

Final Project: Mathematical Statistics (MATH-UA 234)

We will be doing a final project instead of a final exam. The goal of this project is to gain a better understanding of the concepts we learned in this class, as well as to practice communicating about math and statistics. You can work individually or in groups of up to three people total. If you have multiple people, your final result should be more comprehensive than if you're working alone.

There are three main components that will be evaluated:

- | | |
|--|-------|
| (i) project proposal | (10%) |
| (ii) project presentation / review of your peers' projects | (40%) |
| (iii) written final report | (50%) |

Each aspect will be given a rating “does not meet expectations”, “satisfactory”, “good”, and “excellent”. These will roughly correspond to letter grades of F, C, B, and A respectively. Based on your project proposal, I will provide you an estimated rating for your final report, assuming that you complete everything outlined in the proposal.

There are several possibilities for the contents of the project. You should pick *one* of the following:

- Find a real-world dataset. Analyze the data, and find something to say about it. Specially, use two ideas from class as well as one new idea which you've researched.
- Pick a standard topic in statistics that we did not cover in much detail, find one or more textbooks or other resources which cover this topic, and do a self-guided study of the topic.
- Find a paper on a research topic in statistics. Read (part of) the paper and try to implement some of the ideas from the paper.
- Something else you are interested in (please talk to me individually prior to submitting the project proposal if you would like to do something besides what is suggested above)

Proposal (due 11/22)

The first part of the final project is a project proposal, and each group should turn in a single proposal.

The proposal serves several purposes. First, it forces you to start thinking about the project earlier than the last minute. Second, it gives me a chance to make sure your project is on the right track to satisfy the requirements. Finally, it will give you an outline to provide structure for the rest of the project.

Since the exact format for the project is very flexible, the project proposals are very important in making sure that your proposed project will be sufficient. Submit to Gradescope 1-2 pages outlining your plan for the project. Make sure to answer all of the following:

- Who is in your group?
- What is the specific topic / what will you do for your project? Try to be specific as possible. It's not expected that you will know everything (because often what you do is conditional on what happens while you are working on the project), but the more details you can provide, the easier it will be for me to give you feedback on if your project is sufficient.
- Describe what you will put in the final report. Again, try to be as detailed as possible. Minimum suggested requirements are listed below to use as a starting point. However, your proposal should contain a more detailed outline of your plan.

Presentations (12/08, 12/14 in class)

Full details will be provided closer to the end of the semester.

Each group will prepare a poster on their project. The goal of your poster is to help explain what you did for your project to your classmates. Making the poster will also force you to think about how to simplify complex mathematical ideas into something which is understandable to people with less background in the area.

I will split the groups into two or more sections. One section will present while the other goes around to see other groups' posters. You will write a brief review on several of the projects which you visited.

Written report (due 12/21)

In addition to a poster, each group will produce a written report summarizing what you did during your project. The exact requirements for the report are flexible since there are many possible projects. Thus, you should lay out a precise plan for what you will put in the written report in your project proposal. Overall, the project is expected to take the time of several homeworks as it carries 30% of the final grade.

I have listed minimum expectations which you can use as a starting point for your proposal. However, your proposal should contain a detailed description of your plan for the project so I can determine whether it is sufficient.

- If you analyze a real-world dataset, I expect a roughly 5 page summary of what you did in addition to your actual analysis (e.g. code) of the dataset. You should clearly explain what the dataset contains, what analysis you did and why it's appropriate, what you conclusions you were able to make, and potential directions for future work.

- If you choose to study a “textbook” topic, I expect a roughly 10 page summary of what you learned; i.e. explaining the topic and how it fits into what we learned in the class in your own words. Some of this may be your solutions to the textbook problems. You must demonstrate that you have understood the topic well.
- If you read a research paper, I expect a 5-10 page summary of the topic. The length depends on whether the paper is mostly theoretical, or whether you are able to implement the ideas yourself (e.g. code them up). You should clearly explain the context for the paper (why the paper is important), what problem the paper aims to solve, and what the techniques in the paper are.

Project Ideas

Below are a collection of resources you can use to help pick a project.

Datasets


- New York City public data
- CDC COVID-19 data
- Links to several free datasets

“Textbook” topics

- Concentration inequalities
- Kernel density estimation
- Variants of bootstrapping
- ANOVA
- Proof of CLT by method of moments
- Generalized method of moments
- Proof of properties of MLE (consistency, optimality, etc.)

Research topics

- Reproducibility in Learning
 - Reproducibility is a cornerstone of science, but many studies and experiments fail to be reproducible. This paper introduces the idea of reproducible algorithms, and describes how they can be implemented for many problems in learning theory.

- A reproducible learning algorithm is resilient to variations in its samples — with high probability, it returns the exact same output when run on two samples from the same underlying distribution.
- The paper covers a lot of different classes of algorithms, but the simplest example is statistical queries (section 2) which covers a number of the basic parameter estimators we've seen in the class (sample mean, sample variance, etc.)
- COVID related stuff
 - Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21
 - The Vaccine-Hesitant Moment
- Algorithmic bias
 - For the most part, in this class we consider the mathematical aspects of statistics. This allowed us to avoid many of the intricacies of statistics in the real world by making assumptions about our data.
One big issues in learning is that bias in the data results in algorithms and models which reflect the biases in the data, *even if the algorithms themselves are not actively/intentionally biased*. That is, even statistical tools which are mathematically correct when the data satisfies certain assumptions may give in undesirable outputs on real data. This article provides an overview of the topic.
 - On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 
 - * The paper that got Timnit Gebru fired from Google
 - Propagation of societal gender inequality by internet search algorithms
 - * Paper by folks at NYU
- Algorithms for computing sample variances, covariances, and higher moments
 - While it is straightforward to compute sample variances in theory, when algorithms are implemented in finite precision arithmetic, rounding errors can result in inexact computations. Because the sample variance involves the sum of squares of numbers, this can be particularly susceptible to numerical issues.
 - The Wikipedia page is a good starting point with several references