A/B Testing Udacity's Free Trial Screener By James Gallagher

Experiment Design

In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course. If the student indicated 5 or more hours per week, they would be taken through the checkout process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, or access the course materials for free instead.

The unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

Launch criteria was that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial and eventually complete the course.

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

Invariant metrics:

Number of cookies : Since entering the homepage of site doesn't have any correlation with this experiment of free trial screener so it is invariant.

Number of clicks : It will be invariant metric because the clicks happen before the user sees the experiment, and are thus independent from it.

Click-through-probability : Similar to number of cookies and clicks, since the users have not seen the ***start free trial*** page before they decide the click on the button, the click through probability also is not dependent on the test being carried out.

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. Most appropriate evaluation metric because it is directly dependent on the effect of the experiment and allows us to show whether we managed to decrease the cost of enrollments that aren't likely to become paying customers.

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. This could be good evaluation metric as it follows the experiment and is quite affected by it.

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. This will also be affected by the experiment held as it is based on experiment held and attraction provided by it.

None metrics:

Number of userids: That is, number of users who enroll in the free trial. It is not a good invariant metric because the number of users who enroll in the free trial is dependent on the experiment and also not an ideal evaluation metric because the number of visitors may be different between the experiment and control groups, which would skew the results i.e number of user-ids or enrolled users can fluctuate a lot with respect to the number of cookies that clicks start free trial button on a given day.

Also it is a type of raw count which doesn't work in case of different sized control and experiment groups. Raw counts are not as straightforward. A raw count does not have a denominator and can't adjust or normalize to the different sizes, thus not a good metric for this experiment. Instead, the number of user- ids divided by the number of ***start free trial*** clicks, which is the gross conversion, is a better metric as it marginalizes variances in the empirical count of user-ids.

To launch the experiment we need :

(1) decrease enrollments by unprepared students
(2) without decreasing the number of students who complete the free trial and make payment.

Gross conversion addresses goal 1 to have a practically & statistically significant decrease in number of frustrated students who left the free trial because they didn't have enough time to save money and efforts wasted on them.

Net conversion and retention addresses to goal 2 to have statistically significant increase or not significantly reducing the number of students to continue past the free trial and eventually complete the course to increase revenue.

Conversion Net conversion is the number of users to remain enrolled past the 14-days boundary (and thus make at least one payment), divided by the number of unique cookies to click on the "start free trial" button. Since the number of payments can be affected by the experiment, net conversion can be as well. Therefore, I will use it as an evaluation metric. Since this metric is expected to differ between the control and experiment groups, I will not use it as an invariant metric. To launch the experiment, I expect it to be similar across the experiment and control groups. That would mean that the experiment did not affect the number of students who did enroll in the free trial, 1 with having enough time to finish the course successfully.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

To evaluate whether the analytical estimates of standard deviation are accurate and matches the empirical standard deviation, the unit of analysis and unit of diversion are compared for each evaluation metric. A Bernoulli distribution is assumed here with probability p and population N where the standard deviation is given by sqrt(p*(1-p)/N). Lets measure standard deviation of chosen evaluation metrics.

Gross conversion

p = 0.20625 (given) N
= 5000 * 0.08 = 400
std= sqrt(0.20625 * (1-0.20625) / 400) =**0.0202**

Retention

p = 0.53 (given)
N = 5000 * 0.08 * 0.20625 = 82.5
std = sqrt(0.53 * (1-0.53) / 82.5) =**0.0549**

Net conversion

p = 0.1093125(given)
N = 5000 * 0.08 = 400

std = sqrt(0.1093125 * (1-0.1093125) / 400) = **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.

Gross conversion and net conversion both have the number of cookies as their denominator, which is also our unit of diversion. We can therefore proceed using an analytical estimate of the variance.

Although in Retention the unit of analysis i.e number of users enrolled in the free trial is far away from the unit of diversion i.e number of cookies therefore it is expected that analytical and empirical variability could be different.

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. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

As the metrics used in this experiment are highly correlated, I decided against using the Bonferroni correction as it will be too conservative in the figures calculated. I used online calculator to generate the number of samples needed with alpha = 5% and 1-beta = 80%.

From which following information was obtained.

