**Moringa Data Science Course: Core**

**WK4 IP**

**Hypothesis testing report**

**Author: Martha Irungu**

1. **Business Overview**

Autolib’ was an electric car-sharing company owned by Bollore industrial group that operated in France in multiple cities including Paris between 2011 and 31st July 2018. Autolib owned cars could be leased by public and returned after use. Since they were electric cars, there were charging stations located at different locations where customers could pick and drop back the cars after use.

The model of electric cars used by Autolib was:

* Blue car - A passenger electric car.

More background details about this car are available on this [link](https://en.wikipedia.org/wiki/Bollor%C3%A9_Bluecar)

* Utilib car - This is a red-blue car. Apart from the passenger model, it has a delivery variant that has two seats to create more space for cargo. It’s red in color.
* Utilib 1.4

I am working as a data scientist for Autolib, the car-sharing service company and have been tasked by the management to investigate a claim that the population mean of blue cars taken in weekdays is the same as that taken over the weekend, meaning mu1-mu2=0

1. **Problem Statement**

As a data scientist, I joined Autolib one month ago and the management believes that the population mean of Blue cars taken in weekdays is equal to the blue cars taken over the weekend.

Additionally, the management of Autolib is getting pressure from employees who want to take a day off over the weekend. The management is hesitant to allow this as they believe that the mean population of blue cars taken on weekdays is the same as those taken over the weekend. As a new employee,I have been tasked to investigate this claim and by gathering evidence to determine if there is a significance difference.

The hypothesis will be:

* 1. The Null hypothesis is:

Ho:> Blue cars taken weekdays - Blue cars taken weekend=0 (equals)

* 1. The Alternative hypothesis:

H1> Blue cars taken weekdays - Blue cars taken weekend !=0 (not equals to)

This hypothesis is interesting since it will validate or unvalidate the claim that on average the blue cars taken during weekdays is the same as those taken over the weekend. This therefore means that Autolib serves more customers within the span of 2 days as over the weekend we only have 2 days while we have 5days during week days.

Additionally, as a new employee of Autolib, I would like to advise the management, the need to have more staff to serve the customers over the weekends, as is I believe that the number that serves over the weekend is not commensurate to the number of blue cars customers that visit the stations over the weekend.

1. **Data Description**

The Autolib dataset has 16,085 entries and 13 columns. The data types are object and integers. The population of blue cars taken is one of the columns and has a population of 16,085 entries. The cars are taken on weekdays and weekends from various postal codes/locations. Our dataset had no missing values. The blue cars taken had outliers that we dealt with. We randomly generated a sample of 300 entries from blue cars taken in weekdays and weekends out of the possible records of 9,783 to perform hypothesis testing. The columns and description of the dataset is as below:

|  |  |
| --- | --- |
| **Column name** | **Description** |
| Postal code | postal code of the area (in Paris) |
| Date | date of the row aggregation |
| n\_daily\_data\_points | number of daily data poinst that were available for aggregation, that day |
| dayOfWeek | identifier of weekday (0: Monday -> 6: Sunday) |
| day\_type | weekday or weekend |
| BlueCars\_taken\_sum | Number of bluecars taken that date in that area |
| BlueCars\_returned\_sum | Number of bluecars returned that date in that area |
| Utilib\_taken\_sum | Number of Utilib taken that date in that area |
| Utilib\_returned\_sum | Number of Utilib returned that date in that area |
| Utilib\_14\_taken\_sum | Number of Utilib 1.4 taken that date in that area |
| Utilib\_14\_returned\_sum | Number of Utilib 1.4 returned that date in that area |
| Slots\_freed\_sum | Number of recharging slots released that date in that area |
| Slots\_taken\_sum | Number of rechargign slots taken that date in that area |

The dataset is a daily aggregation, by date and postal code, of the number of events on the Autolib network (car-sharing and recharging).The data is sourced from Autolib’ which was an electric car-sharing company owned by Bollore industrial group that operated in France in multiple cities including Paris between 2011 and 31st July 2018. Autolib owned cars which were which could be leased by public and returned after use. Since they were electric cars, there were charging stations located at different locations where customers could pick and drop back the cars after use.

The model of electric cars used by Autolib was:

* Blue car - A passenger electric car.

