**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 under two circumstances:

Circumstance-1: Toll will not be reimbursed when drivers visit both cities.

Circumstance-2: Toll will be reimbursed when drivers visit both cities.

**Evaluation metrics:** ‘Increase in drivers’ number after implementation of Toll’

An explanation to the evaluation metric is that, after implementation of Toll reimbursement if the number of drivers visiting both the cities increases, we’ll consider our experiment to be successful. But if it goes opposite way i.e after Toll reimbursement, if number of drivers visiting both cities decreases/remains same, then we’ll consider our experiment to be not successful.

**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 100 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 |  |  | 0 |  |  |  |  |  |  |
| 2 | 1 |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  | 1 |  |  |  |
| ….. |  |  |  |  |  |  |  |  |  |
| …… |  |  |  | 0 |  |  |  |  |  |
| ….. |  |  |  |  |  |  |  |  |  |
| 100 |  |  |  |  |  |  |  |  |  |

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

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

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Driver Serial No. | Day1 | Day2 | …….. | ……. | ……… | ……….. | ……. | ……… | Day180 |
| 1 |  |  |  | 1 |  |  |  |  |  |
| 2 | 1 |  |  |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |  |  |  |
| ….. |  |  |  |  |  |  |  |  |  |
| …… | 0 |  |  | 1 |  |  |  |  |  |
| ….. |  |  |  |  |  |  |  |  |  |
| 100 |  |  |  |  |  |  |  |  |  |

After data collection we’ll sum the number of drivers those visited both cities on each day (Day1, Day2, ….., Day 180)

**Table-3: No. of drivers visiting both cities on each day (without Toll reimbursement)**

|  |  |
| --- | --- |
| Day | No. of drivers visiting both cities |
| Day1 | 50 |
| Day2 | ….. |
|  | 20 |
|  | ….. |
|  |  |
|  |  |
| Day180 | 80 |

**Table-4: No. of drivers visiting both cities on each day (with Toll reimbursement)**

|  |  |
| --- | --- |
| Day | No. of drivers visiting both cities |
| Day1 | 35 |
| Day2 | ….. |
|  | 70 |
|  | ….. |
|  |  |
|  |  |
| Day180 | 90 |

data\_without\_toll\_reimbursement = [ 50, …., 20, …, …., 80]

data\_with\_toll\_reimbursement = [ 35, .., 70, …., …. , ….. 90]

**Statistical test:**

Here, the evaluation metric depends on the statistical test that based on the observations mentioned above two circumstances. We’ll plot the ‘No. of drivers visiting both the cities each day’ (x-axis) vs ‘No. of days’ (y-axis) We’ll compare the averages of two 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 (ie no. of drivers visiting both cities does not increase after implementation of Toll reimbursement).

**Alternative Hypothesis (H1):** The averages of both distributions are not same (i.e no. of drivers visiting both cities increases with implementation of Toll reimbursement).

**Statistical analysis:**

First, we’ll plot the distribution of ‘No. of drivers visited both cities on each day’ vs. ‘No. of days’ in both circumstances mentioned above and check visually both are normally distributed or not. Also we’ll check by statsmodels in built function

stats.normaltest(distributions without Toll reimbursement)

stats.normaltest(distribution with Toll reimbursement)



**or**

**Code:**

from scipy.stats import Shapiro

def normal\_test(data, alpha):

shapiro\_test\_results = shapiro (data)

t\_statitics = shapiro\_test\_results.statistic

p\_value = shapiro\_test\_results.pvalue

print(‘t\_statistics:’, format(.3f), ‘p\_value: ’, p\_value)

If p\_value > 0.05:

print(“ Data is normally distributed” )

else:

print(“data is not normally distributed”)

return pass

data\_without\_toll\_reimbursement = [80, 50, ……., …..,…………………………………….]

data\_with\_toll\_reimbursement = [ 20, 70, …….., ………, ………,………………………... ]

normal\_test(data\_without\_toll\_reimbursement, 0.05)

normal\_test (data\_with\_toll\_reimbursement, 0.05

After normal test, we’ll carry out the hypothesis test for two conditions:

**Two distributions are normally distributed:**

If both distributions are normally distributed, then we’ll carry out the two-sample test to see they have same distribution or not by comparing their averages. We’ll set the statistical significance label: 0.05. In the distribution if the mean of case-2 (i.e with Toll reimbursement), the p value is < 0.05, then we’ll reject the null hypothesis (that the averages of both the distributions are same) and accept the alternative hypothesis (i.e the averages in two cases are not same and it has increased in case-2).



**The task will be carried out in following steps:**

**Step-1:**

No. of observations, mean and Standard deviation of the distribution (Circumstance-1): **n0,** Graphical user interface, text, application, Word

Description automatically generated**, s0**

No. of observations, mean and Standard deviation of the distribution (Circumstance-2): **n1,** Text

Description automatically generated**, s1**

**Step-2:**

We’ll use the t-test for comparison. So we need to use two formulae here:

Text

Description automatically generated

Sp : Pool standard deviation of the both distributions

**Step-3**

And calculate the t value, then the corresponding p-value applying

**Code:**

p = scipy. stats . t . cdf(t, n0 +n1 -2)

Or

p = scipy. stats . ttest\_ind( case-1 data, case-2 data, equal\_var = True)

If p < 0.05 then we’ll reject the null hypothesis and accept the alternate hypothesis.

**Two distributions are not normally distributed:**

If both distributions are not normally distributed, we’ll use the non-parametric test which is distribution independent, and make use of the median instead of mean.

**The task will be carried out in following steps:**

**Step-1:**

Set the alpha = 0.05

**Step-2:**

From following different types of non-parametric tests we’ll choose one which compares two samples samples distributions:

1. Mann-Whitney U Test
2. Wilcoxon Signed-Rank Test
3. Kruskal-Wallis H Test
4. Friedman Test

### we’ll consider Mann-Whitney U Test.

**Code:**

from scipy.stats import mannwhitneyu

def non\_parametric\_test(data1, data2, alpha):

stat, p = mannwhitneyu(data1, data2)

print('stat=%.3f, p=%.3f' % (stat, p))

if p > 0.05:

print('Both distributions have same mean')

else:

print('Both distributions means are different')

return pass

data\_without\_toll\_reimbursement = [80, 50, ……., …..,…………………………………….]

data\_with\_toll\_reimbursement = [ 20, 70, …….., ………, ………,………………………... ]

non\_parametric\_test(data\_without\_toll\_reimbursement, data\_with\_toll\_reimbursement, 0.05)

**Recommendations to the City operation team with caveats:**

Based on the statistical test, we’ll recommend the City operation team whether to implement toll reimbursement or not. If we observe the significant increase in mean drives’ number after implementation of the toll reimbursement, then recommendation will be to continue implementing it else stop it.

Here one we may remind operation team that the conclusion has been drawn from a sample of 100 drivers, hence we may like to carry out more experiments to confirm the conclusion.