Experiment design on drivers visiting an island with/without toll reimbursement (a hypothetical case study):

Introduction:

An island (hypothetically) in south-eastern part of New York state has been recently turned into an amusement hub hosting a huge adventure park for people of all ages. To enter the island, visitors need to cross a bridge paying toll. The island's governing council is looking forward to imposing a new program to boost the number of visitors into the island. One such strategic program the council is considering to start is 'reimbursement of toll on drivers' passing the bridge, which council believes, will attract more visitors creating more revenue for the island's development.

Here, let me consider myself (hypothetically) working on this program and is assigned to design an experiment to have a data driven decision. The outline of my experiment design will be as follows:

Statistical problem statement:

The implementation of toll reimbursement will increase the number of drivers visiting the island.

Experimental set up:

As the experiment involves hypothesis testing, we'll collect the data under two circumstances:
circumstance-1: **No toll reimbursement** when drivers visit the island.
circumstance-2: **Toll will be reimbursed** when drivers visit the island.

Evaluation metric:

Our evaluation metric will be 'increase in number of drivers because of toll reimbursement'.

An explanation to the evaluation metric is that, after implementation of toll reimbursement if the number of drivers visiting the island 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 the island decreases/remains same, then we'll consider our experiment to be not successful.

Statistical test:

Here, the evaluation metric depends on the statistical test on observations mentioned (above) under two circumstances. We'll formulate the **Null hypothesis** (null means nothing, no effect, no change, equal) and **Alternative hypothesis** as follows:

Null hypothesis (H_0): The averages of both distributions are same (ie number of drivers visiting the island does remain same even after implementation of toll reimbursement)

$$\mu_{\text{with toll reimbursement}} = \mu_{\text{with toll}}$$

Alternative hypothesis (H_1): The averages of both distributions are not same (i.e number of drivers visiting the island increases with implementation of toll reimbursement).

$$\mu_{\text{with toll reimbursement}} \neq \mu_{\text{with toll}}$$

For evaluation, one can take the significance level α to be 0.05 (or 0.01 etc), however, for this case, let us take the α value 0.05.

Data collection:

I decide the data collection duration to be 12 months; **first 6 months without reimbursing the toll** and **subsequent 6 months reimbursing toll**. I'll consider 10,000 drivers (units) as participants in my experiment, and if a driver visits the island, we'll label it as 1 else 0.

Table-1 will record the drivers visiting the island when there will be no toll reimbursement, whereas Table-2 will record the drivers for the period when toll will be reimbursed.

Table-1 (with toll)

Driver Serial Num.	**Day1**	**Day2**	**……..**	**……..**	**………**	**………**	**………**	**……..**	**Day180**
1			0						
2	1								
3						1			
….									
…				0					
….									
10000									

Table-2 (with toll reimbursement)

Driver Serial Num.	**Day1**	**Day2**	**……..**	**………**	**………**	**………..**	**………**	**………**	**Day180**
1				1					
2	1								
3									
….									
….	0			1					
….									
10000									

After data collection, we'll create another two tables (Table-3 and Table-4) for both periods summing the number of drivers visiting the island on each day.

Table-3: Num. of drivers visiting island on each day (with toll)

Day	Num. of drivers visiting island
Day1	5000
Day2	…..
……	2000
……	…..
……	
……	
Day180	8000

Table-4: Num. of drivers visiting island on each day (with toll reimbursement)

Day	Num. of drivers visiting island
Day1	3500
Day2	…..
……	7000
……	……
……	
……	
Day180	4500

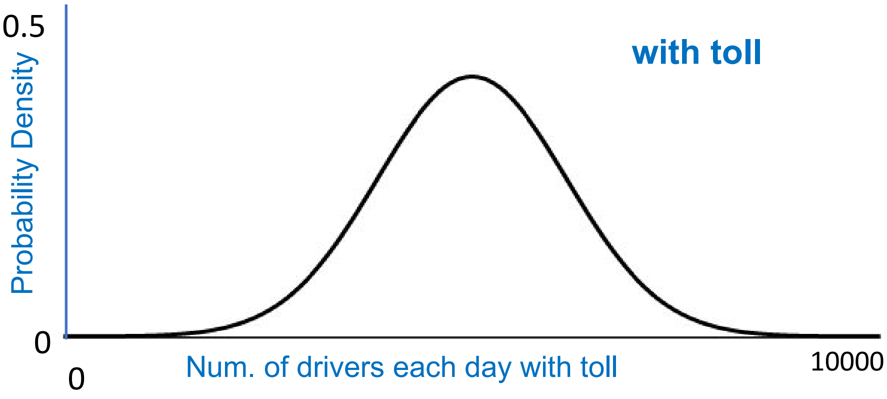
From Table-3 and 4, we'll make two arrays with num. of drivers visiting the island on each day (listing all for 180 days) as follows:

