Project: Capstone Project 1: Statistical Data Analysis

After some exploratory data analysis, one type of product was showing up frequently. Organic food seemed to be very popular and considering produce was the most popular department, this idea had to be explored further. To test if there was a difference in reorder rate based on a product being organic or non organic, a Bayesian approach was chosen to be the most efficient approach. Probability within this sample would allow us to quantify our beliefs on how different organic and non organic products are. In the initial step to using a Bayesian method a beta distribution was created. Looking at the situation as binary, alpha was used as the products that were organic and beta as the non organic products. Samples were generated from this data one thousand times using Numpy's Beta method. Once the distributions were plotted from the samples it was easy to see that the graphs of the two products never converged. This was evidence enough to conclude that in fact these two products had a significant likelihood of having different reorder rates. A binomial distribution was not needed to multiply with the beta distribution. To get a percentage of likelihood, the two samples were subtracted from each other and the mean was found of the difference. Because of the large difference in the distributions every value was over zero and therefore a one hundred percent confidence interval could be established.

Another hypothesis was tested using a Bayesian method. In this case the data was not binary and a different distributions needed to be used. To trade off the distributions used, instead of a beta distribution a Dirichlet distribution was used. To trade the binomial distribution for the distribution for the likelihood, a multinomial distribution was used keeping in mind the conjugate distribution. The hypothesis tested was if the top thirty items purchased were also the top thirty items reordered. Again, dealing with likelihood, a Bayesian method was more appropriate. First a dirichlet distribution provided samples. This could have been done multiple times but there was enough evidence to use just one instance. Also, because the graphs of the data converged so closely, it was evident that this was very likely. The multinomial distribution samples were calculated and its values counted to establish the likelihood of the equation. The dirichlet, prior was then multiplied with the multinomial distribution, likelihood and confidence was established. It was found that only forty percent of the values were over zero maintaining the null hypothesis that there was no difference in the top items purchased and top items reordered. There was little difference in the two.

Two other hypothesis test were performed using inferential statistics. These two hypothesis tested if the order in which the product was picked affected its reorder rate. Items that were picked first were to be tested first and then any order was tested. The means of each sample were calculated and its standard deviation computed by pooling together the variance. Once the standard error could be computed a t value was chosen from a t table and used in the two tailed t test. The confidence interval for both hypotheses tested did not contain zero and therefore the null hypotheses were rejected. It was found that item’s chosen order does have an effect on whether they are reordered.