**REPORT ON HYPOTHESIS TESTING ON BLUE AUTOLIB CARS**

## **1.PROBLEM STATEMENT**

Working as a Data Scientist for the Autolib electric car-sharing service company, we’re tasked to investigate a claim about the blue cars.

The dataset being investigated is provided by the company. Although the dataset contains data from three types of cars, our focus is on the blue cars.

The null hypothesis is that the mean number of blue cars returned in a certain area is different compared to the mean from another area during weekdays.

The alternate hypothesis is that the mean number of blue cars returned in a certain area is not different to the mean from another area during weekdays.

This hypothesis is important as it will give insight to the company to really know if different areas have different means of blue cars returned and this will help the company make better decisions in the distributions of the electric blue car to different areas.

## **2. DATA DESCRIPTION**

The dataset, as mentioned, contains electric car usage of a car-sharing company. The sum of the blue cars returned is the main focus as this is what relates to our hypothesis.

The dataset used was sampled from the main dataset provided by the company and a sample of 1600 entries were chosen as this is 10% of our entire dataset.

The simple random sampling method is used to choose samples from the dataset as this method is not biased and it is a very simple way of selecting samples.

The second sample used is of the two postal code areas randomly chosen.

In addition to the two postal codes, the main focus was also to pick only the weekdays.

Our final dataset, therefore,only contained information on the blue cars returned during weekdays in the above mentioned postal codes.

Our dataset, however, is not normally distributed and this is as a result of so many outliers which could not be removed as they were many and proved useful in our analysis.

## **3. HYPOTHESIS TESTING PROCEDURE**

The following is the procedure taken to test the hypothesis:

* The dataset is loaded, explored and cleaned before conducting univariate and bivariate analysis on it.
* A sample is chosen from the entire dataset. This sample is 10% of all the entries in the main dataset i.e. 1600 entries. This sample is chosen using the Simple Random Method.
* From the chosen sample, two postal codes are chosen using the same simple random sampling method in order to focus on only two areas.
* From the first sample, only weekdays are chosen as this is an objective in our hypothesis. This sample becomes our new dataset.
* Two more datasets are created representing the 2 postal codes initially randomly chosen from the second sampling.
* Z-test is then be used to test for the hypothesis

The null hypothesis, i.e. the mean number of blue cars returned in a certain area is different compared to the mean from another area during weekdays, is important and it came from the fact that different areas will always have different populations of people using the blue car.

It is interesting as it explores the difference in preference of people from certain areas as the blue cars will not always be equally returned in all areas.

The alternate hypothesis discredits this hypothesis although it logically makes sense.

The choice of test static used here is the z-test. This test was chosen because the hypothesis being tested is to determine whether two means are different. The variance is also known and the sample size is large. However, the dataset is not normally distributed hence this assumption was not satisfied.

The alpha level used in this case is 0.05.

## **4. HYPOTHESIS TESTING RESULTS**

The hypothesis test resulted in accepting the null hypothesis.

The value of the z test is -0.261(3 d.p), the p-value is 0.397(3 d.p).

The point estimator of the parameter is as calculated below:

* The population mean of the blue cars returned is 125.91
* The sample mean is 130.36
* The difference in these means i -4.444
* We can therefore state that the point estimator for the mean of the sample underestimates the true mean by 4.444.

Constructing a confidence interval:

* Selecting a confidence level of 95%
* Finding the standard error(SE) of the mean i.e.

SE = s / sqrt( n )

SE = 187.4 / sqrt(1000) = 5.913(3 d.p)

* Finding the critical value using the alpha as 0.05, critical probability of 0.95 and degrees of freedom as 1599(n-1). The critical value is 1.96
* Computing the margin of error(ME) i.e. ME = critical value \* standard error = 11.59
* The 95% confidence level around the mean is 125.91+/-11.59

## **5. DISCUSSION OF TEST SENSITIVITY**

The null hypothesis was accepted as this was the logical argument and discrediting it would have required huge evidence. Changing the alpha level to a higher level would have also resulted in rejection of the null hypothesis.

Using a different sample size i.e. a smaller sample size would have also resulted in rejection of the null hypothesis as the sample size would be so small resulting in change in the z score and p value.

## **6. SUMMARY AND CONCLUSIONS**

This test was successful as the data analysis procedure was followed. The data was cleaned and analyzed correctly, resulting in better results .

The hypothesis test done was fruitful and gained insight on the mean difference for different areas where the blue cars are returned.

The test proved that indeed the means of the blue cars returned actually differs from different areas. The test results might however change if the sample size changes.