Udacity A/B Testing Experiment Analysis

By Long Nguyen

Experiment Design

Metric Choice

For invariant metrics, I chose number of cookies, number of clicks, and click-through-probability.

Number of cookies: this metric should be roughly equal between control and experiment group. It's because the unit of diversion before the user click to enroll. Since this is the unit of diversion, we should expect each cookie to have a 50/50 chance of being put into either control or experiment group. Therefore, this number should be equal across both groups.

Number of clicks: this metric also should not change. This is because before clicking the "Start free trial" button, both groups are exposed to the same course overview page.

Click-through-probability: this metric is invariant across both groups because it's defined as the number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page. Since the number of cookies and the number of clicks are invariant, we should expect this metric to also be invariant.

Number of user-ids: this is not a good invariant metric because user-ids is tracked after user click the "Start free trial" button and see the warning sign. Our goal is to see if the screener is successful in turning down non-committed students, thus the number of user-ids after clicking the "Start free trial" button is expected to be different between control and experiment group. This metric doesn't fit to be an evaluation metric either because it give us similar information to gross conversion, but gross conversion is better. Gross conversion helps normalize the experiment and control grou. For example, a higher number of user-ids in one group doesn't say much if we do not know how many actually clicks. Maybe a lot of users in experiment group just happen to have technical problem and can't click the "Start free trial" button. Gross conversion takes this into account and reduce the risk that one group might have substantially different click profile than the other.

For evaluation metrics, I chose gross conversion and net conversion.

Gross conversion: this makes good evaluation metric because the goal of the screener is to reduce the number of would-not-be-committed students. Thus, if the screener is successful, then students will turn away from paid courses and decide not to enroll as soon as they click the

“Start free trial” and see the warning sign, which makes the number of students who complete checkout after clicking different between control and experiment groups. If the hypothesis held true, we should expect to see significant difference in this metric between control and experiment groups.

Net conversion: this is also a good evaluation metric because another information that we want from the experiment is to see if the number of students who continue past the free trial is significantly affected by the screener at all. Thus, we want to see if the difference in probability of payment, given clicks, is large enough to be considered statistically and practically significant. Also, if our hypothesis is right, then there shouldn’t be a significant difference in this metric between two groups.

Retention is also a good metric. It’s similar to the net conversion, except it uses user-ids as unit of analysis. I excluded this metric because using it will require a very large number of page views, which eventually result in long experiment.

In order to launch the change, two conditions must be met: we must see a decrease which is statistically and practically significant between experiment and control’s gross conversion, and there shouldn’t be a decrease between two groups’ net conversion that is statistically or practically significant.

Measuring Standard Deviation

For gross conversion and net conversion, the estimated standard deviation for 5000 cookies is 0.0202 and 0.0156, respectively.

For both metrics, the unit of analysis is the same as the unit of diversion. In both metrics, the unit of analysis is the number of cookies that click while the unit of diversion is cookie. Because analytical and empirical variability tend to match when the unit of analysis and the unit of diversion match, even though the standard deviation above is estimated analytically, we can safely be certain that this variation is not much different from empirical variation.

Sizing

Number of Samples vs. Power

I won't use the Bonferroni correction in my analysis phase because it;'s not necessary (more details in the summary section). Also, without Bonferroni correction, the experiment needs less page views, and therefore require less traffic and duration. The number of clicks I will need to measure gross and net conversion is 25835 and 27413, respectively. I decided to use 27413 clicks, since this amount of clicks is enough to cover both the gross and net conversion metrics. Using the number of clicks, I then estimate the number of pageviews (based on baseline click-conversion probability), which is 685325.

Duration vs. Exposure

Given that Udacity doesn’t have any experiment running at the same time, I would recommend starting at 50% of traffic. This will substantially shorten the experiment time, while not compromising total user experience significantly. It also reduce the risk of system failure that may affect other users while the experiment is running.

With 50% of daily traffic, or 20000 pageviews per day, it will take approximately 35 days to collect 685325 pageviews needed for the experiment.

