A/B Testing Final Project: Design and Analyze an A/B Test

Max Edwards – 12/2/15 - Data Analyst Nanodegree – Project #7

Experiment Design

Metric Choice

Number of cookies: This is a good invariant metric because it is independent of the experiment and therefore not affected by it. Therefore, there would be no difference between the control/experiment groups. This is not a good evaluation metric as it doesn't provide any information on if the user enrolled in the free trial and if they converted to paying subscriber.

Number of user-ids: This isn't a good invariant metric. The number of people who enroll is not independent of the experiment. The whole idea of adding the survey is to add the reality check prior to clicking free trial and the measure the free trial enrollment effect and the conversion to paying subscriber (user goes passed free trial period and makes one payment). Therefore, measuring the number of user-ids wouldn't be useful. This could be used as an evaluation metric because we could measure the difference between the number of enrollments in each group. However, this isn't the best evaluation metric as it isn't normalized (the amount of people who click free trial will likely be different in each group). Therefore, I choose not to use this as an evaluation metric.

Number of clicks: This is a good invariant metric as there wouldn't be anything different between the control and experiment groups. The difference starts after the "Start free trial" button is clicked. This is not a good evaluation metric as it doesn't provide any information on if the user enrolled in the free trial and if they converted to paying subscriber.

Click-through-probability: This is a good invariant metric as there wouldn't be anything different between the control and experiment groups. The difference starts after the "Start free trial" button is clicked. This is not a good evaluation metric as it doesn't provide any information on if the user enrolled in the free trial and if they converted to paying subscriber.

Gross conversion: This isn't a good invariant metric. This number of people of people who enroll is affected by the experiment. This would be a good evaluation metric because it measures if adding the survey decreased the amount of enrollments but did not decrease conversions to paying subscribers.

Retention: This isn't a good invariant metric. This number of people of people who enroll is affected by the experiment. This would be a good evaluation metric because it would allow for us to measure the outcome of the change. You would hope to find that there was no effect or a positive effect on retention by implementing this change. That is, the ratio of user-ids to subscribe compared to free trial enrollments to remain the same or increase as a result of implementing the "time investment check" survey.

Net conversion: This isn't a good invariant metric. This number of people of people who enroll is affected by the experiment. This would be a good evaluation metric because it would allow for us to measure the outcome of the change. You would hope to find that there was no effect or a positive effect on retention by implementing this change. That is, the ratio of user-ids to subscribe compared to the unique clicks on "Start free trial" to remain the same or increase as a result of implementing the "time investment check" survey.

Overall, I think Gross Conversion and Net Conversion are the best evaluation metrics for this experiment. Although Retention is a useful metric, it is comprised of net conversion divided by gross conversion so it is easier to examine each of these individually. I will look for the gross conversion to decrease more than the given d_{min} = 0.01 as the idea of the experiment is to temper a prospective student's expectations and have them re-evaluate if starting the free trial is really the right thing for them. I will look for net conversion to increase more than the given d_{min} = 0.0075. If it remains steady, it is still beneficial for Udacity as they have less students to manage (less students in free trial) but the same amount of subscribers (and thus revenue). If the net conversion increases, then there is the added bonus of additional subscribers (more revenue) and the additional teacher capacity. Ultimately, in order to consider launching the experimental feature, gross conversion and net conversion both need to surpass their given minimum differences.

For convenience, the below table is provided as a summary of the chosen metrics as discussed above.

Invariant Metrics	**Evaluation Metrics**
• Number of cookies • Number of clicks • Click-through-probability	• Gross conversion • Net conversion

Measuring Standard Deviation

The standard deviations for my chosen evaluation metrics are shown below.

Evaluation Metrics	**Standard Deviation**
Gross Conversion	0.0202
Net Conversion	0.0156

I would expect the empirical standard deviation and analytical standard deviation to be comparable if the unit diversion is the same as the unit of analysis. Net conversion and Gross Retention both use cookies and the unit of diversion is also a cookie. Therefore, these metrics each should have comparable analytical and empirical standard deviations.

Sizing

Number of Samples vs. Power

Bonferroni Correction. I opted not to use this in my analysis.

Gross Conversion. I used the calculator from Reference (1) with the following parameters to calculate sample size for one group:

- Baseline conversion rate = 20.625%
- Minimum Detectable Effect = 1%
- β = 0.20
- α = 0.05

The result was a sample size of 25,835 for one group. Therefore, I will need double that (51,760) for two groups (experiment and control). Given there is only an 8% chance that the user actually makes it down the funnel far enough and clicks the "Start Free Trial", I will need 645,875 pageviews (51,760 / 0.08) to achieve an adequate sample for this metric.

Net Conversion. **.** I used the calculator from Reference (1) with the following parameters to calculate sample size for one group:

- Baseline conversion rate = 10.9312%
- Minimum Detectable Effect = 0.75%
- β = 0.20
- α = 0.05

The result was a sample size of 27,413 for one group. Therefore, I will need double that (54,826) for two groups (experiment and control). Given there is only an 8% chance that the user actually makes it down the funnel far enough and clicks the "Start Free Trial", I will need 685,325 pageviews (54,826 / 0.08) to achieve an adequate sample for this metric. Also, given this is higher than the Gross Conversion pageviews needed, it is the metric driving the amount of pageviews needed.

