Answers for:

Part 2 - Experiment and metrics design

The neighboring cities of Gotham and Metropolis have complementary circadian rhythms: on weekdays, Ultimate Gotham is most active at night, and Ultimate Metropolis is most active during the day. On weekends, there is reasonable activity in both cities.

However, a toll bridge, with a two-way toll, between the two cities causes driver partners to tend to be exclusive to each city. The Ultimate managers of city operations for the two cities have proposed an experiment to encourage driver-partners to be available in both cities, by reimbursing all toll costs.

1. What would you choose as the key measure of success of this experiment in encouraging driver partners to serve both cities and why would you choose this metric?

2. Describe a practical experiment you would design to compare the effectiveness of the proposed change in relation to the key measure of success. Please provide details on:

    A. how you will implement the experiment
    B. what statistical test(s) you will conduct to verify the significance of the observation
    C. how you would interpret the results and provide recommendations to the city operations team along with any caveats.

Answers:

Question 1:
To choose a metric we need to first consider the goal of this experiment. We want to encourage inter-city service among drivers, and we will attempt this by reimbursing the drivers for the two-way bridge toll between the two cities.

Why do we believe this is beneficial?
We believe this is beneficial because of the complementary circadian rhythms of the two cities: during weekdays, one city gets busy at night while the other gets busy during the day. Encouraging drivers to cross city limits during busy times, instead of staying idle, means that the drivers will meet more of the demand of the riders in each city.

How can we measure the success of this experiment?
We can measure it by calculating the average number of rides completed during weekdays. We only include weekdays because during the weekends both cities have similar demand, which means inter-city service wouldn't have an effect.

What are the underlying assumptions of choosing this metric?
We are assuming no change in demand (the number of riders/customers), we are assuming no change in supply (the number of drivers), and we are also not taking profitability into account.

Question 2:
To implement our experiment we will begin by putting the toll reimbursement plan into effect. We then calculate the average number

of rides completed during weekdays for a number of weeks. We then state our null hypothesis as ‘The toll reimbursements did not affect the average number of rides during weekdays’. To test this statistically, we will calculate the average number of rides completed during weekdays for all of our past data up until we implemented the reimbursement plan. From these historical averages, we can then construct a distribution- which will be normally distributed according to the Central Limit Theorem. Using this normal distribution and our observed values after the implementation of the reimbursement plan, we can perform a one-sided p-value test to calculate the likelihood of these observations being observed given the null hypothesis. If the values are statistically significant we will reject the null hypothesis, if not, we will fail to reject the null hypothesis.

The p-value is usually taken as 5% by convention, so we will do the same.

Using all of our past data to construct the distribution is a conservative approach because this way we are taking into account past data when demand wasn’t similar and we’re also taking into account the first days of the company which must have been lower in demand and, we’re assuming, don’t reflect the current state of the company in general.

Since we are calculating the averages after implementing the reimbursement plan for a number of weeks, meaning we will have multiple observations, it is important to keep in mind that one of these observations could produce significant results just as a result of randomness. One way to guard against this is to declare the results significant only if the majority of the observations are significant.