I think this experiment is not very risky:

Even though there exist privacy risk as the gross and net conversion metrics require the recording of user-ids (if the experiment fails because of technical reason, several users might not be able to access the site), this risk is present whether the experiment is run or not, not risks that increase as a function of any element of the experiment. Also, the risk of privacy data leakage is not the risk that introduced by the experiment itself, but is the risk that exists on any website that takes user-ids. Because of these reasons, I would say that there doesn’t seem to be inherently risky.

Experiment Analysis

Sanity Checks

Number of cookies:

- **Lower bound:** 0.4988
- **Upper bound:** 0.5012
- **Observed:** 0.5006
- **Pass:** Yes

Number of clicks on “Start free trial”:

- **Lower bound:** 0.4959
- **Upper bound:** 0.5041
- **Observed:** 0.5005
- **Pass:** Yes

Click-through-probability on “Start free trial”:

- **Lower bound:** -0.0013
- **Upper bound:** 0.0013
- **Observed:** -0.0001

- **Pass:** Yes

All the invariant metrics behave as expected.

Result Analysis

Effect Size Tests

Gross conversion:

- **Lower bound:** -0.0291
- **Upper bound:** -0.0120
- **Statistically significant:** Yes
- **Practically significant:** Yes

The range between the lower bound and upper bound of gross conversion doesn't include 0, which means that we are certain 95% of time that there is a difference between control and experiment gross conversion. Also, the least difference is 0.0108, which is larger than our required minimum difference of 0.0, indicates that the result is also practically significant.

Net conversion:

- **Lower bound:** -0.0116
- **Upper bound:** 0.0019
- **Statistically significant:** No
- **Practically significant:** No

The range of lower bound and upper bound of net conversion includes 0, which means that there is a possibility that within 95% of time, we can observe that the control and experiment net conversion is equal. This indicates that the difference is not statistically significant. However, the lower bound of the confidence interval is larger than the practical significance level, while the upper bound is smaller. This means that in 95% of time, we may see net conversion increase, but not that much (less than 0.75%); we may also see a decrease in net conversion, and this decrease is practically significant (larger than 0.75%).

Sign Tests

Gross conversion:

- **P-value:** 0.0026
- **Statistical significance:** Yes

Net conversion:

- **P-value:** 0.6776
- **Statistical significance:** No

Using the definition of success as when experiment gross conversion is smaller that control gross conversion, the sign test shows that the probability of observing the number of days that experiment gross conversion is less than control gross conversion, due to random chance, is less than 0.3% (less than our of 0.025), indicates that such observation is too rare to happen by chance. On the other hand, net conversion doesn’t pass the sign test, suggesting that the difference in the number of days that experiment net conversion is less than control net conversion could be due to random chance.

Summary

The primary purpose of using Bonferroni correction is to reduce the Type I errors, or false positives, when using multiple evaluation metrics. However, in this experiment, we do not depend on at least one metrics to be statistically significant to make our decision, thus using Bonferroni in this experiment is unnecessary, In other words, we expect to see only gross conversion metric to be statistically significant, and not net conversion. Assuming the metrics are independent of each other, the chance that one metric is statistically significant and the other is not is less than 5% (which is within our required alpha). This means that if we observe that gross conversion is statistically significant while net conversion is not, we can safely say that the fact that gross conversion is statistically significant is not due to random chance, and we can launch the change. If we observe that both metrics are statistically signficant, we wouldn't launch the change because we do not know if gross conversion (the metric we expect to change) or net conversion (the one we did not expect to change) happen to be statistically significant by accident. If we observe that net conversion is statistically significant while gross conversion is not, we wouldn't launch either because we are pretty certain that the fact that net conversion is statistically significant doesn't happen by chance ($p < 0.05$), and that the change would affect students' retention after the free trial. In brief, in order to launch, we only expect gross conversion to be statistically significant and not net conversion, that is, we expect only one out of three cases to happen. Therefore, we do not increase Type I error in our experiment, and Bonferroni correction is not necessary.

Recommendation

I would recommend conducting more test to see what’s the probability of seeing net conversion decrease. In our theory, we suspect that adding the screener will reduce the number students signing up for free trial, without significantly reduce the number of students to continue past the free trial and eventually complete the course. Form the result, we can see that the gross conversion (which the probability of enrollment, given clicks), substantially reduced in experiment group (more than 1%), and the difference is statistically significant. Also, even though there is a difference in our net conversion as well (which is the probability of payment, given enrollment), the difference is not statistically significant, which means that the difference could be due to random chance. However, for net conversion, the 95% confidence interval includes negative numbers, which means that a drop in students who make first payment after the free trial can occur (which is the case we don't want because it hurts profit). Nevertheless, this reduction in net conversion is practically significant (larger than our minimum of 0.75%). Because our 95% confidence interval for net conversion includes many scenarios: the metric doesn’t change, the metric slightly increase, and the metric decreases significantly, I suggest we conduct more tests to determine what is the likelihood that each of these scenarios will occur. If the likelihood that net conversion will decrease is larger than the probability of other possible outcomes, then I would suggest not launching the experiment, because it is more likely that launching the experiment would hurt Udacity’s revenue.

Follow-Up Experiment

In the above experiment, time-commitment is hypothesized as one of the major factors that lead to students dropping out of the program. One possible follow-up experiment that Udacity can perform is to see which other factors might affect the dropout possibility. A possible factor can be the convenience of the coaching system. Currently students have to set up appointments with coaches 24 hours in advance; however, this might cause some frustration for students. When a student tries to fit doing Udacity project in a tiny time slot (let’s say during lunch break), not being able to see the coach immediately can be a problem. This can potentially delay the project (which causes frustration), and students might not remember very well what their problem was when they meet up with the coaches (due to time long time period). Thus, Udacity can test changing the coaching system as a way to retain students.

One intervention feature that Udacity can adopt is allowing students to contact coaches on the go. Students who have quick questions for coaches can either call in or chat with coaches on multiple platforms. Considering that coaching resources are limited (there are many more students than coaches) the coaching system can be set up so that each student's request is screened, and quick request will be directed to coach immediately, while longer assignment help will be delivered through email. The system can also be set up by taking advantage of the number of students using the platform, where qualified students who have completed certain projects have the option to join the coaching system for small rewards. Again, a smart screener will forward suitable questions to these coach assistants, while more difficult ones will be directed immediately to coaches. These will help the coaching system capable of rapidly answer students' inquiry on the go, making the learning process much easier.

The hypothesis is that a more convenient coaching system will significantly reduce the number of dropouts after students sign up for paid courses. If the hypothesis is true, Udacity could potentially earn more money as well as referrals, since students who committed to the end of the course are happy students, and they are more likely to come back or refer the platform to other people.

This experiment will use the dropout probability, or the total number of user-ids that drops out during courses divided by the total number of user-ids enrolled in those courses and made the first payment. Since each course is different and each student is different, a course might be easier for one student than for the other. Thus, a particular course or student might have higher dropout probability than the rest. By using total, I hope to offset this effect and get an unbiased estimate of dropout probability. Another metric that can be useful is the rate of using coach, which is defined as the number of user-ids sign up for coach meeting the second time divided by the number of user-ids enrolled in course. If our hypothesis is true, we would see a drop in dropout probability and a rise in the rate of using coach.

The unit of diversion would be user-ids, since the student might be using the coach system several times, this will ensure consistent user experience in case the user decided to switch to another platform or clear their browser history. Also, it’s unlikely that a student would enroll in several paid courses at once, using user-id as unit of diversion would reduce the risk of disruptive experience in using the coaching system.

In brief, I believe that the experience would make a valuable follow-up on Udacity efforts to improve user experience and retention. I have experienced the coaching system myself, and I often end up doing long research on my own because the system do not offer that level of convenience that I need. I would love to see the Udacity carry out the experiment.