Gross conversion

Baseline conversion rate	d_min	Sample size needed	Number of pageviews needed
20.625%	1%	25,835	645,875

Retention

Baseline conversion rate	d_min	Sample size needed	Number of pageviews needed
53%	1%	39,115	4,741,212

Net conversion

Baseline conversion rate	d_min	Sample size needed	Number of pageviews needed
10.93125%	0.75%	27,413	685,325

I selected to proceed with are **Gross conversion** and **Net conversion** because to achieve sufficient pageviews i.e 4,741,212 for the **Retention** metric, it would take too long for an A/B test.

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. (These should be the answers from the "Choosing Duration and Exposure" quiz.)Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

Since required no of page views for net conversion is 685325 which quite large, I suggest to divert 100% or whole fraction of traffic to this experiment. In 100% diversion around 18 days would be needed to run the experiment.
It is not much risky as its just an additional information provided to user that nanodegree requires some efforts and can never change mind of the person who actually wants to enroll. Also 18 days time is not very long duration to conduct any experiment and also not much technical and financial support is required.
Since it is only an informative experiment it wouldn't cause any harm in security and duration criterias.
So risk is bearable for this experiment.

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.)

All the three invariant metrics passed the sanity checks with following 95% confidence interval:

Number of cookies:
CI- [.4988, .5012]
Observed .5006

Number of clicks on "Start free trial":
CI- [.4959, .5041]
Observed .5005

Click-through-probability on "Start free trial":
CI- [.0812, .0830]
Observed .0822

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. (These should be the answers from the "Effect Size Tests" quiz.)

Gross conversion:
CI- [-.0291, -.0120]

This metric is statistically significant as the interval does not include zero, and is practically significant as it also does not include the practical significance boundary.

Net conversion:
CI- [-.0116, .0019]

This metric is not statistically significant as it included zero, and therefore not practically significant either.

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. (These should be the answers from the "Sign Tests" quiz.)

Gross conversion: 0.0026, statistically significant
Net conversion: 0.6776, not statistically significant

So on 0.05 cutoff Gross conversion passed the Sign Test and Net Conversion fails using online calculator.

Summary

Bonferroni correction is used when multiple independent tests were performed simultaneously. One of multiple metrics will be falsely positive as the number of metrics increases. However, we would only launch if all evaluation metrics must show a significant change. In that case, there would be no need to use Bonferroni correction as to approve this experiment what was expected that both gross conversion and net conversion to be significant, so Bonferroni could be too much conservative that is why it is not used.

Based on the practically significance of the effective size and sign tests, gross conversion will decrease while net conversion will not be significantly impacted. So no discrepancies were found b/w the effect size tests and the sign tests.

Recommendation

Launch criteria is aligned with the goals :

(1) decrease enrollments by unprepared students (Gross conversion)
(2) without decreasing the number of students who complete the free trial and make payment. (Net conversion)
Gross conversion turned out to be statistically and practically significant. This is a good outcome because it lower costs by discouraging trial signups that are unlikely to convert i.e decrease enrollments by unprepared students.
Since Net conversion unfortunately ended up being statistically and practically insignificant and the confidence interval includes negative numbers, so launching that experiment is not a good idea as it is very risky to just depend upon reduction of free trials or decrease in Gross conversion.

Follow-Up Experiment

A potential follow-up experiment could be testing enroll now with a discount. Each nanodegree take a good part of a year on average and there is currently an offer of receiving 50% of tuition paid back if the program is completed within a year. It will be compelling to users who are already determined to take the course and ready to jump in directly.

The hypothesis is that by providing this direct enrollment with a discount for completion in a set time frame of 12 months. They will have an expectation set on what an average completion time frame looks like and work towards that.

Also this experiment will also provide students with the motivation to complete their nanodegree and reduce number of cancellations.

Following are components for this analysis: unit of diversion: cookie
This follow-up experiment should use cookie as the unit of diversion as experiment is shown to all cookies that enter the course page. This ensures that a cookies are not both in the control and experimental group.
invariant metric: number of cookies, click through probability, no of clicks
Since all these events happen before enrolling into the course therefore these don't effect i.e invariant metrics.
evaluation metric: Retention, gross conversion, net conversion
Finally we only want to reduce cancellation so best evaluations metrics would be all of the above mentioned as explained in the previous experiment.