Chege John

Jedidah Ochieng

Antonny Muiko

1. **Problem statement:**

This dataset was an autolib dataset that contained details about the operation of cars within Paris. It showed a compilation of dates when the blue cars were picked from and returned to the particular addresses. The claim being investigated was whether or not the average number or blue cars taken was different from the average number of blue cars returned during that period.

**Null hypothesis**: the average number of blue cars taken is not different that of the blue cars returned.

**Alternative hypothesis:** the average number of blue cars taken is different that of the blue cars returned.

The interest of this hypothesis is crucial for the understanding of the data and its distribution.

1. **Data description**

The dataset I used for this investigation was an open dataset about cars in Paris. It contains variables like the postal code of the area which was Paris, the dates of data collection. The dates ranged between January and July of 2018. With also had the number of daily data points that were available for aggregation on the particular days of aggregation within the specified time periods. The days of the week were the usual Monday to Friday with the specifications and special assignments of days. Weekday or weekend the dataset had the specific days within the time period. The blue cars that were taken and returned, the utilib data and the slots set of data were also contained in the dataset. The problem under investigation was on the averages which would make the null and alternative hypotheses.

It was a set of data that was already collected. However, if i were to collect such comprehensive data, i would use my data response team to go out in the field, collect the data and perform the analysis from which conclusions would later on be made.

1. **Hypothesis testing procedure**

The dataset is large with 16, 085 \* 13 entries. Therefore, I picked a sample from it using a stratified sampling method. I used this method of sampling because with it, the sample would be unbiased. Stratified sampling required me to group the data into different sets of strata and then randomly pick a sample from the strata. However, considering I used python programming to generate my sample, it was quite easy and efficient. Since the dataset only had the count and dates of picked and taken bluecars, I used those to determine my hypothesis.

The logic behind my null and alternative hypothesis is that, since it was not going to be easy for me to manually group the data or identify the clusters and the sample in them, I decided to work with an average. It was interesting for me to know whether the average number of cars that were picked in a day was similar to the average number of cars that were returned on that very day. This way it would be easier to determine the future trends of business operations relating to blue cars in the autolib electric car sharing company.

From my stratified sampling, I got a sample of 296 \* 4 entries. This meant that n > 30. As such, I used the z-score to determine the p-value. I did perform normality tests on the data since the statistic I was using demanded so. The alpha level of significance that I used was 0.05.

1. **Hypothesis testing results**

From the hypothesis test, we found that there was not sufficient evidence to prove that the average means of the blue cars taken and the blue cars returned are not equal. The z-score was -1.3163 and as a result, the null hypothesis was not rejected. The z-critical value was 1.959963984540054 with the confidence interval being

Confidence interval:

(38.67676993633314, 80.92458141501822)

The p-value was 0.0940 which was greater than the significance level.

1. **Discussion of 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 91%.

1. **Summary and conclusions**

The project was comprehensive and demanding. I performed exploratory data analysis with the hypothesis testing as its implementation. Conclusively, I failed to reject the null hypothesis because there was not enough evidence for me to reject the null hypothesis.