LAB 3: Google Analytics
MSAN 631: Design and Analysis of Experiments

Arda Aysu, Mikaela Hoffman-Stapleton, Zefeng Zhang, Tim Zhou

June 19, 2017

Introduction

We designed a simple website containing short biographies of five current MSAN students: Arda Aysu, Samuel Haaf, Mikaela Hoffman-Stapleton, Zefeng Zhang, and Tim Zhou. The website consists of two pages. The first is interactive, and contains a banner of first names and sections for each of the students, where each section consists of a picture and clickable text that expands to show the student descriptions. There is a clickable link, which reads "RABBIT HOLE," located in the last position of the banner. This leads to the second page, which is not interactive, and just shows some interesting GIFs.

The purpose of this website is twofold: (1) to give a brief introduction to the five MSAN students, and (2) to create a website that measures user traffic and activity. We hope that users will be excited to learn previously unknown facts about the highlighted students, and hope that the website generally fosters a more open and communicative environment among current MSAN students. As users engage with the website, we will track their activity through the use of Google Analytics. Since the website essentially acts as a content publisher, we are most interested in encouraging engagement and frequent visitation. This informed our choices in the experimental setup described in the next section.

Experimental Design

We decided to design a multivariate experiment to determine whether there was a superior version of the website. While there are potentially multiple factors that influence user engagement, we were most curious to see whether changing whose section appeared first on the website had any effect on this factor. Given the limitation of varying only a single factor, this seemed like the most appropriate factor to vary since it is the most obvious difference we could have made beyond changing the overall stylistic theme of the website. Naturally, we had five different levels, each of which corresponded to a different student section appearing first. The remaining sections were kept in their original order.

Since we are interested in maximizing user engagement, we chose session duration as our response variable. This makes sense since session duration directly measures user engagement – we can

assume that the longer a user spends on the website, the more they read about each of the students and the longer they spent being mesmerized by the GIFs. We did, alternatively, consider using bounce rate as the response variable since this more directly characterizes a user's initial reaction to our varying factor. However, we were more interested in determining whether the entire content of the first section – not only the student's picture, but their description as well – played a role in overall user engagement. Assuming the user clicks to expand the first description, we theorized that perhaps a more intriguing first section would encourage them to move onto the next section. Our experimental setup is summarized as follows:

- response: session duration
- factor: student section that appears first on the website
- levels: Tim's section, Mikaela's section, Arda's section, Sam's section, and Zefeng's section

Implementation

The specific type of multivariate experiment we carry out is commonly known as the *multi-armed bandit*. This approach differs from the "traditional" A/B testing paradigm in how it determines the sample size for each variation. In a standard A/B test, each variation is assigned some fixed proportion of the overall sample population. This proportion remains fixed for the lifetime of the experiment.

The multi-armed bandit, on the other hand, *dynamically adjusts* the proportion of traffic to each variation as the experiment is running. This adjustment is performed based on a performance metric/KPI; if a variation performs better relative to this KPI than others, the multi-armed bandit will send accordingly more traffic to that variation.

Using a multi-armed bandit, rather than standard factorial design, presents its own advantages and disadvantages.

Advantages

- By redirecting traffic dynamically based on performance, the multi-armed bandit minimizes the effect of poor performing variations. This is very desirable for long-running experiments where the performance metric is tied to something of immediate value, such as revenue, or conversion rates.
- In traditional A/B testing, if a variation performs poorly (e.g. significantly less than control), it would still maintain the same proportion of the sample population throughout the experiment. If performance is tied to revenue, this could result in incurring a large amount of (lost) opportunity cost, just for the sake of maintaining the experiment.

Disadvantages

- By directing traffic to the better performing variation(s), the proportion directed to the poor performing variation(s) are reduced. This means it will take a longer time for the sample size in those variations to reach a size large enough for statistical significance, especially with a large number of arms and more severe Bonferroni corrections.

Statistical significance is desirable in poorly performing variations because we would want to know if the poor performance is actually due to the variation, and not just random noise. This disadvantage is irrelevant, however, if we only care about finding variations that perform *well* with statistical significance, and don't care too much about making Type I errors on the ones that seem to be performing poorly.

In the context of our experiment, a standard factorial design approach would first test if there is *any* difference at all between the variations. This is usually done using an ANOVA F-test or χ^2 test of independence. However, such tests can only be run *a posteriori* - that is, after we have collected our data! With a multi-armed bandit, we can *see* how the variations perform as the experiment is running (though not necessarily with statistical significance).

In a classical A/B test experiment, we would calculate the needed sample size for each variation, given a desired significance level and test power. From a practical perspective, we would enforce this by assigning incoming traffic to the variations with a probability proportionate to each variation's sample size; this proportion remains unchanged throughout the experiment. For this experiment, assuming a significance level of 5 / 1 = 1%, test power of 95%, 60 second session duration in control with 15 second standard deviation and an effect size of 20 seconds, we would need a sample size of 20 people per variation.

