This quarter, I participated in a Directed Reading Project co-mentored by Shirley Mathur and Leon Tran, focusing on the methodology outlined in the paper “Assumption-lean and Data-adaptive Post-Prediction Inference.” This initiative explored advanced statistical techniques that ensure valid inference in settings where machine learning (ML)-generated data is used for downstream analyses.

We began by understanding the context and motivation for the project: the challenges inherent in using ML-predicted labels or covariates in statistical inference. Traditional methods often fail to incorporate the variability introduced by these predictions which can lead to biased estimates or overly conservative confidence intervals. The study focused on bridging this gap through methods like the “Post-Prediction Statistical Adjustment” (PSPA) framework and other techniques such as Post-Prediction Inference (PPI) and Efficient Influence Functions (EIF*).

To build a foundation, we revisited topics such as probability and linear regression, ML predictions and bias analysis as well as inference frameworks.

The implementation phase involved analyzing simulated data generated via a Random Forest regression. Key characteristics of the Random Forest model included focusing on regression, variance as well as mean squared residual. By incorporating the PSPA framework into our analysis, we observed narrower confidence intervals compared to classical methods, particularly in high-accuracy prediction scenarios. This finding demonstrated the framework’s ability to balance rigor and efficiency.

Through this project, I gained a deeper appreciation for the synergy between machine learning and statistical inference. Applying these methods illuminated how ML predictions can amplify the utility of incomplete datasets without compromising validity. The structured approach to integrating ML-derived data has broad implications, from genomics to education, where our study’s simulated example—estimating juniors’ test scores—highlighted PSPA’s robustness in practical applications. This experience advanced my technical skills, particularly in combining ML outputs with advanced inference methods, and expanded my understanding of statistical methodology in real-world contexts.