regression section, giving you up to sixteen tasks. However, don't worry about possible errors and potholes you might fall into. If you complete and get twelve of them accepted, you will receive full credit for this section. This allows you to focus on the tasks that interest you most while ensuring you cover a good range of what you have learned and also be forced to go beyond that.

The quality of your work is just as important as the quantity. You can't conduct meaningless analyses and expect to receive a good score. Your approach and execution should demonstrate your understanding of the concepts and their relevance to your project.

(1) Visualization and Summarization of Dataset Variables

- Perform exploratory data analysis by visualizing categorical, numerical, and mixed variables. Use appropriate plots like bar charts for categorical data and histograms for numerical data.
- Analyze the distribution and relationship of variables. Create scatter plots, box plots, and QQ-plots to understand variable dependencies.
- Create pair plots to visualize relationships between multiple numerical variables simultaneously.
- Identify and report any dependent or independent factors discovered during analysis.
- Generate heatmaps to visualize the correlation matrix of numerical variables.

- Explore and visualize metrics that capture relationships beyond linear correlation, such as Spearman's rank correlation, Kendall's Tau, or mutual information.
- If the dataset comes from a published source, include and discuss any visualizations provided in the publication.

(2) Parametric Inference and Estimation

- Utilize interpolation methods to predict missing or unobserved data points, if necessary.
- Conduct parametric inference methods where applicable, explaining the choice of parametric tests.
- Apply estimation techniques to various variables, such as point estimation or maximum likelihood estimation.
- Conduct goodness-of-fit tests to determine how well the data fits a specified distribution.
- Calculate and interpret confidence intervals for the estimated parameters. (It might be a good idea to create plots to visually represent confidence intervals)

(3) Hypothesis Testing and Statistical Analysis

- Formulate and test hypotheses based on the dependency of variables. Clearly state the hypothesis, the rationale for testing, and the method used.
- Perform power analysis to determine the sample size required for detecting effects with desired power.
- Report all statistical findings, including test statistics, p-values, and conclusions drawn.
- Employ bootstrap or resampling methods for hypothesis testing, especially in cases of non-normal data distributions.
- Apply non-parametric tests such as the Mann-Whitney U test or the Kruskal-Wallis test for data that do not meet parametric assumptions.
- Conduct Analysis of Variance (ANOVA) tests to compare means of different groups. Use correction methods like Bonferroni or Tukey's HSD for multiple comparisons, if applicable.
- If the dataset comes from a published source, Critically analyze and discuss any tests presented in associated publications and compare them with your findings.

(4) Regression Analysis and Reporting

- Perform regression analysis to investigate relationships between variables. Explain the choice of regression model and interpret the results.
- Apply regularization techniques such as Lasso or Ridge regression to handle multicollinearity and enhance model performance.
- Perform diagnostic tests on regression models to check for assumptions like homoscedasticity, multicollinearity, and independence of errors.
- Include a detailed analysis of regression findings from the publication, if available.

  Note: If you haven't studied regression, you can instead choose 4 Visualization (Section **1**) tasks and 6 tasks from Estimation and Testing (Sections **2** and **3**)

Bonus Part

You will apply your full statistical skills once you have secured approval and chosen a published claim to explore. Hypothesis testing, confidence intervals, parametric or non-parametric tests, and regression analyses are all fair game. Visualizations such as scatter plots, box plots, or correlation maps can make your findings more transparent, helping you and your audience see the underlying patterns in the data. You may find it helpful to test alternative models, run diagnostics, or consider bootstrapping to confirm whether the original claim remains under varied assumptions. (you can use TODO part to find more information)

If you chose a paper and wanted to implement the bonus part, this section is for you. Remember that this part is worth 45% of your total score, making it a substantial component of your project. It is unnecessary to identify something explicitly wrong in the paper; you can instead choose to explore aspects or analyses not mentioned in the paper. It is recommended you select a Q1 or Q2 paper, as the quality of your source is as important as the quality of your work. Your conclusions and conducted analyses must be valid and meaningful, avoiding any superficial or nonsensical interpretations.

You will weave your observations into a cohesive discussion when you conclude your investigation. Begin by recalling what inspired you to question the claim first, then walk your reader through the methods and tests you employed. Share whether your evidence supported or contradicted the authors' conclusions, being clear about any limitations or nuances you discovered. If the published claim is included in your analysis, you will enrich it by demonstrating its robustness. If it does not hold, you will have unearthed a path for further inquiry, allowing others to explore the new insights you have revealed.

Final Notes

- Students must submit a document (draft) detailing their project work, methodologies, and results. In academic terms, a 'draft' refers to a preliminary version of your work. It should comprehensively outline your project's progress and findings, though it may not be in its final, fully polished state. This draft serves as an essential step in developing and refining academic reports.
- Students must upload three essential components for their project submission: the well-documented project code, a comprehensive draft detailing their methodologies and preliminary results, and a concise, straight-to-the-point, and academic PowerPoint presentation summarizing their key findings. For more details on the formatting of these submissions, refer to your course guidelines.
- This project will have an in-person (or online) hand-in in which you will be asked questions about what you turned in to assess how fluent you are in your own work. This is where you will present the presentation you made.
- Final Word: Besides adhering to the instructions in the four outlined sections, students must engage in further research and decision-making as data scientists. Should certain aspects within these sections not be fully explained, you are encouraged to seek the necessary information and decide on appropriate approaches independently. While support will be available as it's been so far in the course, your initiative in exploring beyond the provided guidelines and justifying your methodologies is essential and highly valued in this project. Think of this as your first work beyond "Courses and Grades." Do not worry about grades, and focus on doing work you are interested in. But remember, no effort shall go unrewarded (or ungraded!)

May your curiosity and efforts lead to great discoveries!
Your Teaching Team