THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

Statistics Colloquium

Yiqun Chen

Stanford Data Science
Stanford University

"Advancing Biomedical Data Science: Testing data-driven hypotheses post-clustering & Leveraging ChatGPT for genomics data."

Monday February 5, 2024, at 11:30 AM
Jones 303, 5747 S. Ellis Avenue
Pre-Seminar refreshments will be served at 11:00 AM in Jones 303

Abstract

My research centers around bringing statistical insights and understanding to the practice of modern data science, and I will cover two projects related to this research vision in this talk. The first part of the talk is motivated by the practice of testing data-driven hypotheses. In the biomedical sciences, it has become increasingly common to collect massive datasets without a pre-specified research question. In this setting, a data analyst might use the data both to generate a research question, and to test the associated null hypothesis. For example, in single-cell RNA-sequencing analyses, researchers often first cluster the cells, and then test for differences in the expected gene expression levels between the clusters to quantify up- or down-regulation of genes, annotate known cell types, and identify new cell types. However, this popular practice is invalid from a statistical perspective: once we have used the data to generate hypotheses, standard statistical inference tools are no longer valid. To tackle this problem, I developed a conditional selective approach to test for a difference in means between pairs of clusters obtained via k-means clustering. The proposed approach has appropriate statistical guarantees (e.g., selective Type 1 error control).

In the second part of the talk, I will consider how to leverage large language models (LLMs) such as ChatGPT for biomedical discovery. While significant progress has been made in customizing large language models for biomedical data, these models often require extensive data curation and resource-intensive training. In the context of single-cell RNA-sequencing data, I will show that we can achieve surprisingly competitive results on many downstream tasks via a much simpler alternative: I input textual descriptions of genes into an off-the-shelf LLM, such as ChatGPT, to obtain low-dimensional representations of the genes, or "embeddings." I then use these embeddings as features in downstream tasks. A similar approach enables LLM-derived embeddings of cells. This work highlights the potential of LLMs to provide meaningful and concise representations for biomedical data, and also raises a number of challenging statistical questions. Addressing these questions requires bringing principled statistical thinking to the practice of modern data science.

This talk features joint work with Lucy Gao (University of British Columbia), Daniela Witten (University of Washington), and James Zou (Stanford University).

Information about building access for persons with disabilities may be obtained in advance by calling Shannon Kokesh, Department Secretary, at 773-702-8333. If you wish to subscribe to our email list, please visit the following website: https://lists.uchicago.edu/web/info/statseminars.