Unit 7: A/B Testing - Final Project
By Iwan Thomas

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:

- Number of cookies
- Number of clicks:
- Click-through-probability:

Number of cookies, number of clicks and click-through-probability are appropriate as invariant metrics and inappropriate as evaluation metrics because they are measured before any intervention. Control and experiment groups should be randomly assigned and the invariant metrics in both groups should be approximately equal. These metrics will provide good sanity checks.

Evaluation Metrics:

- Gross Conversion: Useful for measuring any difference in the proportion of students who enroll in the free trial when using the control and experiment page.
- Net Conversion: For the entire funnel in the experiment and control groups, what is the proportion of students enrolled beyond the two-week free trial in the control and experiment groups

Gross Conversion and Net Conversion are appropriate as evaluation metrics and inappropriate as invariant metrics because they are measured after the intervention.These metrics will be useful in assessing the impact of the experiment.

The aim of the experiment is to see whether setting clearer expectations for the students up front results in fewer students leaving during the free trial. Retention would be a valuable metric to measure this. Unfortunately, using the variable retention would require an excessively large number of pageviews (over 2 million) and result in the experiment lasting too long. This is why it has not been selected as an evaluation metric.

The hypothesis of the experiment is that implementing this change could reduce the number of enrollments by unprepared students who leave the free trial, without significantly reducing the number of students who continue past the free trial. The aim therefore is to reduce gross conversion without significantly reducing the net conversion. In analysing the results of the experiment, I will be looking to see a statistically and practically significant decrease in gross conversion **and** net conversion must not decrease.

Metrics that weren't used:

- Number of user-ids who enroll in the free trial:
    - Not appropriate as an invariant as the experimental change might cause a change in the number who enroll
    - Not appropriate as an evaluation metric as the experiment and control group sizes could be different.
        - As the metric is not normalised for the group size, small changes in the experiment and control group sizes, which are to be expected, will not be accounted for. Therefore, although a higher proportion of users might enroll in the free trial in the experiment group, the number of user-ids might be higher in the control group because of the difference in group sizes. This could be misleading.

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.

Evaluation Metric	Standard Deviation
Gross Conversion	0.0202
Net Conversion	0.0156

The key consideration when determining whether the analytic estimate would be comparable to the empirical variability is whether or not the samples are likely to be independent. Samples tend to be independent when the unit of diversion and the unit of analysis match. The unit of diversion for the experiment is cookies and, for both gross conversion and net conversion, the unit of analysis is cookies. Therefore, the analytic estimate is likely to be comparable to the empirical 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 you experiment appropriately.

I did not use the Bonferroni correction during the analysis phase.

I used the following website to calculate the number of pageviews needed to power the experiment,

http://www.evanmiller.org/ab-testing/sample-size.html

and the following parameters:

- alpha = 0.05
- beta = 0.2
- The baseline conversion rate provided in the “Final Project Baseline Values” spreadsheet
- And the minimum detectable effect as specified for each evaluation metric.

I computed the necessary pageviews for both evaluation metrics and chose the result that yielded the largest number of pageviews.

The calculator linked above returned a sample size of 39,115. The unit of analysis for net conversion is the viewers who click the “Start Free Trial” button, 0.08 of the unique cookies to view the page. Accounting for this funneling, and multiplying the result by two to account for the experiment and control pages, the number of pageviews required to power the experiment was computed as 685,325.

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.

Typically an experiment is considered to be high risk if it could cause harm to somebody or if it collects sensitive information. This experiment is doing neither and can therefore be considered to be low risk. Although there is a financial risk to this experiment, this is being monitored and the risk has been mitigated.

In light of all this and the client’s desire to see this experiment completed in a few weeks, I will divert all of the traffic to this experiment, resulting in a duration of 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.

Invariant Metric	Lower Bound of CI	Upper Bound of CI	Observed Value	Passes Sanity Check?
Number of cookies	0.4988	0.5012	.5006	Yes
Number of clicks	0.4959	0.5041	0.5005	Yes
Click-through-probability	-0.0013	0.0013	0.0001	Yes

As all invariant metrics have passed the sanity check, I can proceed to analyse the results of the A/B Test.

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.

Evaluation Metric	Lower Bound of CI	Upper Bound of CI	Statistical Significance	Practical Significance
Gross Conversion	-0.0291	-0.0120	Yes	Yes
Net Conversion	-0.0116	0.0019	No	No

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.

Evaluation Metric	p-value	Statistical Significance
Gross Conversion	0.0026	Yes
Net Conversion	0.6776	No

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

I did not use the Bonferroni correction as it is excessively conservative. My criteria for launching the experiment is a statistically and practically significant decrease in gross conversion and no significant change in net conversion. Both metrics must be satisfied to launch the experiment. As such, false negatives will have the greatest impact, as one false negative could cause the experiment to not be launched. The Bonferroni correction tends to reduce the likelihood of type I errors (false positives) at the expense of increased type II errors (false negatives). As false negatives will have the greatest impact in our case, we do not want to increase their likelihood and will therefore choose not to use the Bonferroni correction.

As a matter of interest, I calculated the confidence interval for gross conversion using the Bonferronni-derived alpha value of 2.5 to be:

[-0.0303, -0.0108]

The result is still statistically and practically significant at alpha = 2.5.

The results of the effect size hypothesis tests and the sign tests are in agreement.

Recommendation

Make a recommendation and briefly describe your reasoning.

I would recommend not launching the change.

There is evidence to suggest that the change led to a statistically significant drop in gross conversion, the probability of a user signing up for the free trial. There is evidence that the net conversion decreases, although this decrease is not statistically nor practically significant.

The confidence interval for the net conversion is skewed negative, meaning that the true value is likely to be negative. This means that if this change were launched, there would be a chance of a decrease in paying customers. This is not desirable for the client and therefore the change should not be launched.

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.

Another possible reason why a student might cancel early in the course is that a student does not meet course prerequisites.

In the experiment, if the student clicked “Start Free Trial”, he/she would be asked to take part in a placement test in order to identify whether he/she met the course prerequisites. Another option would be for the student to self-assess. If the student scored above a certain threshold on this test, they would proceed to the checkout process as usual. If however the student failed to reach a certain standard on this test, a message would appear indicating that this course requires more background knowledge, and suggesting a number of other free courses that the student might like to complete first. At this point, the student would have the option to continue to enroll in the free trial, or access the preparatory material instead.

The hypothesis is that by testing for prerequisites upfront, the number of students disheartened by difficulty of course is reduced resulting in a lower attrition rate. What’s more, the student ultimately gets a better learning experience by first addressing the basics before moving on to the more advanced material. This helps Udacity achieve its goal of educating and empowering its students!

The metrics I would want to measure are similar to those used in the A/B Test discussed in this report.

Invariant Metrics

- Number of cookies: The number of unique cookies to view the course overview page
- Number of clicks: The number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger)
- Click-through-probability: The number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page

Evaluation Metrics

- Gross conversion: The 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.
  - I would expect this to drop for the experiment group
- Retention: 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.
  - I would expect this to increase for the experiment group as it has been screened.
- Net conversion: 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. (dmin= 0.0075)

- I would expect this to increase for the experiment group.

My unit of diversion would be a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward.