Udacity Data Analyst Nanodegree A/B Testing Project

Arif Hikmet Onat Balta

Project instructions can be found here.

Experiment Design

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here.

Invariant metrics:

- **Number of cookies:** That is, number of unique cookies to view the course overview page.
- **Number of clicks:** That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger).
- **Click-through-probability:** That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.

Evaluation metrics:

- **Gross conversion:** That is, number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button.
- **Net conversion:** That is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button.

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

- **Number of cookies** is a good invariant metric because cookies are created before the "start free trial" clicks, so it will be the same in control and in experiment groups. As it won't change, this metric can not be considered as an evaluation metric.
- **Number of user-ids** is neither a good invariant metric nor a suitable evaluation metric because the number of users who enroll in the free trial is dependent on the test results and it can be different across control and experiment groups because of the default diversion.
- **Number of clicks** is a good invariant metric because it happens before the pop-up window therefore it won't change between control and experiment groups. It is not a good evaluation metric because of its unchanging condition.

- **Click-through-probability** is a good invariant metric because of the same reason explained above. It happens before the test, therefore it is independent from the test and won't change between control and experiment groups. It is not a good evaluation metric because it won't change during the test.
- **Gross conversion** is a good evaluation metric because it depends on the test results and shows the effects of the test by explaining whether asking students their available time has an effect on enrollments or not. It is not a good invariant metric because of its dependent structure.
- **Retention** is a good evaluation metric because it is dependent to test and not a good invariant metric because the number of enrolled and paid students are affected by the test results.
- **Net conversion** is also a good evaluation metric because it changes according to test results and explains whether asking students their available time reduces the number of frustrated students who left the free trial or not. Because of this it is not a good invariant metric.

In the following parts, I will use **Gross conversion** and **Net conversion** metrics for evaluation and try to investigate: "Would this test reduce the number of frustrated students who left the free trial because they didn't have enough time ?"

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics.

- Gross Conversion: 0.0202
- Net Conversion: 0.0156

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

In both metrics, the denominator is a cookie, which is also Udacity's unit of diversion, therefore variance can be analytically estimated. If any of the denominator was something different than a cookie, then we had to empirically estimate the variance.

Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately.

I didn't use Bonferroni correction during my analysis.

The required pageviews: 685275, which is the total number of pageviews for control and experiment groups.

Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment.

50% of Udacity's traffic would be diverted to this experiment and it would take 35 days to be completed.

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

Although it seems like a clear A/B test, any mistake or bug can prevent students enroll on courses, which will negatively affect Udacity. I decided to divert only a half of the traffic to this experiment so if something goes wrong, only one half of the traffic will notice it.

Experiment Analysis

Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check.

- Number of cookies:
  - It is expected to be in range [0.4988, 0.5012] with a 95% confidence level.
  - The actual observed value (0.5006) is inside the range.
  - This metric is statistically significant and passes the sanity check.

- Number of clicks on "Start free trial":
  - It is expected to be in range [0.4959, 0.5041] with a 95% confidence level.
  - The actual observed value (0.5005) is inside the range.
  - This metric is statistically significant and passes the sanity check.

- Click-through-probability on "Start free trial":
  - It is expected to be in range [0.0812, 0.0830] with a 95% confidence level.
  - The actual observed value (0.0822) is inside the range.
  - This metric is statistically significant and passes the sanity check.

Result Analysis

Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant.

- Gross conversion:
  - Difference between the experiment and control groups are expected to be in range [-0.0291, -0.0120] with a 95% confidence level.
  - Since the range is not including 0, difference is statistically significant
  - The difference is also practically significant since dmin = 1% is inside the range.
- Net conversion:
  - Difference between the experiment and control groups are expected to be in range [-0.0116, 0.0019] with a 95% confidence level.
  - Since the range is including 0, difference is not statistically significant
  - The difference is also not practically significant since dmin = 0.75% is not inside the range.

Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant.

- Gross conversion has a p-value of 0.0026, which is lower than the α level (0.05), therefore result is statistically significant.

- Net conversion has a p-value of 0.6776, which is higher than the α level (0.05), therefore result is not statistically significant.

