Date: 11/5/2021

Data Report: Autolib Hypothesis Report

Name: Andrew Wairegi

Role: Data scientist

Github Repository:

https://github.com/andrew-wairegi-school/core-week4-IP

Collab Notebook:

https://colab.research.google.com/drive/1T8Bl3TDlR2nUKHGqymnQgZdq\_F10vhqN?usp=sharing

Contents

[Problem statement 3](#_Toc87037350)

[Data description 3](#_Toc87037351)

[Hypothesis testing procedure 4](#_Toc87037352)

[Hypothesis testing results 4](#_Toc87037353)

[Test sensitivity 5](#_Toc87037354)

[Summary 6](#_Toc87037355)

[Conclusion 6](#_Toc87037356)

# Problem statement

I want to create a hypothesis. So that I may be able to tell whether one location has higher values of blue car rentals than another. This statement will be created based on the data that I find out during the data exploratory phase of my notebook. This will allow me to create a hypothesis that I can test based on existing data. As opposed to creating a hypothesis of which I have no basis on. That is without any prior information. Which would mean that my hypothesis has no basis, and may be wrong without any experimentation.

After doing my exploratory analysis. I found that there were two points which I could test in my hypothesis. Which would allow me to have a fair hypothesis that might be true or not. Based on the hypothesis test that I propose.

I questioned, that whether the location with Postal code 75017 would have the same value as 75015. Meaning in the same area, with a margin of error. That was my alternative hypothesis. According to the data, my null hypothesis. Was that the location with postal code 75017 would not have the same number of car rentals as 75015. But would be different. That means it would not fall there, even if there was an error.

# Data description

According to the graph that I plotted of mean car rentals per month, of different postal codes. 75015 had the highest number of car rentals. While 75017, had the third highest number of car rentals. So I wanted to determine whether, it could be counted as part of the peak. Since they were not too far behind.

In terms of the data for the hypothesis test. I got the standard deviation of Postal code 75015, which was 4452. That meant that any value that was between 23320, with an error of 4452. Could be inclusive of the peak, as that means it would not have missed the mark of between 18868 and 27772, per month. However, this would not work with all values, as we have an average over several months. So as the year progresses, this actually means that people might have changed their spending habits throughout the year. Which is normal.

Our data, the postal code that we are selecting is 19096. Which means it does fall in this boundary. But it might not be part of the peak (or the high peak). However, it is still a good value that we can look at. To see if it can be counted as part of the high peak.

Regarding the other data, the mean of the postal code location is 23320. While X is the number of bluecars that were taken from the 75017 postal code. This will allow us, to calculate the P-value. Using the Z-score method. As we already have the standard deviation, and mean from the data set.

# Hypothesis testing procedure

I will create the P-value, using a z-score. The z-score will allow me to get the deviation of the point, from the postal code bluecar count. As we will take it as the mean. Since this is the point we are measuring from. Then I will put the z-score in to a normal distribution. To allow me to calculate the probability of getting that z-score, from the postal code. This will then be the p-value that I will use in my calculations.

This p-value allows me to know the probability of getting the bluecar value, from the postal code. Given that the null hypothesis is true, which is what the p-value is calculating. The null hypothesis was that, the number of bluecars from the 75017 postal code are different from the 75015 postal code, and they cannot be put in the same range. Meaning that they do not have the same number of cars, and that it was not some kind of error when counting the number of cars.

After getting the P-value, I will test it with the significance level. Which I have set at 5%. This will allow me, to get an accurate result. As I will only allow it to have an error of 5%. For it to be allowed, that they may have had the same value.

As the procedure goes if they have a p-value of over 5%, I will reject the alternative hypothesis. Moreover if the p-value is 5% or less, I will accept the alternative hypothesis. As well as reject the null hypothesis.

**Alternative hypothesis**: Can the number of bluecars in Postal code 75017 be in the same range as 75015 ?

**Null hypothesis**: The number of bluecars in Postal code 75017 is different from the postal code 75015

**Significance level:** 0.05 (5%)

# Hypothesis testing results

The p-value that I obtained was 0.17137. Which meant that we had a 17% chance that the value of 19096 bluecars, could be within the postal code 75015. Which meant that it was wrong. As we were only allowing a chance of 5% and below. For it be accepted that it is within the 75015 postal code. This means that we accept the null hypothesis and reject the alternative hypothesis.

It also means that we have a confidence interval of only 83%. That the 75017 postal code lies within the 75015 postal code, in terms of blue cars. So that means it can be outside by 17%, in other words. This is our confidence interval.

# Test sensitivity

The sensitivity of the test is very accurate. Because it has been proven through research. That the normal distribution gives us an accurate score of the probability of an event happening, when it is given a z-value. This has been proven by research. That is why we use the normal distribution & z-value or t-value.

In terms of the significance level. I choose a rate of 5%. So that the chance of you getting a value that is outside of the given mean, is only 5%. Which means that it would fall in a confidence interval of 95%. As it would be within a 5% chance, of being close to the mean.

However, we will not be calculating the ranges themselves. Using backwards mathematics.

# Summary

In summary we are creating the p-value using the z-score and normal distribution curve. But if the p-value is above our significance level, we will reject our alternative hypothesis. While if it is below, then we will accept it. As that is a narrow bar of 95% confidence interval. That it is within our range. This will determine whether our postal code is in range of the peak postal code, or not.

# Conclusion

In conclusion our p-value was over 5%, as it was 17%. Which means that it did not fall in our confidence interval of 95%. But 83%. Which means that we are not sure if it would fall within our range. While it may fall within our range if it was 90%. We want to know whether it can be counted as among the high peaks. Which means it has to be within 5%, of the peak. That means it was neither within our targeted 5%, or the alternative 10%. Which would have allowed to be within the peak itself.