**Take home challenge:**

**Part-2: Experiment and metric design**

**Experiment design (case study):**

Here we are going to design an experiment to see whether the implementation of ‘Toll reimbursement” is helping the drivers to visit both cities or not? So, there will be collection of data in two circumstances:

1. Toll will not be reimbursed when drivers visit both cities.
2. Toll will be reimbursed when drivers visit both cities.

**Data collection:**

Let us collect the data for 12 months. First 6 months without reimbursing the toll and subsequent 6 months reimbursing the toll. We’ll consider 50,000 drivers for the study.

**Table-1 (without toll reimbursement)**

If driver has visited both cities label: 1, else 0.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Driver Serial No. | Day1 | Day2 | …….. | ……. | ……… | ……….. | ……. | ……… | Day180 |
| 1 |  |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |  |
| ….. |  |  |  |  |  |  |  |  |  |
| …… |  |  |  |  |  |  |  |  |  |
| ….. |  |  |  |  |  |  |  |  |  |
| 50000 |  |  |  |  |  |  |  |  |  |

**Table-2 (with toll reimbursement)**

If driver has visited both cities label: 1, else 0.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Driver Serial No. | Day1 | Day2 | …….. | ……. | ……… | ……….. | ……. | ……… | Day180 |
| 1 |  |  |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |  |
| ….. |  |  |  |  |  |  |  |  |  |
| …… |  |  |  |  |  |  |  |  |  |
| ….. |  |  |  |  |  |  |  |  |  |
| 50000 |  |  |  |  |  |  |  |  |  |

We’ll sum the number of drivers those visited both cities on everyday (Day1, Day2..Day 180)

**Evaluation metrics:**

Evaluation metrics will compare the averages of both distributions. Based on that we’ll formulate the Null hypothesis and Alternative Hypothesis as follows:

**Null hypothesis(H0):** The averages of both distributions are same.

**Alternative Hypothesis (H1):** The averages of both distributions are not same.

Set the statistical significance label: 0.05. Plot the difference between the two means. If in the distribution, the p value for extreme difference is < 0.05, then we reject the null hypothesis that “The averages of both the distributions is same”.



**Statistical analysis:**

First, we’ll plot the distribution of “No. of drivers visited both cities” vs. counts in both circumstances mentioned above and check visually both are normally distributed or not.

Case-1: If both the distributions are normally distributed, then we’ll carry out the two-sample test to see both averages are same or not.



Case-2: If both distributions are not normally distributed, we’ll use the non-parametric test (permutation test) and plot the difference of their means. If the p < 0.05, then we’ll reject the null hypothesis and accept the alternative hypothesis. Else we’ll not reject the null hypothesis and accept it.