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.

Part I: Defining a Key Measure of Success

1. We can tell where drivers are accepting and fulfilling trips by looking at the location stamps of their accepted and/or completed rides. Right now, because most drivers are staying exclusive to one city, we would expect most of their accepted rides to be linked to their primary city. If we looked at a ratio for each driver of n_rides_Gotham/n_rides_total over a given time period, we would expect most drivers to have a ratio near to 1 (if they primarily work in Gotham) or 0 (if they primarily work in Metropolis). If the toll-reimbursement initiative is effective for encouraging drivers to work in both cities, we would expect to see these ratios shift. Both cities have equal demand, on average, so a perfectly effective initiative would show a ratio of 0.5 between n_rides_Gotham/n_rides_total for every driver. By this logic, a meaningful measure would be the absolute value of the difference between 0.5 and the driver's "city ratio". The image below shows the derivations for this measure, which I'll call the "City Ratio Differential (CRD)."

On weekends, we don't need to see drivers moving back and forth between cities because both cities have equal demand at the same times of day; in fact, seeing a lot of movement back and forth could cost Ultimate Inc. a lot of unnecessary expense *if* the current behavior of drivers on the weekends is satisfactory. Additionally, if having a driver stay exclusive to one city on the weekends is acceptable behavior, including weekend trips could dilute the value of the CRD measure. Therefore, I'd propose to try this initiative solely for weekday driving as a starting place, and base all ratio calculations on weekday driving behavior.

$$G_{ratio} = \frac{n\ rides\ in\ Gotham}{total\ rides} = \frac{n_G}{n_{total}} \qquad M_{ratio} = \frac{n\ rides\ in\ Metropolis}{total\ rides} = \frac{n_M}{n_{total}}$$

$$G_{ratio} + M_{ratio} = \frac{n_G}{n_{total}} + \frac{n_M}{n_{total}} = \frac{n_G + n_M}{n_G + n_M} = 1$$

The City Ratio Differential measures how far a driver's behavior is from the ideal 50:50 weekday split. Though it only uses the G_{ratio} measure, it captures behavior in both cities.

$$City\ Ratio\ Differential\ (CRD) = |0.5 - G_{ratio}|$$

If the toll-reimbursement initiative has the desired effect of making driver partners available in both cities, we would expect to see the magnitude of the CRD decrease over an experimentation period.

Part II: Defining a Practical Experiment

This business case is a great candidate for an A/B Hypothesis Test, which will test to see if there is a statistical difference between the CRD (defined above) for 2 randomly assigned groups of equal size, where one group is the control group that will have no changes to their current conditions and one group is the test group with whom we'll implement the toll-reimbursement initiative.
Our test hypothesis would be: reimbursing driver-partners for toll fares on the bridge connecting Gotham and Metropolis will decrease their City Ratio Differential, that is, will improve driver-partner availability in both cities during weekdays.
The null hypothesis is that there is no difference in the CRD between the two groups (i.e. the initiative did not have the desired behavior); if we can reject the null hypothesis based on our data, then we can conclude that the initiative does have the desired effect on driver availability in both cities.

a) how to implement the experiment:
    i. determine the power of the test, the significance level of the test, and the minimum detectable effect:
        i. 80% is standard for power (that is, the probability of *not* rejecting the null hypothesis if there is indeed an effect),
        ii. 5% is standard for significance level (that is, the probability of rejecting the null hypothesis when there actually *isn't* an effect)
        iii. minimum detectable effect (MED) should be decided with stakeholders; how much of a difference in the CRD do we need to see to be substantial to Ultimate Inc.?
    ii. Based on the selections of power, significance level, and MED, calculate the necessary sample size and duration.
        i. Based on the calculated sample size, select users that proportionally represent all sectors of the driver partner population based on business insights: e.g. age, weekly drive time, type of car, reviews, etc. Stratify a split based on the determined business insights to create two identical groups of driver-partners: one group as the control and one as the test group.
        ii. Determine how long we want to run the test for *before* we begin the experiment. There are some formulaic guidelines we could look at, and in general we want to avoid too short a duration and too long a duration (to avoid novelty effects and maturation effects, respectively.)
    iii. Running the experiment: once the control and test groups of driver-partners have been assigned and a duration of the test has been determined, we can run the test by offering the toll-reimbursement modification to the test group only. We'll record the CRD of each driver in both groups over the duration of the experiment.

b) I'd conduct a one-tailed Z test on the CRD distributions of the control and test groups to determine the probability that the test group's CRD distribution was significantly less than the control group's CRD distribution by the predetermined MED level.
c) Based on the significance level of 5%, I'd reject the null hypothesis if the p-value from the above test was $p < 0.05$ and conclude that the new weekday toll-reimbursement initiative did result in driver partners being more available for rides in both cities during weekdays. If $p > 0.05$, then we fail to reject the null and can't make the conclusion that the toll initiative has an effect on driver availability in both cities.

i. In the case where the A/B Test shows a significant result, I'd report that result to the CityOps team. We'd conclude that reimbursing for toll costs does have the effect of encouraging drivers to be available in both cities. I'd also do some subsequent EDA on the driving behavior and Ultimate Inc.'s net revenue between the two groups to help ensure that implementing the change would make good business sense for Ultimate's bottom line.
ii. In the case where the A/B Test does not show a significant result, I'd report that the test was inconclusive. If this initiative was very important to CityOps, it might be reasonable to rethink the test parameters and conditions and try running a similar test again *with the caveat* that we can't simply run the exact same test again or continue the current test's duration in hopes of getting a different result. Such practices would not be statistically sound.