On the other hand, in a multi-armed bandit, we would choose initial proportions based on prior experience (or just start off with uniform proportions). However, as the variations start performing differently, we *adjust* the proportion of traffic going to each accordingly.

We ran the experiment for 4 days, at the end of which statistical significance had not yet been achieved. The results are shown in Figure 1.

Results and Conclusion

By default, Google Analytics ends an experiment when it is 95% certain that one arm performs better than control. According to the documentation, this is computed empirically by Monte Carlo simulation. As seen in Figure 1, this did not happen, so no conclusive winner was determined. Given the time frame of this project and the limited traffic, this is not surprising.

At the time when this screenshot was taken, the version with Sam's section appearing first seemed to be the most likely to reach the 95% threshold. However, we do not have enough data to say this with confidence; indeed, the sample size in three of the arms are in the single digits (less than the 20 needed for a pairwise test), and it is possible that given more time, they could end up outperforming the current winner.

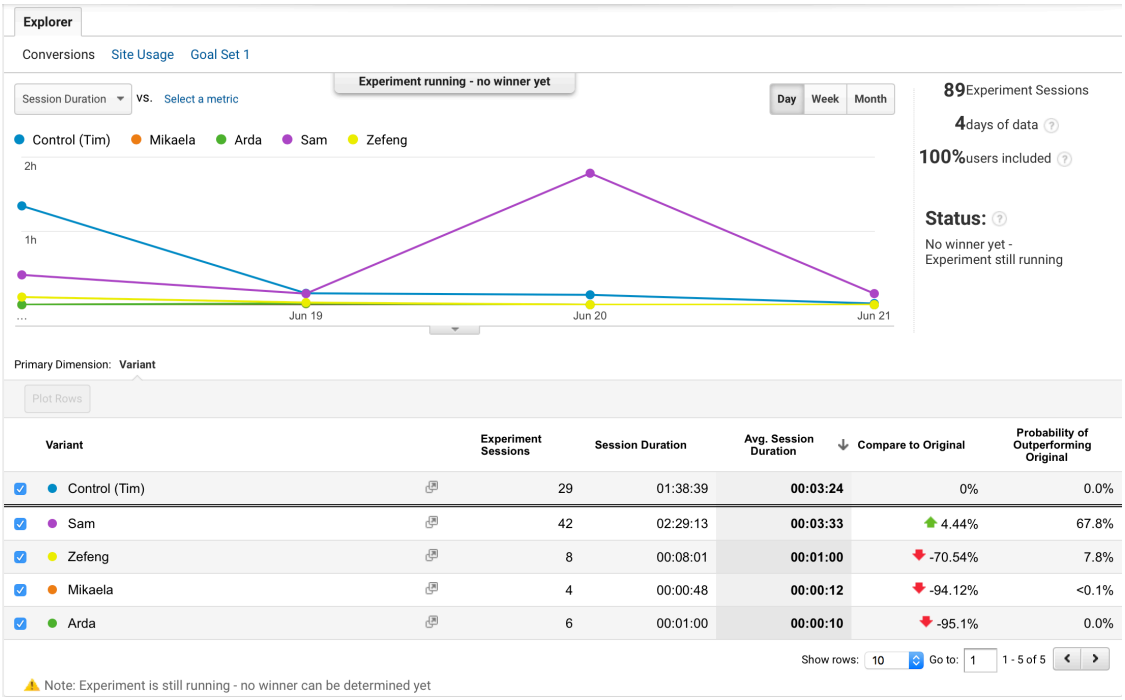

Figure 1: Screenshot of Google Analytics experiment results as of June 21, 2017 at 8:25 PM.

One important experimental flaw – which we only noticed after the fact – was that when some of us shared the website link on Facebook (and perhaps other social media feeds), there ended up being *two* distinct ways that users could access the website. One was through the actual blue highlighted link, and the other was by clicking the picture that appears with the link. If the user happened to click on the actual link, then they were randomly assigned one of the versions as intended by the experiment. However, if they clicked on the picture, they were sent the same version of the website that the user who posted it was assigned. This meant that there was a direct link to certain variations, rather than Google Analytics assigning the proportion according to the bandit arms. This resulted in increased traffic to that version, moreso than would have been naturally assigned by the bandit. This might explain the high number of experiment sessions associated with the "Tim" and "Sam" versions relative to the rest, and the higher probability corresponding to the "Sam" version. Ultimately, it is unknown whether this is actually the case, since we cannot be sure how users accessed the website. However, it is enough to invalidate any results – it is a good thing we did not come to any conclusions.

Additionally, we point out that throughout this experiment, some of us often checked the website to make sure that it was working properly (both for this class as well as Diane's class). This meant that our sessions were counted as data in the experiment. These data points do not actually contain any information, and potentially even threw off our results. Ideally, these sessions would be excluded in a formal analysis. This further supports our belief that our experiment did not deem any version as optimal.