More background details about this car are available on this [link](https://en.wikipedia.org/wiki/Bollor%C3%A9_Bluecar)

* Utilib car - This is a red-blue car. Apart from the passenger model, it has a delivery variant that has two seats to create more space for cargo. It’s red in color.
* Utilib 1.4

1. **Hypothesis Testing Procedure**

Hypothesis is a claim that we are trying to investigate. The basis of the hypothesis that we will investigate is:

The management of Autolib are getting pressure from employees who want to take a day off over the weekend. The management is hesitant to allow this as they believe that the mean population of blue cars taken on weekdays is the same as those taken over the weekend.

I have been tasked me to investigate this claim to gain evidence that there is no significance difference between the two. This will strengthen their basis of management team argument as they believe the company need good number of employees to serve the high number of customers on the weekends (as this has only 2 days), while weekdays has 5 days.

**Tests**

Given the different tests, I will use both z- and t-test to perform the hypothesis testing.

**Z-test**

We choose this by benchmarking it against central limit theorem, since we have more that 2 data points, we expect that it will follow a normal distribution. This is also 2-tailed as we may have data falling on either side of the distribution. The samples from each population assumed to be independent of one another. The populations from which the samples are taken, has sample sizes is larger than 30 (i.e. n1≥30 and n2≥30

**We will create the hypothesis as below:**

Null Hypothesis: Ho:=> mu1-mu2=0 # the mean are the same Alternative Hypothesis: H1: => mu1-m2!=0. # the mean are not the same**.**

In this case we will use Z-score two tailed test**.**We set our level of significance (alpha) to 0.001 area of rejectionon z score will be two sided test critical z value is 3.09

1. **Hypothesis Testing Results**
   1. **Using Z-test**

Below are the results of using z-test.

Our observed value is -11.6533333336 this is less than 3.09 so we FAIL TO REJECT the null hypothesis, which stated that the mean of blue cars taken in weekdays is the same as the mean of blue cars taken over the weekend. We therefore have enough evidence that the mean of cars taken in weekdays minus the mean of cars taken durind weekdays is zero

**b) Using t-test**

We also used test, we assumed the values of the population are not known, so we used the sample data to compute t-test value and the p-value. The following conditions were also met

* The data is continuous (not discrete).
* The data follow the normal probability distribution.
* The two samples are independent. ...
* Both samples are simple random samples from their respective populations.

Below are the results of using z-test.

The P-value is 0.0024288, this is less than our alpha which is 0.05, in this case we will REJECT the null hypothesis.

**c)Construct a confidence interval around the parameter**

On conducting the confidence interval of the parameter

i)At 0.05 significance level, meaning 95% confidence level, the outcome was an interval of

-17.92505125173453 <= mu1 - mu2 <= -5.381615414932142

Therefore, there was no chance that zero will fall in this estimation.

ii)At 0.01 significance level, meaning 99% confidence level, the outcome was an interval of

-23.470167703979488 <= mu1 - mu2 <= 0.16350103731281607

Therefore, there is no chance that zero will fall in this estimation.

* + 1. **Discussion of Test Sensitivity**

**Type I and Type II Errors as tabulated below:**

|  |  |  |
| --- | --- | --- |
| Conclusion from the analysis | Effect found in the population | |
| yes | no |
| Reject the Null Hypothesis | Correct decision | Type 1 error |
| Fail to reject the Null Hypothesis | Type ii error | Correct decision |

There is a likelihood to commit these errors when making conclusions

**Type 1 Error**

This error occurs when we reject the null hypothesis when we should have retained it. That means that we believe we found a genuine effect when in reality there isn’t one. The probability of a type I error occurring is represented by α and as a convention the threshold is set at 0.05 (also known as significance level). When setting a threshold at 0.05 we are accepting that there is a 5% probability of identifying an effect when actually there isn’t one

**Type II error**

This error occurs when we fail to reject the null hypothesis. In other words, we believe that there isn’t a genuine effect when actually there is one. The probability of a Type II error is represented as β and this is related to the power of the test (power = 1- β). Cohen (1998) proposed that the maximum accepted probability of a Type II error should be 20% (β = 0.2).

* + 1. **Summary and Conclusions**
       - In summary we have implemented our hypothesis testing using the following steps:
* Clearly stating a null hypothesis and alternative hypothesis
* Clearly stating an alternative hypothesis
* Give the value of the test statistic
* Report the P-value
* Clearly state your conclusion (i.e. ‘Reject the Null’ is not sufficient)
  + - * From our implementation, it is evident that there is likelihood of making the wrong conclusion due margin of error when making the conclusion. This is the reason why there is always a provision of standard error. When we used z-test we failed to reject the Null hypothesis and when we used the t-test we were supposed to reject the null hypothesis.