A/B Testing Final Project Submission

by Frank Corrigan

Experiment Design

Metric Choice

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

Invariant metrics are metrics that shouldn't change across experiment or control groups and are used for sanity checking our populations. The following metrics were used as invariant metrics:

- Number of cookies defined as the number of unique cookies to view the course overview page. Since cookies are randomly assigned to a group, I expect the split between experiment and control to be roughly 50/50. Hence, we should see an equal number in each group.
- Number of clicks defined as the number of unique cookies to click "Start Free Trial" button. Similar to cookies, I expect clicks to be randomly assigned between the groups. As such, if our 'assign group algorithm' is working properly, we should see an equal(ish) number in each group.

Evaluation metrics are metrics that may change across experiment or control groups such as business metrics (revenue, # of users, or market share) or user metrics (# users not finishing course) and are used to draw conclusions about effect of the A/B test. The following metrics were used as evaluation metrics:

- Gross conversion 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. Immediately, this metric shows us at what rate the pop-up message was able to divert students that were not able to allocate appropriate time to studying. A difference in this metric is an initial look to see if the pop-up message had any effect at all. I would expect gross conversion to be lower in the experiment group.
- ~~Retention[1] defined as the 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. In the event, every user in the experiment group clicked "Start Free Trial", I would expect retention to be the same in both groups because the pop-up window failed to divert students without adequate time to dedicate. However, since I believe the pop-up window will divert some students, I expect that retention should be higher in the experiment group (thus agreeing with the hypothesis that this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial.~~
- Net conversion defined as the 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. I believe these numbers will be about the same in the end. Students that might have

[1] Originally, I wanted to use this as an evaluation metric. However, after I ran the calculations for number of required pageviews and saw that if I used retention the experiment would take nearly 4 months, I reconsidered.

been diverted will sign up, pay, but then drop out sometime after the first 14 days because they did not have the appropriate expectations. This means lower coaching bandwidth in that time period and wasted money on behalf of the student.

Two metrics that were not assigned to either group:

- Number of user-ids defined as number of users who enroll in the free trial. This definitely couldn't be an invariant metric since I do not know ahead of time how many users in each group will create a user-id. While it potentially can be used as an evaluation metric, the evaluation metrics chosen above give us more information about the likelihood of improving the overall student experience and improving coaches' capacity to support students who are likely to complete the course.
- Click through probability 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 this metric is derived from the # of cookies and the # of clicks (both invariant metrics) using this as an invariant metric would be redundant.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. 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.

Page Views:	40,000
Observations:	5,000

Calculate analytic estimate of standard deviation for each evaluation metric

	Probability	Unit of Analysis	Standard Error	Standard Dev
Gross Conversion	0.2063	unique cookies	0.0072	0.0202
Retention	0.5300	enrollments	0.0194	0.0549
Net Conversion	0.1093	unique cookies	0.0055	0.0156

If unit of diversion, in this case a cookie, and unit of analysis are different the empirically computed variability will be higher than the analytically computed variability. This difference is telling us something about the underlying distribution of the metric. In the table above, we see that the unit of analysis for gross conversion and net conversion is unique cookies to click the 'Start Free Trial' button' per day and the unit of analysis for retention is enrollments per day. As such, I would expect analytically and empirically calculated variability to be the same for gross and net conversion. For retention, it will be better to use the empirically computed variability.

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

Up until this point, I had committed to using gross conversion, retention, and net conversion as evaluation metrics. However, after calculating the number of pageviews required using statistical power of 80% and alpha of 5% combined with a 53% retention rate and 1% practical significance boundary, it was evident that with an average of 40,000 viewers per day the experiment would need to be run for nearly 4 months[2]. For this reason, I revised earlier assumptions and will use only gross and net conversion as evaluation metrics.

I will not use the Bonferroni correction during this analysis. I believe that gross conversion and net conversion will move together, in other words they are covariant. After significance tests are run, I will only implement the change if the null is rejected in both evaluation metrics.

Statistical Significance:	80%
Alpha:	5%

	Gross Margin	Net Conversion
Probability	20.63%	10.93%
Practical Significance	1.00%	0.75%
Sample Size Needed[3]	25,835	27,413
Page Views Needed	645,875	685,325
Days Needed	16	18

The table above shows that using a statistical significance level of 80% and alpha of 5%, the number of pageviews required per group is 322,937.5 or 685,325 total.

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.

I don't foresee any risk in this experiment for Udacity. If anything, I believe people will interpret it as the company looking out for the customer's well being. For this reason, I would divert 100% of users through this experiment (either into the control or experiment group) in order to obtain results as

[2] I had considered reducing the percentage of traffic diverted to ≈ 15% in order to keep this as an evaluation metric. However, at ≈ 15% I worried that the user set captured may not be representative of the population.
[3] Calculator used: http://www.evanmiller.org/ab-testing/sample-size.html

quickly as possible. At an average rate of 40,000 users per day, this experiment would take 685,325 / 40,000 = 17.13 days. Rounding up to the nearest day, the experiment would take 18 days.