Duration vs. Exposure

I will divert 60% of the traffic to the experiment which means there will be 24,000 pageviews per day in the experiment group. The 24,000 pageviews per day results in 29 total days needed to reach the requirement of 685,325 pageviews. This seems reasonable as it covers a decent amount of time in order to address user behavioral differences by day (i.e. weekday vs. weekend) but isn't too long to the point where the experiment duration is impractical.

In terms of risk, this experiment is rather moderate as current paying subscribers are not affected and the site functionality is the exact same with the exception of the time commitment question. There is the possibility that you could lose some of the users that would of signed up for the free trial if they were not asked the question of how much time they have available. Therefore, diverting 60% of the traffic to the experiment helps limit potential adverse impacts.

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. (These should be the answers from the "Sanity Checks" quiz.)

The following table provides the 95% confidence intervals for the invariant metrics. For each case, the observed value (p_exp) is between the LCL and UCL.

	Pageviews	Clicks	Click-through probability
p_cont	0.5000	0.5000	0.0821
p_exp	0.5006	0.5005	0.0822
SE	0.0006	0.0021	0.0005
m	0.0012	0.0041	0.0009
UCL	0.5012	0.5041	0.0830
LCL	0.4988	0.4959	0.0812
Pass?	YES	YES	YES

Result Analysis

Effect Size Tests

The results of the Effect Size Test are displayed in the table below.

	Gross Conversion	Net Conversion
p_cont	0.21887	0.11756
p_exp	0.19832	0.11269
d	(0.0206)	(0.0049)
SE_pool	0.00437	0.00343
m	0.0086	0.0067
UCL	(0.0120)	0.0019
LCL	(0.0291)	(0.0116)
Statistically significant?	YES	NO
dmin	0.01	0.0075
Practically significant?	YES	NO

Sign Tests

To perform a Sign Test for my evaluation metrics, I subtracted the control values from the experiment values for Gross Conversion and Net Conversion. For Gross conversion, there were 4 positive signs out of 23 observations. The probability of this happening by chance is 0.0026 (two-tailed p-value) and therefore we can conclude there is a statistical significant difference (given an alpha of 0.05). For Net conversion, there were 10 positive signs out of 23 observations. The probability of this happening by chance is 0.6776 (two-tailed p-value) and therefore we can conclude there is not a statistical significant difference (given an alpha of 0.05).

Digging into the sign test a little deeper, it was observed that the 4 gross conversion positive differences occurred consecutively between Thursday, October 28th and Friday, October 31st. This could have happened by chance or there could be something going on that increased gross conversion during this time frame. This would be something I'd investigate before proceeding.

Summary

I did not use Bonferroni's correction because I'm considering each of my evaluation metrics individually as this is a requirement of the hypothesis. In order to decide to launch this change I need both metrics to be significant. By making this decision I'm increasing the chance of a Type II error because I'm more likely to fail to reject the null hypothesis if I need both metrics to be significant.

My effect size hypothesis test for Gross Conversion resulted in a statistically significant difference and was practically significant. These results aligned with the sign test which also found the difference to be significantly different.

My effect size hypothesis test for Net Conversion was not statistically significant different and was not practically significant. These results aligned with the sign test which also found the no statistically significant difference.

Recommendation

Based on the results of the experiment, **I would not launch the change** to the website. Although there was a statistically and practically significant difference in Gross Conversion, the Net conversion confidence interval contained zero with a lower confidence of -0.0116. This actually exceeds the negative of the practical significance level for this metric (d_{min} = 0.075). In other words, there's a possibility that Net Conversion decreased to the point where it impacts the business negatively enough for managers to care – obviously not a good thing.

Overall, we observed the outcome that we expected where less people enrolled. However, the potential adverse impact to Net Conversion means it is not worth the risk to implement this change. It may be worth investigating other ways to test out the idea of surveying student's time prior to enrolling in the free trial in future experiments that don't have a possible negative impact on Net Conversion.

Follow-Up Experiment

To follow up the results from the previous experiment, I would want to design an experiment that examines a change post enrollment during the free trial period that improves student experience to increase the amount of subscribers. One feature that could be added to do this would to add a feature that pops after the first week of the trial asking if the student wants "pause" and restart the free trial in a week. The hypothesis is that this change will allow for frustrated students to pick the free trial back up at a more convenient time as they didn't realize the time investment and need to. When they come back, they would be ready to complete the free trial. This change is essentially an attempt to eliminate the time pressure of using the 14 day free trial period for those that didn't realize the amount of work required.

The unit of diversion for this experiment would be User ID as the group split is at the point in the funnel where an individual is already registered. To perform the test, each student would be randomly assigned to an experiment group or control group. The students assigned to the experiment group will receive a popup at the seven day period asking if they want to pause their free trial for a week and the control group will receive the baseline free trial treatment.

The invariant metric would be the User ID count as the number of people who enroll in the free trial shouldn't be any different between the control and the experiment groups. The evaluation metric would be retention (the number of User IDs to remain enrolled past the 14 day period

divided by the number of User IDs to enroll in the free trial period). The duration of the experiment would need to take into account that a portion of the experiment group users will be in the free trial period for 21 days (if they elected to pause).

In order to launch the experiment, a statistically significant difference in retention (positively) would need to be observed. If not statistically significant, a practically significant difference in retention would need to be observed at a threshold (min difference) set by Udacity where they feel it is worth it to implement this change.

References

1. http://www.evanmiller.org/ab-testing/sample-size.html
2. http://graphpad.com/quickcalcs/binomial1.cfm
3. https://en.wikipedia.org/wiki/Bonferroni_correction