Barrett Yeretzian
12/09/2018

Module 2 Final Project Write-Up

This project was a test of my ability to use SQL as well as my understanding and application of hypothesis testing. Using Microsoft's Northwind Database, I obtained the relevant data and answered four questions through hypothesis testing with Python.

The first question was whether or not a discount had an effect on the quantity of a product purchased. To start, I created dataframes after querying the order ID, product ID, quantity, and discount from the "Order Details" table. One dataframe contained orders of products that were not discounted, and the other was of discounted products. I defined my null and alternative hypotheses, and determined that a two-sided test was necessary to determine whether discount affected quantity. My analysis began by viewing the distributions of each group, and this step provided a visual representation from which I inferred that the means were different. Then, I decided to use Welch's t-test to find the answer to my question, since this test is more accurate when sample sizes and variances are unequal. The result was a t-statistic of -4.79 and a p-value very close to zero, which indicates that the null hypothesis should be rejected and that discount does in fact have a statistically significant effect on order quantity. Finally, I found Cohen's d for this test in order to quantify the magnitude of difference. A d of 0.335 indicates a small to medium difference. The second part of this question was to determine which levels of discount were effective, and I basically used the same process to find a p-value and Cohen's d for each level. Interestingly, the 20% discount was the only one that did not have a statistically significant effect on quantity.

My next question was also regarding discount, but focused on whether or not the discount had a positive effect on reorder level. Now, I used a one-sided t-test since my alternative hypothesis was that discounted items have a higher mean level of reorder than non-discounted items. To query the right information for this question, I joined the Product and Order Details tables on Product ID, and again created dataframes of discounted and non-discounted items. My methodology was the same, but my results were very different. With a p-value of 0.41, I concluded that discount does not have a statistically significant effect on reorder level.

My third question was whether the mean price of seafood items was higher than that of meat/poultry items. This query was relatively simple, but I first had to determine which category ID corresponded to each type of product. I did so by querying everything from the Category table, and then created dataframes with the product name and unit price from the Product table for the two types of items. Realizing that I had a small sample, I checked if this size was acceptable. Since it wasn't, I created samples from normal distributions with the mean and standard deviation of each group and used these samples to conduct my hypothesis testing. For this test, I decided to use a student's t-test, since the sample sizes were now the same and to try something new. I then wrote a function that would take distributions and return the p-value, decision, and visualization for a test. The results were that the p-value was very close to zero, and that the null hypothesis should be rejected – seafood products do generally cost more than meat/poultry products.

Finally, my fourth question was whether orders shipped to North America had a higher mean total price (quantity * unit price) than items shipped to Europe. This query gave me a fair amount of trouble, as I had problems with joining the tables and using the SUM aggregation. I found that rearranging the syntax of my query allowed me to get the right dataframes to answer this question. I created one dataframe with orders which had the ship region of North America, and used the LIKE function to include Northern, Southern, Eastern, and Western Europe in the next dataframe. I again used a two-sided test because my alternative hypothesis was that the two groups did not have the same mean total price per order. Also, I used Welch's t-test again to ensure that my results were accurate, and found that the p-value of the test was 0.043, small enough to reject my null hypothesis. I also found Cohen's d again, and the value of 0.21 indicated that the effect was relatively small.