AUTOLIB DATASET.

HYPOTHESIS TESTING REPORT

BY

KELVIN NJUNGE

Problem Statement

Autolib is an electric car sharing company. The company has 3 main types of electric cars: blue cars, utilib counter and the utilib 4 counter. These cars operate in Paris, France.

The claim is that the mean number of blue cars returned in a certain area is equal to the mean of the returned blue cars from another area during weekends.

The random variable taken into consideration for this problem was the "BlueCars_taken_sum" which is the number of blue cars taken at a particular date in an area (Postal Code within Paris).

The Null Hypothesis(Ho) for this test was that the average number of blue cars taken during the weekdays in a randomly selected postal code will not be different to the average taken from another randomly selected postal code. Conversely, the Alternate Hypothesis(Ha) was that the average number of blue cars taken during the weekdays in a randomly selected postal code will be different to the average taken from another randomly selected postal code:

- Null Hypothesis(Ho): $\mu(\text{taken weekday}|PC1) = \mu(\text{taken weekday}|PC2)$
- Alternate Hypothesis(Ha): $\mu(\text{taken weekday}|PC1) \neq \mu(\text{taken weekday}|PC2)$

where μ is the mean number of cars taken during weekdays, PC1 and PC2 are the two randomly chosen postcodes from the dataset.

The above test was chosen as a means of focusing the analysis carried out on the dataset to gain invaluable insights. In answering the above question we are able to shed light on potential differences in operational zones/areas which can inform decisions on how to allocate resources and possible opportunities that are untapped.

Level of significance:

Refers to the degree of significance in which we accept or reject the null-hypothesis. 100% accuracy is not possible for accepting or rejecting a hypothesis, so we therefore select a level of significance that is usually 5%. This is normally denoted with alpha and generally it is 0.05 or 5%, which means your output should be 95% confident to give a similar kind of result in each sample.

Data description.

Usage of electric cars has seen growth with the advancement of technology. Unlike the fuel cars, electric cars are more environmentally friendly and thus sustainable. An increase in electric car usage will mean a well-kept environment due to reduced air and noise pollution and a lot of other factors.

The data set is provided by the Autolib car sharing company. It contains a daily aggregation, by date and postal code, of the number of events on the Autolib network.

We performed exploratory data analysis on the data set where we found that it contains 14 columns and 16065 rows which are as described below.

Column name	**Explanation**
Postal code	postal code of the area (in Paris)
date	date of the row aggregation
n_daily_data_points	number of daily data points 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 blue cars taken that date in that area
BlueCars_returned_sum	Number of blue cars 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 recharging slots taken that date in that area

The data set had neither missing values nor duplicated data. Most of the data types were integers. We dropped some columns that we were not going to use for our analysis. Any syntax errors were corrected and all the column names changed into lowercase. The date column which was an object data type was converted into date-time format.

The random variable used for our testing purpose was the 'BlueCars_taken_sum'. The general analysis showed that the mean number of blue cars taken to be about 38.45 for the two postal codes randomly selected with a standard deviation of about 25.74 and a minimum number of 0 cars and maximum of 97 cars in a day taken over the duration of the dataset. Given the wide range of about 90 cars taken in a given day, it is expected that the Null hypothesis will be rejected and that there's bound to be a significant difference in mean number of cars taken a day between the randomly selected postal codes (this will be verified analytically).

Hypothesis Testing Procedure

Hypothesis testing for this project will be carried out by first selecting randomly two postal codes, taking the new data as the samples, then further subsetting of the data so as to have two independent samples (corresponding to the respective postal code selected). The new datasets (samples) will thereafter be tested according to the relevant test that is based on the kind of data distributions they closely resemble and the size of the sample (n). A p-value will then be calculated and compared to a significance level of 0.05 chosen for this test (most common level of significance in statistical calculations). The p-value obtained will inform whether to accept or reject the null hypothesis.

In emerging businesses, it's not common to see significant differences in product uptake per region, assuming that pre-launch surveys and studies were done properly which resulted in data driven rollouts of the product. In this test, we want to check the hypothesis that there aren't significant differences in the mean number of blue cars taken in different postal codes (two randomly chosen for our case).

Given the size of the dataset (n = about 16,000 records) and assuming the normalcy of the data distribution then a Z - test is preferred as the assumptions/requirements have been satisfied, even though we don't have the population parameters (Population here means all Autolib blue car sharing data since they started). The distribution requirement will be checked in the workbook.

Hypothesis Testing Results

The results for the test that was based on the two random postal codes selected: 92320 and 94450 was that the null hypothesis was rejected as there was enough evidence to show a significant difference between the average numbers of blue cars taken in the two postal codes.

The p-value calculated using the z test was approximately equal to zero, which was the same as the t-test used as a check. Since zero is less than the significance level we had set (0.05), this means that the result is significant and that the null hypothesis is rejected and the alternate hypothesis is accepted.

Discussion of Test Sensitivity.

The test was done on a sample statistic, checking if the mean was different in different postal codes. The significance of the result obtained is that we can infer to some degree that there are differences (significant ones) in blue cars taken depending on postal codes.

This should push the company to review all postal codes to either re-think/ re-strategise the viability of some postal codes or redistribute resources according to demand so as not to waste them needlessly. It should be highlighted that changing sample size will affect the standard error of the test. Increasing the sample size will yield more accurate results (minimizing standard error) and reducing the size will have the opposite effect.).

Summary and Conclusions

The project entailed checking the hypothesis that for two randomly selected postal codes, will their mean number of blue cars taken in a day be different. To perform the test the dataset was sampled by use of selecting two random postal codes (using a random selector of the ordered postal codes contained in the dataset) and their accompanying data.

Subsequently, the data was constrained to weekdays and two separate data frames (PC1 and PC2) were created to store the same according to the two different postal codes. These two data frames were then used as two samples to carry out an independent two-tailed Z-test which was verified with an Independent two-sample T-test.

The results of the hypothesis test done in this project was a rejection of the null hypothesis (p-value < 0.005) conclusively and acceptance of the alternate hypothesis that there is a difference in the mean number of blue cars taken in a day in different postal codes (92320 and 94450 in this case