Jessica Wulzen
Ultimate Takehome, Question 2 - Experiment and metrics design

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?*

We'd like our metric as the key measure of success to be something that reflects Ultimate's best interest as a company for making this change. From the information in the question, I can think of two potential goals for Ultimate in making this change. The first potential goal is to improve customer experience by reducing wait times for rides - as drivers would have increased flexibility under the new system, there could be fewer situations where passengers in Gotham are stuck waiting for rides while drivers are sitting idle in Metropolis, or vice versa. The second potential goal is to improve driver efficiency, as drivers will be able to travel to the city with higher demand to give rides instead of sitting around in the city with lower demand.

Both of these goals seem feasible, and before carrying out further analysis, it would be important to discuss the company's motivation so that the findings can be most useful. In the context of this problem we cannot find out more information about the company, so we'll proceed considering each of these two goals as being potentially important to measure. For the goal of reducing rider wait time, the metric can be rider wait time for driver arrival. For the goal of improving driver efficiency, the metric can be percentage of driver time spent providing rides versus waiting for new rides.

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

For the purposes of this experiment, we'll make the assumption that it is trivial for Ultimate to measure and record the information necessary (rider wait time, driver activity time, etc). We'll use A/B testing to run these experiments, by giving a random sample of drivers reimbursements for bridge tolls while the rest of the drivers still have to pay. Then, we compare the data between the two groups. It is crucial to limit potential confounding factors. It is known that usage patterns are different between nighttime and daytime on weekdays and weekends, so we'll only compare between findings from the same timeframe. Thus, we'll compare weekend nights only with weekend nights, weekday daytimes only with weekday daytimes, and so on.

To evaluate the impact from our A/B testing, and to ensure that the observed difference is caused by our toll reimbursement and not random chance, we'll use t-testing. The data from our experiments should have the properties necessary for a t-test to be appropriate, and will allow us to determine the likelihood of seeing results as extreme as our findings.

From this experimental structure and testing, we'll be able to comment on if the proposed reimbursement plan leads to a statistically significant change in our metrics of choice.

However, in the communication with the city operations team, it is important to keep in mind that statistical significance is not the only aspect that matters. Reimbursing all drivers all toll payments will be expensive. In addition to the results from the t-tests, it will be important to communicate the improvement in our metrics in relation to the amount paid in reimbursements. For example, it is possible that weekend evenings and weekday evenings both cost the same for Ultimate in reimbursements, but that the metrics improve far more on weekdays than weekends. Ensuring that these details are reported will help make the most useful recommendations possible.