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.

	Lower Bound	Upper Bound	Observed	Passes?
# of cookies	0.4988	0.5012	0.5006	yes
# of clicks on "Start Free Trial"	0.4959	0.5041	0.5005	yes

All sanity checks pass.

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.

Metric	Lower Bound	Upper Bound	Stat Significance?	Practical Significance?
Gross Conversion	-0.0291	-0.0120	yes	yes
Net Conversion	-0.0116	0.0018	no	no

While gross conversion is both statistically and practically significant, net conversion is neither statistically or practically significant.

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	Net Conversion
p-value[4]	0.0026	0.6776

[4] Calculator used: http://graphpad.com/quickcalcs/binomial1.cfm

Significant?	yes	no

In agreement with the effect size tests, while gross conversion is statistically significant, net conversion is not.

Summary

In multiple hypothesis testing, as we increase the number of hypotheses being tested, we also increase the likelihood of a rare event, and therefore, the likelihood of incorrectly rejecting a null hypothesis[5] (a false positive). In this situation, we have two evaluation metrics and hence two hypothesis tests related to our bipartite hypothesis that the screener will A) divert a portion of unprepared students that may become frustrated during the free trial and not complete the course B) without significantly reducing the number of students to continue past the free trial and eventually complete the course. In order to launch this change, our hypothesis demands that all metrics, both gross and net conversion, need to be relevant in order to launch. Since we are not testing each metric individually, the Bonferroni Correction (which reduces the risk of Type 1 errors) is not necessary.

I did not find any discrepancies between effect size hypothesis and the sign test.

Recommendation

Based on my logic in the metric section at the start of this analysis, the hypothesis for each evaluation metric were as follows:

Gross Conversion:

- H0: The metrics are the same
- H1: The metrics are different

Net Conversion:

- H0: The metrics are different
- H1: The metrics are the same

In the absence of the Bonferroni Correction, all null hypotheses must be rejected in order to declare a statistical effect is in order. Independently it would seem that my expectation that gross conversion would be different between groups was true (effect size test indicated both statistical and practical significance) and my expectation that net conversion would be the same between groups was false (effect size test indicated neither statistical or practical significance). If net conversion is not the same between groups, perhaps we are reducing the the number of students to continue past the free trial and eventually complete the course. This is problematic from a business perspective.

If we recall the confidence interval for net conversion, -0.0116 to 0.0018, and the practical significance boundary indicating a meaningful change for the business, 0.0075, we realize that the practical significance boundary sits inside the lower bound of our confidence interval. This tells me

[5] https://en.wikipedia.org/wiki/Bonferroni_correction

that while we have declared no statistical or practical significance, it is possible that net conversions went down by an amount that would matter to the business.

Alas, in order to launch this change, our hypothesis demands that all metrics, both gross and net conversion, need to be relevant in order to launch. According to both the effect size tests and sign tests results, this is not the case. Additionally, the fact that it is possible that this experiment has decreased net conversions will be concerning to the business. As such, the recommendation is not to launch this change.

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.

If a gentle screener does not work to deter unprepared students from starting the free trial and consuming coaching resources, perhaps the screener needs to be more direct and/or more aggressive. I would run an experiment that splits cookies that click 'Start Free Trial' into two groups - one that can immediately create a user-id and start the free trial (control group) and one that is challenged to take a short prerequisite course (similar to what Udacity does for Nano-degrees) as the experiment group. I would let the user know that in order to *better prepare them for the course*, you'd like them to spend 20 minutes working through some prerequisite material (again, this makes it seem like you are looking out for the student).

My own experience when signing up for the Data Analyst Nanodegree was comparable. I started going through the 'Are you prepared section' and after the first few questions, stopped and took a few weeks going through the Udacity's Intro Statistics courses and brushing up on my Python. When I eventually signed up for the Nanodegree, I was fully prepared to succeed.

My hypothesis would be that the screener would divert unprepared students (pointing them to material they could review to become prepared) and improving coaches' capacity to support students who are likely to complete the course.

While, again, I would capture gross conversion, retention, and net conversion I would focus on retention as the evaluation metric. I would expect retention to be significantly higher for students that signed up for the free trial after they took the prerequisite course (experiment group retention > control group retention). This would tell me if this screener is helping us to retain more students (increasing payments) and decreasing pressure on coaching resources. In order to complete this A/B test in a timely manner, I would divert only 15% of Udacity traffic through this screener.