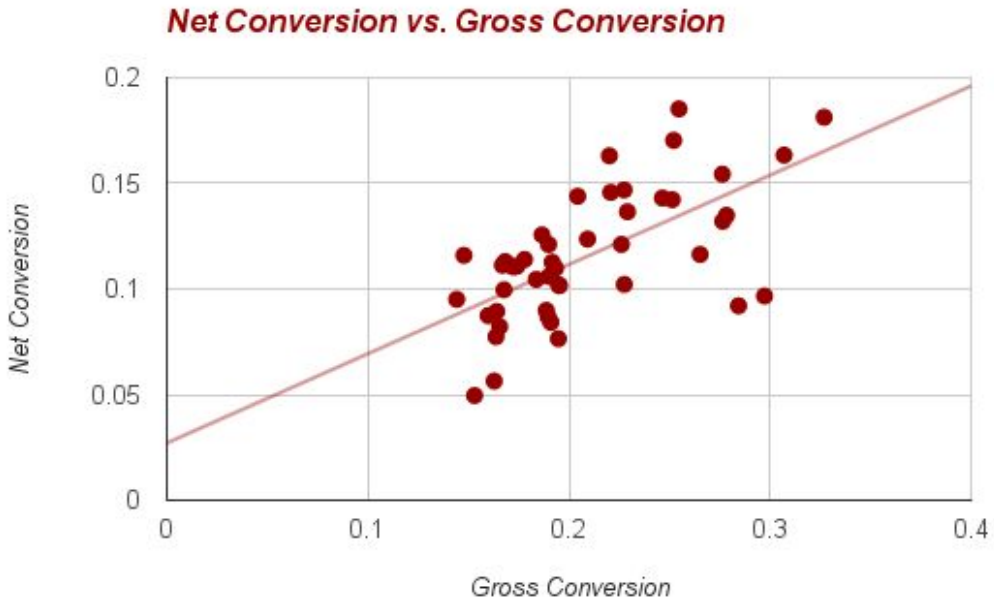
Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

I did not use the Bonferroni correction. The explanation is as below.

Selected evaluation metrics, gross conversion and net conversion, are dependent on each other. It can be also said that there is a positive correlation between these two metrics. Therefore, the metrics are expected to move together.

Net conversion versus gross conversion graph can be seen below.



As stated in Wikipedia: “The Bonferroni correction can be somewhat conservative if there are a large number of tests and/or the test statistics are positively correlated.”[1]. In other words, applying Bonferroni correction where it is unnecessary, like this case, may cause us to be overly conservative in our tests' conclusions and potentially lock us out of useful experimental outcomes [2].

Recommendation

Make a recommendation and briefly describe your reasoning.

The first evaluation metric, gross conversion, is both statistically and practically significant. The difference between experiment and control groups is negative, telling that there are less students enrolled in courses after seeing the pop-up window, which asks students their available time. This will help Udacity improve coaches' capacity to support students who are likely to complete the course. However, the second evaluation metric, net conversion, is neither statistically nor practically significant. This means there isn’t enough evidence to say that this test has an effect on Udacity’s revenue. In other words, we can say that this test has no effect on increasing or decreasing net conversion, which is highly related with the number of paying students, because the difference between experiment and control groups is not significant.

As a result, if Udacity’s only goal is to increase the likelihood of course completion, then this test can be carried out but if Udacity also wants to increase the number of paying students, then they might consider doing more follow-up experiments. For example, besides this test, Udacity can also test whether sending emails about job opportunities after graduation or student success stories to newly enrolled students will increase net conversion rate or not. So these two

experiments might help Udacity to decrease the number of frustrated students while increasing the number of paying students.

Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

In a computer based education, motivation is one of the key point. Each time you find a quiz difficult and struggling or each time you get a returning project, it is your motivation that forces you to continue. Thanks to Udacity's awesome resources, like discussion forum, webcasts and office hours, it becomes easier to overcome such difficulties. Therefore, I believe each Udacity student with a high motivation, can complete any course in Udacity. So my follow-up experiment idea is about increasing students motivation by sending regular motivational emails thus aiming to increase the number of paying students, which could also decrease the number of subscription cancellations.

Udacity could consider implementing a mailing program for the Nanodegree's. When a user joins the trial program, he or she will receive an email from Udacity, in which there could be a sincere welcome video from Udacity graduates and some motivating things like their thoughts about the course, how Udacity changed their lives, what are some good practices, useful tips to new students, how they achieved certain things and how to struggle with difficulties etc. Also, if possible, some contact info or social media account links of graduates could be added. Then, these kind of student success story emails with different graduates could be sent regularly.

New students that enroll in a Nanodegree will randomly assigned to either control or experiment group. In control group everything will remain same, where in experiment group users will receive regular motivational emails from Udacity and its graduates

For this experiment,

- Null hypothesis: "Sending motivational emails to new students will not change Retention rate."
- Unit of diversion: user-ids because the experiment is held after students enroll in free trial
- Invariant metric: number of user-ids. Creating an account (and getting an unique user-id) is before the enrollment process so we don't expect this to change between control and experiment groups.
- Evaluation metric: Retention, that is, number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout and is a good evaluation metric for this test because we want to test whether motivational emails increase the number of enrolled students that pay or not and this metric is explaining exactly this.