A Report on Hypothesis testing

1.Problem Statement

We will be analyzing the dataset from Autolib. Autolib' was an electric car sharing service which was inaugurated in Paris, France, in December 2011. The company operated throughout the week. The dataset we have contains details about the operation of electric cars within Paris. The data set shows the sum of a blue car picked everyday from each area in Paris represented by postal code. The claim being investigated is that the highest usage of blue cars is on the weekends. This is according to an article written by John Voelcker on ‘Green car Reports’.

- **Null hypothesis**: the average number of blue cars taken on weekends is less or equal to that of the blue cars taken on weekdays
- **Alternative hypothesis**:the average number of blue cars taken on weekends is greater than that of the blue cars taken on weekdays(CLAIM)

2.Data Description

In this project we used already collected data that was stored in the form of a csv file. The dataset had 13 columns and 16,085 rows. The entries were according to dates that the electric cars were taken and returned. The dates range from January 2018 to June 2018. The day of the week was represented in numerical form where 0 represented Monday and 6 represented Sunday. The dataset goes further and specifies if the day was a weekday or weekend on the column “day type’. Different areas in Paris are represented by their postal code.

The dataset has three different types of cars, blue cars, utilib and utilib 14. The dataset also shows the sum of the number of recharging slots that were freed on specific days. It also had the number of daily data points that were available for aggregation on the particular day of aggregation within the specified time periods.

The datatypes of the columns were as expected apart from the date column that we changed from ‘string’ to ‘datetime ns’. There were no null values or duplicates in this dataset. Some columns were dropped for they were not to be used in the analysis. We found the column blue cars taken to have some outliers.

3.Hypothesis Testing Procedure

The dataset at hand is very large, so before doing the hypothesis testing I first selected a sample to use. I used a stratified sampling method. I used this method because I needed to group the data according to day type and select samples from the stratas.

I first started by grouping the data by day type. There were only two day types: weekend and weekday. Then from the two stratas I randomly selected a sample. The process was quite easy because I used python programming to do so.

I decided to work with an average. It was interesting for me to know whether the average number of cars that were picked over the weekday was similar to the average number of cars that were taken during the weekends.

After sampling, I did a normality test of the two subsets of data that I extracted for comparison. My sample size was more than 30 ($n > 30$) and the two samples are from the same population. As such, I did z statistics and used the z-score to determine the p-value. This was done using the confidence level of 95% and level of significance was 5%.

4.Hypothesis Testing Results

From the hypothesis test, we found that there was sufficient evidence to prove that the average means of the blue cars taken over the weekends is greater than the mean of blue cars taken on weekdays. The test gave a p-value of 6.1068632144455e-05, this being way less than the significance level 0.05 the null hypothesis was rejected.

5.Test Sensitivity

Sensitivity in a statistical test is the measure of performance of a binary classification test. It measures the proportion of the actual positive i.e. the probability of a null hypothesis being true. In this case the sensitivity was 95%.

6.Summary

In summary;

- The data had no missing values or duplicated records .
- Large number of outliers on variables.
- The numerical variables did not follow a normal distribution.
- There was a significant drop in blue car usage in the month of february.
- In hypothesis testing, the p_value obtained after the test statistic was less than the level of significance, providing enough evidence that the average usage was higher on weekends than weekdays.