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.

This prompt, while clear, is missing one crucial piece of information: *why* do we want to encourage driver availability in both cities? This information is critical because it will determine which methods are effective toward achieving our *overall* goal, and which methods merely increase the "availability" metric without providing any real value to the company.

For example, one motivation could be to increase the uniformity of service and strengthen the brand image by unifying the standard of service offered in both cities, with the ultimate goal of increasing customer satisfaction and market share in both cities. Another motivation could be to reduce driver idle time and therefore reduce the number of drivers needed in all, thereby reducing driver training and management costs (without necessarily increasing market share at the same time).

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?

Without knowing anything about our higher-level goals, the most reasonable measure of success would be found in how well each driver's time spent in each city mirrors the actual demand in each city. For example, if Metropolis accounts for 70% of rides in the bi-city area and Gotham accounts for 30%, then a "fully-metropolitan" driver would be spending roughly 70% of their time in Metropolis and 30% in Gotham.

Many drivers may have their own reasons for not proactively seeking to serve different areas of the bi-city area, no matter how many incentives the company offers. Therefore, our standard of success should not be to make every driver "fully-metro."

Instead, we can establish a series of tiers for how metropolitan a driver might be, and seek reasonable increases in the higher tiers.

For example, a "fully-metro" driver's city ratio might be **±10%** of the 70/30 demand split. A "growing metro" driver's ratio might be **±20%, etc.** Naming a small number of tiers, showing progress up through those tiers, and giving them snappy names will help non-data people within the company understand, remember, and feel invested in these metrics.

If this variable is not appropriate for any reason, the next best choice is to record each instance of a driver crossing the bridge, and analyze the change in the frequency of this event under the new policy as a Poisson variable. However, I have declined to suggest this as our first choice because it does not capture the most important parts of driver behavior, and in some cases might work against our real goals. For example, drivers who are willing to cross over the bridge on one ride but will only accept rides back to their original city on the next ride will likely have the *highest* frequency of bridge-crossings, but this is not necessarily our desired behavior.

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

The best way to tell if a policy affects individual behavior is A/B testing, where some drivers get toll reimbursements and some don't for a certain period of time. As long as we randomly select a representative subset of drivers (in terms of where they are based, demographics, general driving habits, etc), then A/B testing will allow us to see in real-time whether and how the reimbursements affect driver behavior. One way to achieve this would be to pair up all our existing drivers into most-similar pairs (where each driver is paired with the driver most similar to them across the relevant metrics) and then randomly choose one driver from each pair to be in the B group, which receives reimbursements. This process would be opaque to the drivers; they would never know who they were "paired" with.

However, because this policy might be seen as a perk by the drivers, I would consult with the training and management departments first to see whether this would be seen as unfair by the drivers in the control group. If so, then we could do the slightly-less-rigorous limited-time test, where all drivers get toll reimbursements for a certain period of time. This will allow us to see whether driver behavior changes, but since there is no control group, we won't be able to say for certain how much of the behavioral changes are due to the new policy and how much are due to external factors that just happened to occur around the same time as the test.

If we split the drivers into A/B groups, we will compare each drivers' ratio of time spent in each city. If we put all drivers into the B group at once, then we can compare the each drivers' ratio of time in the two cities during the most recent period before the test as well as with year-on-year data. This will be our best way of controlling for factors unrelated to the experiment (seasons, holidays, etc).

b) what statistical test(s) you will conduct to verify the significance of the observation

We can verify the significance between the city ratios of the A and B conditions using a one-sided t-test on the difference in each distribution's absolute difference between the drivers' behavior and the desired behavior.

For example, if 70/30 is the desirable Metropolis/Gotham split, we can take each driver's Metropolis ratio, subtract 0.7, and take the absolute value to determine that driver's difference from the desired ratio. If our experiment was successful, this difference will be lower on average in the B group than in the A group. Thus we can perform a one-sided t-test on the distributions to determine our confidence level that the A group's differences are actually lower than the B group's differences.

c) how you would interpret the results and provide recommendations to the city operations team along with any caveats.

If the A and B distributions are significantly different, I would chart and package the following statistics for a presentation:

1. Charts of the two distributions (possibly with maps of each city overlaying each other like a Venn diagram to visually demonstrate how much more overlap there is in the B group than the A group.
2. Big numbers showing the increase in time that Gotham drivers spent in Metropolis, and vice versa.
3. Information on how many fares were paid, to give an idea of the ROI.
4. Hopefully we'd also be able to survey the drivers to see how they feel about the program, what they learned under the new policy, and what else might help them accept the desired driver behavior more.
5. Any other interesting discoveries made along the way.

I would be sure to include the following caveats:

1. If the results were significant, that does not necessarily mean that the behavior we saw in the test will continue when the policy is rolled out. There's still a low chance that we only observed random variation in the data. (I wouldn't put too find a point on this, to prevent our budget from getting slashed.)
2. Emphasize that we should continue to monitor driver behavior and feedback as drivers' behavior changes over time. After they become used to the policy, their behavior might change again.