Independent Project (Week 4) - Hypothesis Testing

The Google Colaboratory notebook used for this analysis can be found here.

Problem Statement

The data used for this analysis is from Autolib, and electric car sharing company based in France. The dataset contains an aggregate count of bluecars taken and returned per day and the postal code for each station that the data is taken from.

The variable under investigation here is the number of Bluecars taken in a day from a specific area.

The null hypothesis is that there is no difference between the number of Bluecars taken from stations with postal codes under 80000 and those over 90000.

The alternative hypothesis is that there is a statistically significant difference in the number of Bluecars taken from stations with postal codes under 80000 and those over 90000.

This information is essential for Autolib to understand where their product is most popular, and can help them to understand how frequently their service is used at different stations. They make decisions on whether to continue the service in those areas, or how to increase use of Bluecars in certain areas.

The main assumption is that the number of Bluecars taken is similar to the number of Bluecars returned, and therefore we can use one of them as a metric for activity at those stations.

Data Description

The dataset is an aggregation of data collected by Autolib and made available to the public. They regularly collected data on the number of Bluecars taken and returned, and the number of slots available at each station. This data was automatically uploaded to the main database multiple times a day. The postal code for each station is provided for analysis of regional use of the service. The data also has information on the day of the week when the data was collected, and whether it was a weekday or weekend.

The variable under investigation (number of Bluecars taken) is the number of Bluecars taken that date in that area. The data is not normally distributed, with a skewness of 2.38 and kurtosis of 6.02.

Some dates had multiple data points missing, so dates with more than 3 missing data points for aggregation were dropped from the dataset.

The dataset generally has more records from the weekend, so data from dates during the weekend were used since the service is used most on these days. Data form the weekends was also more complete since most of the dates with missing values that were dropped were weekdays.

From the bivariate analysis, there seems to be a correlation between postal code and number of bluecars returned.

Hypothesis Testing Procedure

The bivariate analysis showed that the number of Bluecars taken was higher, on average, at stations with postal codes below 80000 than stations with postal codes above 90000. The difference was evident from the visualizations, but hypothesis testing was required to confirm that the difference is statistically significant. This would help inform business decisions on how they market their service and how to distribute their cars to different locations according to demand. The hypotheses to test the significance of this difference were formulated as follows:

H_0: There is no difference between the number of Bluecars taken from stations with postal codes under 80000 and those over 90000.

H_a: There is a statistically significant difference in the number of Bluecars taken from stations with postal codes under 80000 and those over 90000.

For this question, a Mann-Whitney U test will be used to compare the mean number of Bluecars taken from stations with postal codes below 80000 with those from stations with postal codes above 90000 since the Bluecars data is not normally distributed and the groups are independent of each other.

The alpha level used is 0.05.

Hypothesis Testing Results

The test statistic U = 27190.000, and the p-value = 0.000.

The p-value is lower than the significance level (α = 0.05), so the null hypothesis can be rejected in favour of the alternative hypothesis.

Based on this analysis, there is a significant difference in Bluecars taken at postal codes below 80000 and those above 90000.

Summary and Conclusions

The exploratory data analysis provided a null hypothesis of no difference between Bluecars taken in different regions. The null hypothesis was rejected after a Mann-Whitney U test and the result shows that there is a significant difference.