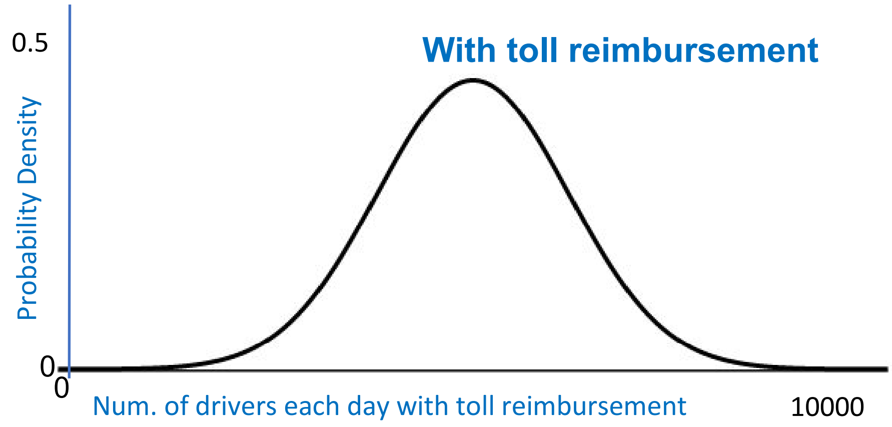
Num_of_drivers_each day_with_toll = [5000, …, 2000, …, 8000]

Num_of_drivers_each day_with_toll_reimbursement = [3500, …, 7000, …, …, …, 4500]

Statistical analysis:

First, we'll plot the distribution of '**Num. of drivers visiting the island on each day'** (along '**Probability Density of Num. of days'**) under both circumstances (with/without toll) and check visually both are normally distributed or not. In addition to that, Python's packages like statsmodels (or SciPy) built-in functions for normality tests can also be used.





Python code to test normality of each groups’ data distribution:

```
stats.normaltest (Num_of_drivers_each day_with_toll)

stats.normaltest (Num_of_drivers_each day_without_toll_reimbursement)
```

or

```
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 > alpha:
        print(“data  is normally distributed”)
    else:
        print(“data is not normally distributed”)
    return pass

normal_test(Num_of_drivers_each day_with_toll, 0.05)
normal_test (Num_of_drivers_each day_with_toll_reimbursement, 0.05)
```

Now, we’ll carry out the hypothesis test on the data pair, and the type of test we’ll choose taking into consideration the normality of the data. The test will be under any one the scenarios:

1. When both distributions are normally distributed:

If both distributions are normally distributed, then we'll carry out the two-sample test to see if they have same distribution or not by comparing their averages.

We'll set the statistical significance level: 0.05 (or 0.01)

In the distribution if the mean of case-2 (i.e with Toll reimbursement), the p value is < significance level, then we'll reject the null hypothesis (i.e means of both distributions are same) and accept the alternative hypothesis (i.e means of both distributions are not same and it has increased/decreased as per the difference in means). For this case study, let us choose a significance level of 0.05.

The task will be carried out in following steps:

Step-1

Calculate the following:

Num. of observations, mean, standard deviation of distribution (with toll): n_0, $\bar{x}_0$, s_0

Num. of observations, mean and standard deviation of distribution (with toll reimbursement): n_1, $\bar{x}_1$, s_1

Step-2

We'll use t-test for comparison and use the following two formulas to calculate the 'pool standard deviation of both distributions' (s_p) and t-value:

$$s_p = \sqrt{\frac{(n_0 - 1)s_0^2 + (n_1 - 1)s_1^2}{n_0 + n_1 - 2}}$$

$$t = \frac{\bar{x}_0 - \bar{x}_1}{s_p\sqrt{1 / n_0 + 1 / n_1}}.$$

Step-3

We can also calculate the t-value, and corresponding p-value using Python's built-in functions:

Python code:

```
p = scipy.stats.ttest_ind( Num_of_drivers_each day_with_toll, Num_of_drivers_each
day_with_toll_reimbursement,  equal_var = True/False)
```

when equal_var = True (performs a standard independent 2 sample test that assumes equal population variances)
equal_var = False (performs Welch’s t-test, which does not assume equal population variance)

Depending on the means and variances of both data, we need to carry out the test accordingly.

Here, if the output $p < 0.05$ then we’ll reject the null hypothesis and accept the alternate hypothesis.

2. When both distributions are not normally distributed:

If both distributions are not normally distributed, we’ll use the non-parametric test (also called distribution-free test). The non-parametric tests work well with skewed distributions that are better represented by the median (instead of mean).

The task will be carried out in following steps:

Step-1:

Set alpha = 0.05

Step-2:

There are many non-parametric tests to consider, out of which few are listed here:

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

Let us consider to use Mann-Whitney U Test for this case.

Python 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


non_parametric_test (Num_of_drivers_each day_with_toll, Num_of_drivers_each
day_without_toll_reimbursement, 0.05)
```

Recommendations for the island’s governing council:

Based on the outcome of hypothesis testing, I’ll recommend the island’s governing council. If I observe, there is a significant statistical difference between both means and an increase in drives’ number with reimbursement of toll, my recommendation will be to continue implementing reimbursement of toll, else withdrawing the toll reimbursement program.

In addition to the recommendation, I’ll request the council to propose for more set of experiments to have a concrete conclusion on this.