Special Topics Comp Stat & Pro MAT5999 and Computational Stats & Prob. AIM 5002 Spring 2022 Final projects

You have two options to choose from: you can either take a final exam OR work on a final project. You will only get credit for one of these two options. This document is about the final projects. We will discuss the specifics of the final exam near the end of the semester.

General rules about the final project. If you work on a final project, you will have to write and submit a report. The report will be shared with the class because all topics are of interest to everyone. The length of the report is a poor indicator of the quality of the work but still to have an idea, I would suggest to write a report of approximately 5 to 10 pages length. The report should be typed in and not hand written. There is no requirement on the text editor that you use but using RMarkdown file and producing a pdf for a final report is encouraged (and is required if a substantial part of the project involves R). Latex is preferred to Microsoft Word but Microsoft Word is also accepted, especially if you have not used Latex before.

Everyone should work on a different project. If you want to choose a topic from the list below, send me an email and I'll mark it as "taken" on the website (except for project 1 which can be chosen by any number of students as long as everyone works on a different dataset). You can only choose topics that are not marked as "taken".

I would estimate that you want to work on a project for about 3 weeks but again, this highly depends on the intensity of the work. The deadline for choosing a project is the last day before Passover break (April 13th) and the deadline for submission of the report is the last day of classes (May 13th). I am happy to answer any questions that you may have.

Project topics

1. **A statistical study of your choice** If you are interested in application of the statistical methods that we studied in class, you can go ahead and find some data set yourself on which you perform statistical analysis. The analysis should be reproducible and the report should be written in an R Markdown file. Some good ideas are listed on this site: https://spring2022.data606.net/assignments/project/. I do not insist on all formal requirements listed on this link but they are a good guiding principle and can help you writing a nice report.

2. **Glivenko Cantelli theorem** The Glivenko Cantelli theorem is a law of large numbers for empirical distributions. Present the theorem and its complete proof. It is important that you understand the proof and not just copy paste a proof from some textbook. Exemplify the Glivenko Cantelli theorem by Monte Carlo simulations in R.

3. **Kolmogorov Smirnov theorem and test** The Kolmogorov-Smornov theorem is a central limit theorem for empirical distributions. Thus it is a refinement of the Glivenko Cantelli theorem as above. Since this theorem is deeper and the proof is much harder, in this project you do not need to present the proof. Instead present the theorem itself and apply it to some data sets. For example: pick an

actual random sample from https://www.openintro.org/data/ and test whether the dataset comes from a normal distribution using the Kolmogorov-Smirnov test.

4. **Cramer-Rao inequality and Fisher information** The Cramer-Rao inequality provides a lower bound on the variance of unbiased estimators. If it is attained by an estimator, then that estimator is guaranteed to be the MVUE. Present the Cramer-Rao inequality and its proof and work out two examples (such as Bernoulli, Poisson, exponential, etc.) A reference is https://www.randomservices.org/random/point/Unbiased.html **Taken by Liteshwar.**

5. **Multivariate normal distribution and chi-squared tests** Discuss basic properties of multivariate normal distributions using linear algebra. Then discuss the multivariate central limit theorem. Prove that the chi-squared test statistic has asymptotic chi-squared distribution if the null hypothesis holds. You can use this set of notes: https://math.bme.hu/ marib/tobbvalt/tv2.pdf Don't just copy paste it. E.g. you can skip the proof of Proposition 1 but may want to add a proof for Propositions 2, 5. **Taken by Yonah.**

6. **Analysis of variances** Analysis of variances is a standard statistical topics that we do not cover in class for lack of time. Summarize the theory presented in Sections 13.1 - 13.4 of our textbook #1 (Wackerley, Mendenhall, Schaffer) and solve at least 2 even numbered exercises from Section 13.4. **Taken by Dovi.**

7. **Introduction to the theory of machine learning**. This topic is related to linear regression that we will cover in early April. Machine learning is a vast generalization and deserves to be studied in a separate semester. Nonetheless, you can get some basic ideas by reading Chapters 1 (A gentle start) and 2 (A formal learning model) from "Understanding Machine Learning: From Theory to Algorithms" by Shalev-Shwartz and Ben-David Write a report on what you found the most important in these two chapters. Solve at least 2 exercises from these chapters. **Taken by Alejandro.**