Evan Okin's Blog Post – Module 3 Final Project

As a data scientist, it is important to understand how to perform hypothesis testing. Essentially, hypothesis testing is the use of statistics to determine whether or not a given hypothesis, or "thesis", is true, and with what level of certainty. For any hypothesis test, the user formulates what we call the null hypothesis (often denoted Ho) along with an alternative hypothesis (often denoted Ha). For example, let's say that a high school decides to give after-school SAT tutoring to half of the student body. We can use a hypothesis test to determine whether or not SAT performance is impacted (statistically speaking) from the tutoring. In this case, the null hypothesis (Ho) would be "There is no difference in SAT performance among students who get tutored or do not get tutored." The alternative hypothesis (Ha) would be "There is a significance difference in SAT performance after tutoring."

We then try to "rule out" the possibility that the null-hypothesis is true. We do this by computing a p-value. The p-value is the probability that a test statistic is at least as large as the one observed, observing that the null hypothesis is true. A low p-value implies that we can reject the null hypothesis, in favor of the alternative hypothesis. We compare our p-value to a significance value, call it an alpha threshold, and this is our cut-off for where we will decide whether to accept or reject the null hypothesis.

It is standard to use an alpha threshold of roughly 5%. For example, in our SAT tutoring example, if we calculated a p-value of 3%, this is less than our alpha of 5% and means that we can reject the null hypothesis, in favor of the alternative hypothesis. We can thus claim - with statistical backing - that tutoring does impact SAT performance.

For the final project for Module 3, I analyzed the Northwind Traders dataset. Northwind is a fictitious company of which we have a lot of sales data - such as order information, employee information, and supplier information. This presented a perfect opportunity to perform a hypothesis test, in order to act as a sort of business consultant to Northwind.

The data, pulled into python from a SQL database, includes 2,155 order details, of which 61% had no discounts and 39% had discounts (of any amount).

Do discounts result in an increase in sales? Intuitively, it seems plausible – customers need to pay less for goods or services with discounts, so, all else equal, we should see an increase in sales. This is where we can create a hypothesis test. The null hypothesis is that there is no difference between quantity of sales with and without discounts, while the alternative hypothesis is that there is a difference between quantity of sales with discounts. We "hope" that we can reject the null hypothesis instead of the alternative hypothesis because, as mentioned above, this would be an intuitive result that we can understand.

After setting up our test, we come to a p-value of 5.656*10^-10. This is .0000000005656, with nine zeros to start after the decimal. The p-value is almost zero! This means that there is very strong evidence to reject H0. In fact, with a p-value so small, it is almost impossible that we would observe that the same results for sales with and without discounting by chance. We thus

used a hypothesis test and drew the conclusion that discounts do result in an increase in sales (and it's statistically strong!).

It is also reasonable, as a next step, to see if the discount amount has an impact on sales. For example, it is reasonable to assume that a discount of $0.25 should result in more sales than a discount of $0.05. The overwhelming majority of the data have discount amounts of $0.05, $0.10, $0.15, $0.20, and $0.25 – very few data exists for other increments (8 data points total out of more than 2,000). Thus, we can drop these outlier data points as they are not indicative of the entire sample.

Our hypothesis test yields a p-value of roughly 10%, which exceeds our alpha threshold of 5%. We thus fail to reject the null hypothesis. We cannot confidently say that there is a significant difference between discount amounts and quantity sold. This information is helpful for Northwind Traders, because they can issue discounts towards the low-end of the spectrum (such as $0.05). By discounting less and generating the same amount of sales, Northwind Traders can profit more.

I also used hypothesis testing to test if orders differ by month, if orders differ by country, and if orders differ by which company ships packages. These examples show us how we can use hypothesis testing for "real-life" business problems.