Statistical inference Analysis of capstone project 1

The main objective of my capstone project 1 is to evaluate the correlation between seismic activities and the injection activities. To evaluate whether seismic activities and injection activities are strongly correlations, I apply two methods: confidence interval and T-test. The data we are focusing on is the total seismic numbers and total injection volumes between 2008 and 2017 in Oklahoma states. For the sake of illustration, I will only choose the data from zone 2 (there two zones of data sets in my analysis based on the location). The hypothesis we are making is that seismic distribution and injection volume distribution are the same.

1. Confidence interval

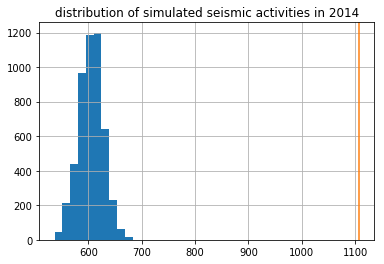
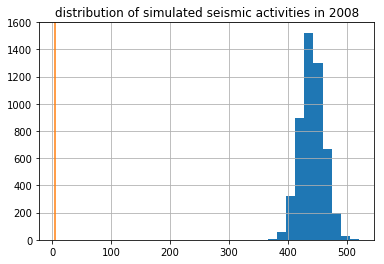
This method is what I learned while taking the course. The basic idea is that assuming the seismic follows the distribution of injection volumes, then we take random simulations for thousands of times and utilize the distribution of injection volumes to get a distribution of estimated seismic activities. If our assumption is true, the real seismic data should be within the confidence interval of the estimated seismic distribution. Otherwise, our assumption is not correct, which indicates the distribution of seismic activities and injection volumes are statistically different.

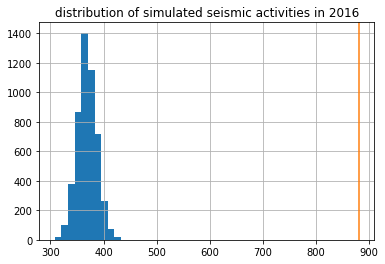
The following the functions to perform random simulations, n is the total seismic activities between 2008 and 2017, Inj\_sum\_2\_normalized is the percentages of injection volumes between 2008 and 2017.

def simulate(n):

return pd.DataFrame({'year': np.random.choice([2008,2009,2010,2011,2012,2013,2014,2015,2016,2017], size=n, p=Inj\_sum\_2\_normlized)})

For each random test, I can obtain a group of seismic data between 2008 and 2017. I did 5000 tests in total and obtained 5000 data sets of seismic numbers for each year between 2008 and 2017. Then I plotted the distribution of simulated seismic activities and the real seismic activities for a year. If our assumption is correct, the real seismic number should be within the distribution of the simulated seismic numbers. Here I show the results for three years: 2008, 2014, 2016 for the sake of illustration though I tested for all the years and the conclusion are the same. In these plots, the yellow line indicates the real seismic data and the blue blocks shows the histogram plots of the simulated seismic activities. The results indicate that the real seismic number is far away from the simulated seismic numbers, which indicates that the distribution of seismic number and injection volume is statistically different.





1. T-test

T-test compares two means of the distribution and tells the difference between each other. I utilized scipy built-in function scipy.stats.ttest\_ind to compare the distribution of seismic activities and total injection volume by evaluating t-values and p-values. T-values tells how large the difference is between two groups: the larger the t score, the more difference there is. P-values is the probability that the results occurred by chance. To test my hypothesis that the seismic distribution and injection volume distribution are the same, I compared two groups of distributions: one is the simulated distribution of seismic number using 5000 random simulations by utilizing the percentages of injection volumes between 2008 and 2017, which I already obtained in the 1st method. The other data sets I selected is the simulated distribution of seismic number using 5000 random simulations by utilizing the percentages of seismic numbers between 2008 and 2017. Thus, these two data sets have the same sample size.

I tested the t-values and p-values for three years: 2014, 2016, and 2017. In 2014, the t-value is 1131 and p-value is 0. In 2016, the t-value is 984 and p-value is 0. In 2017, the t-value is 98 and the p-value is 0. The extreme large value of t-values and 0 p-values indicate that we are very confident that seismic numbers and injection volumes are statistically different.