Mentee: Max Bi
Mentor: Nina Galanter
Topic: Optimal Treatment Rules, Causal Inference and Statistical Learning

Often in statistics and probability courses, the majority of the time is spent focusing on the mechanics of calculating probabilities or deriving the tools that are used to calculate them. This unfortunately often leaves little time to actually discuss much of the context around when these tools can and should be used, such as an assessment of the outcome of a randomized experiment. Throughout the course of the Statistics and Probability Association's Directed Reading Program, I was given the opportunity to put time into understanding the fundamentals of statistics by reading statistical literature, learn about basic applications of said fundamentals through optimal treatment rules, as well as reinforce concepts learned from other statistics adjacent courses such as machine learning.

The term "correlation does not imply causation" often gets thrown around because it makes a lot of sense intuitively, but advancing beyond this base level of understanding is slightly more uncommon. Through reading statistical literature on the topic, I have learned the vocabulary with which to describe this difference in more detail. Causation is directly related to the outcomes at a population level, had every individual in that population received the same exact treatment. That is to say, a causal comparison can be done in a hypothetical scenario where we can apply the same treatment to all members of a population and analyze the result versus another hypothetical world where the population received a different treatment. Correlation, on the other hand, is concerned with the outcomes on a population in which not everyone received the same treatment. An example scenario could be an observational study on a certain medical treatment in which under treatment A, 3 of 5 people die, whereas under treatment B, 7 of 15 people die. In this scenario, treatment A is associated with a higher risk of death, but if we knew that 10 / 20 would die had either A or B been given to the entire population, we can say that treatment A does not cause more deaths than treatment B. Furthermore, I learned the conditions that are required to make causal inferences, such as randomization in a controlled experiment, or when certain conditions hold in an observational study. These two basic concepts of statistics, simple as they may be, are vital to have a meaningful understanding of the field, and I was able to attain an understanding of them through the literature in the DRP.

I was able to apply my knowledge of causal inference to the concept of an optimal treatment rule (OTR). Briefly, OTRs state that in the context of healthcare, for example, it can be more optimal to treat a certain subset of patients with one specific treatment and another subset of patients with another treatment, than to give everyone the same treatment. OTRs are learned by performing calculations on covariate data from these patients, and there are a variety of ways in which they can be learned.

In this DRP, the primary method I was introduced to to learn OTRs is called Q-learning. Q-learning involves using linear regression to solve for the coefficients of a function that is equal to the expected mean outcome given certain treatment and values for covariates. Different techniques, such as regularized regression (e.g.ridge, LASSO) can be applied to the loss function in order to achieve more accurate predictions given some additional information about data. For example, if expert analysis suggests that some covariates might not have much of an effect on the mean outcome, then a loss function that encourages a sparse solution (LASSO) to the vector containing the coefficients of the regression function can be used. I was able to apply the concepts regarding regression that I learned in machine learning to a new context in the form of Q-learning, and reinforce them through designing simulations to give empirical data in support of said concepts.

Although the DRP happened to take place for me in an already very busy quarter, I'm extremely grateful for the opportunity to take a look at statistics in a non-classroom environment. It was an extraordinary asset to have a mentor who I could ask all of my questions regarding the material we covered in the program, as well as material beyond it. I want to extend my gratitude to my mentor Nina Galanter for being there to help me this quarter with my understanding of the literature, the creation of my final presentation, and for being understanding regarding my busy academic schedule.