Robust statistics refer to statistical methods and measures that remain effective and reliable, particularly even when the underlying assumptions about the data are violated, examples being the presence of outliers or non-normal distributions. Throughout this quarter, we learned many mathematical and theoretical ideas from robust statistics, but at the end of the quarter I decided to work on a simulation study.

The goal of this simulation study is to see which unbiased estimator is most robust at estimating a parameter μ from a distribution F_ϵ given data from this distribution where the underlying assumptions of that data have been violated through the presence of outliers. In this simulation study, we specifically consider the unbiased estimators of the mean, median, 0.1-trimmed mean, and midrange.

Each time you sample from this distribution, you have a probability $1 - \epsilon$ to sample from the $N(5, 1)$ distribution, and a probability ϵ to sample from the $N(5, \rho^2)$. We have $\mu = 5$ here. For this simulation study, we considered many pairs of values for ϵ and ρ, and then measured how the unbiased estimators performed in each pair of values. A higher ϵ means there's more outliers in the data, and higher ρ means the outliers are more extreme.

To get their performance for each ϵ and ρ, we sampled 1000 data points from F_e each time for 100 replications. Then, for each replication, we used the 1000 data points to obtain a value for each estimate of Mu (mean, median, 0.1-trimmed mean, and midrange). With 100 replications, we then have 100 values for each estimate of μ. We can use these values to calculate the expectation and variance for each estimate of μ. The expectation of an estimate for μ being close to the true $\mu = 5$ is good, and lower variance for an estimate of μ is also good, as then the estimate is more robust.

My results showed that the mean estimator was most robust for estimating μ when we had $\epsilon = 0.000$, meaning the data wasn't really violated with the presence of outliers. However, when ϵ was positive and nonzero, the 0.1-trimmed mean estimator was the most robust. These results didn't change between $\rho = 2$ and $\rho = 10$. Some other important takeaways were that the midrange performed really poorly, as its expectation wasn't close to the true $\mu = 5$ and the variance was really large. Also, from $\rho = 2$ to $\rho = 10$, the median and 0.1-trimmed mean didn't see much difference in performance, but the mean and midrange seemed to perform worse.

In conclusion, the purpose of this simulation study was to see which unbiased estimator is most robust at estimating a parameter μ from a distribution F_ϵ given data from this distribution where the underlying assumptions of that data were violated through the presence of outliers. The main result from this simulation study is that the 0.1-trimmed mean generally seemed to be the most robust, while the mean and median weren't that far behind, and the midrange performed really poorly compared to these other unbiased estimators.