# 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.
- 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.
- 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. However, it is difficult to tell whether this would be a positive or negative change, since we would expect the number of people that click the “start free trial” button would be the same in both cases, however the total number of people to enroll could be unchanged (even though retention is higher). If net conversion is the same but retention is higher, this may mean that the people who ended up enrolling, could have ended up enrolling anyways (regardless of the screening).

## Measuring Standard Deviation

Analytical Standard Deviations:

- Gross Conversion = 0.0202
- Retention = 0.0549
- Net Conversion = 0.0156

If the unit of analysis (denominator) and the unit of diversion (numerator) 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.

Of the evaluation metrics chosen, net conversion and gross conversion both have different units of diversion (cookie) and analysis (user-id). Therefore, it would be best to compute the empirical variability for these metrics.

## 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.

# 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 |
| Retention | 0.0081 | 0.0541 | True | False |
| Net Conversion | -0.0116 | 0.0019 | False | False |

### Sign Tests

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

### Summary

I did not use a Bonferroni correction for any of the calculations. Although using the correction could have compensated for the probability of getting a false positive being slightly higher when doing hypothesis testing with three metrics, it also has the parallel effect of increasing the probability of false negatives. Since we are only using 3 evaluation metrics, and we want to maintain the statistical power as much as possible for the time being I did not apply the correction.

Interestingly, the results for practical significance of the effect size tests match the statistical significance results of the sign tests. This leaves the only difference to be the statistical significance of the effect size test for retention. Although the effect size for retention was statistically significant it was was not practically significant, and it was not statistically significant for the sign test. This could be due to the fact that the variability of the retention metric was so much higher that the other metrics (its 95% confidence interval had a range of of 4.6%, while the next highest range was Gross Conversion with 2.79%). This meant that although the confidence interval did not include 0, it was too wide not to include the practical significance boundary, and there was a greater chance that the metric was not consistently higher in the experiment than the control on a day to day basis.

## Recommendation

Based upon these results, I would roll out the change to production. Although all 3 metrics did not exhibit significant results, the change had a statistically and practically significant effect on gross conversion. In the real world, this means that although the change may not affect the proportion of individuals coming to the site that end up paying for coaching, the proportion of people who sign up for a free trial is actually lowered, which will save coaching resources from investing time in providing services to non-paying customers (who likely will never pay). This is quite positive, because there actually could have been a situation where the proportion of individuals signing up for

the free trial was lower (good), and therefore the proportion of people who end up paying for the service is also lower (bad). Ideally there would have also been a significant increase in the proportion of paying customers (retention and net conversion), but a lack of significant change in these metrics is actually ok when paired with a significant decrease in gross conversion.

## Follow-Up Experiment

Based upon the results of this experiment, I would want to focus on how to improve retention and gross conversion, so that those who do enter the free-trail are converted into paying customers more often. To do this, I would offer 21 free days (instead of 14) to see if students become more attached to the class after a longer period of time with “no strings attached”. More specifically, I would look a significant positive change in retention.

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.

To ensure there are not any unintended consequences of this change I would also want to keep an eye on the gross conversion. It would be unfortunate to enable this change, and undo the effect that the last change had of reducing the gross conversion rate.