Udacity Project 7: Design an A/B Test

David Hey

Experiment Design

Metric Choice

Invariant Metrics:

- Number of cookies - since the experience of a user does not change before they view the course overview page, we expect around the same number of cookies to view it. The change is not revealed until after they click the "start free trial" button". A change in this metric could indicate something fishy with our experiment.
- Number of clicks - similarly we expect a relatively similar number of unique cookies to click the "start free trial" button, since this happens before the screener is revealed
- Click through probability - given that we expect the number of cookies and the number of clicks to remain consistent, logically we would also expect the click through probability (clicks / cookies) to also remain stable.

Evaluation Metrics:

- Gross conversion - since the change implemented intends to screen out individuals who may not be able to commit enough time to complete the course, we are looking to test if the number of user-ids that complete checkout and enroll in the free trial divided by the number of unique cookies is lower. This would mean people were deterred from even beginning the free trial and the screening worked well. In order to launch we will need to see a significantly lower gross conversion rate.
- Net conversion - screening out individuals less prone to committing the necessary time could also mean that the number of individuals who enroll after 14 days divided by the number of unique cookies that clicked the start free trial could change. In this experiment, it will be important that the net conversion does not decrease. If it did, this could lead to less paying customers in the long term. In order to launch we will need to confirm that there is not a significant decrease in net conversion.

Unused Metrics:

- User ID - The main reason it is not used as an invariant metric is that we expect the number of user IDs that enroll in the free trial to change when the screening is done. Although it could technically be used as an evaluation metric, but in the case of this experiment using a rate or probability metric as opposed to a sum or count is a better option. For example, we could see a significant increase in users signing up for the free trial, but could only be due to the fact that there was more traffic to the site (perhaps due to some seasonal effect) and more people were clicking the "start free trial" button. This would lead to a significant difference in the number of user ID’s but not in gross conversion.

- Retention - since the screening intends to deter individuals who may not be as willing to commit time to the course (and therefore are more willing to give up in the first 14 days), one would hope that screening out these individuals also increases the proportion of people who enroll in the free trial and actually end up paying after staying on for more than 14 days. However, this metric requires far more page views then both of the conversion metrics, and would cause the experiment to take far too long. Therefore, it is being excluded from the evaluation metrics.

Measuring Standard Deviation

Analytical Standard Deviations:

- Gross Conversion = 0.0202
- Net Conversion = 0.0156

If the unit of analysis and the unit of diversion in an evaluation metric are different, the empirical variability will tend to be higher than the analytical variability. However, when they are the same the analytical estimate can be used. This is due to the fact that many distributions operate under the assumption of independence, but when units are different there is some uncertainty and correlation that is introduced, because there could be multiple cookies that are in fact the same user-id.

For this experiment the unit of diversion is cookie, as was decided during the design phase of the experiment. Since the unit of analysis for both metrics is cookie as well, the empirical variability and analytical variability should be consistent. Therefore, there is no need to calculate an empirical variability for these metrics as well.

Sizing

Number of Samples vs. Power

No Bonferroni correction was used during the analysis phase, in which case the number of page views needed would be 685,325.

Duration vs. Exposure

Assuming 685,325 page views are required, if 90% of the traffic were diverted the experiment would take 20 days.

Ideally, less traffic would be diverted to the experiment just in case there was some sort of bug that it introduced. However, by diverting a higher fraction of traffic, a result with the confidence and power needed can be achieved more quickly, so the change can be rolled out (or not) in a more timely fashion. Luckily, the risk introduced by this experiment is relatively low, so we do not have to be too concerned about diverting this percentage of traffic. There is very little risk of anyone getting injured, and there is not any sensitive data being collected, in fact there is really no additional data being collected that was not already being collected. The main thing we would have to worry about this this proportion of traffic being diverted is a bug being introduced to the website, and hopefully this is something the Quality Assurance team would have found that before hand.

Experiment Analysis

Sanity Checks

Metric	Lower Bound	Upper Bound	Observed	Passes?
Number of cookies	0.4988	0.5012	0.5006	True
Number of clicks on “Start free trial”	0.4959	0.5041	0.5005	True
Click-through-probability on “Start free trial	0.0812	0.0830	0.0822	True

Result Analysis

Effect Size Tests

Metric	Lower Bound	Upper Bound	Statistical Significance	Practical Significance
Gross Conversion	-0.0291	-0.012	True	True
Net Conversion	-0.0116	0.0019	False	False

Sign Tests

Metric	P-Value	Statistical Significance
Gross Conversion	0.0026	True
Net Conversion	0.6776	False

Summary

I did not use a Bonferroni correction for the calculations. Although we are using two evaluation metrics to try and make a decision, we need all the results of both metrics to match our expectations to launch the change. Bonferroni corrections are good for avoiding Type I error (false positives), but in our case we are not as worried about this as much since we would need two false positives to incorrectly launch the change. As a counterpoint, if we only needed one of the metrics we were evaluating to meet our criteria, there would be a much higher risk that one false positive would lead us to launch a change in error, and it would be important to counteract this with the Bonferroni calculation.

There are no discrepancies between the effect size or sign tests for these two metrics, which will help simplify the interpretation of the results in the upcoming recommendation.

Recommendation

Based upon these results, I would not roll out the change to production. The primary reason for not launching is that the confidence interval for net conversion includes the negative practical significance (-0.75%). Practically, this means that there is a chance that the change is deployed and causes an undesirable decrease in net conversion that would be considered impactful. This is a risk that could possibly lead to less subscribers, and therefore should be avoided.

Even though the results of the gross conversion test showed that there would be 3.03%-1.08% reduction in this metric (a statistically and practically significant result), the change can not be rolled out because all the metrics needed to be confirmed by the test. Since such a positive change was seen in the gross conversion, it could make sense to try and rerun the experiment with more power, or go back to the drawing board and consider what factors could be leading to a reduction in subscribers. One reason could be that the screening is actually scaring off a very small proportion of users who would actually have ended up being paying subscribers.

Follow-Up Experiment

Based upon the results of this experiment, a logical next step would be to focus on how to reduce early cancellations (i.e. students giving up during the 14 day free trial, and cancelling their subscription). My guess would be that 14 days is not enough time for some people to make meaningful progress, in turn making them believe that they cannot learn anything meaningful with the paid subscription. However, if they were presented with the option to get their free trial extended 7 days when they go to cancel their account, that may give them the time they need to make meaningful progress and “get hooked”. The most obvious metric for examining whether or not the goal of reducing early cancellations was achieved would be retention (number of user-ids enrolled past the 14 day boundary / number of user-ids to complete checkout). In this case the metric would have to be altered slightly to make the comparison fair, and would be as follows: number of user ids enrolled past free trial period (14 or 21 days) / number of user ids to complete checkout.

Null Hypothesis: There is no difference in retention when the free trial period is 14 days, compared to when the free trial period is 21 days.

Alternative Hypothesis: There is an increase in retention when the trial period is changed to 21 days rather than 14 days.

For retention the unit of diversion is user-id and the unit of analysis is also user-id.

References

https://discussions.udacity.com/t/p7-empirical-variance-and-anlytical-variance/38868
https://graphpad.com/quickcalcs/binomial2/
https://onlinecourses.science.psu.edu/stat464/node/32
http://pandas.pydata.org/pandas-docs/stable/index.html
https://en.wikipedia.org/wiki/Bonferroni_correction