# A/B Testing Udacity's Free Trial Screener

Casimir COMPAORE

## Experiment Design: A/B Test

### Metric Choice

List of metrics I will use as invariant metrics and evaluation metrics here:

**Invariant metrics:** Number of cookies, Number of clicks, Click-through-probability

**Evaluation metrics:** Gross conversion, Retention, Net conversion

Explanations of why the metric were or were not chosen:

- Number of cookies: Good invariant metric as a cookie is created even before a user can see the experiment.
- Number of user-ids: Not a good invariant metric as the number of users enrolled in the free trial is dependent on the experiment. Not a good evaluation metric because the number of visitors may be different between the experiment and control groups, which would skew the results.
- Number of clicks: Good invariant metric as the clicks happen before the user sees the experiment, and are thus independent from it.
- Click-through-probability: Good invariant metric as the clicks happen before the user sees the experiment, and are thus independent from it.
- Gross conversion: Not a good invariant metric as the number of users who enroll in the free trial is dependent on the experiment. But it is a good evaluation metric because it is directly dependent on the effect of the experiment and it can be used as an evaluation metric to check if the experiment makes a significant difference in the enrolment.
- Retention: Not a good invariant metric as the number of users who enroll in the free trial is dependent on the experiment. But it is a good evaluation metric because it is directly dependent on the effect of the experiment, and it can be used as an evaluation metric to check if the experiment makes a significant difference in the financial outcome.

- Net conversion: Not a good invariant metric as the number of users who enroll in the free trial is dependent on the experiment. But it is a good evaluation metric because it is directly dependent on the effect of the experiment, and also shows positive financial outcome of the change.

I will consider the evaluation metrics Gross conversion and Net conversion. The metric Gross conversion will indicate whether the cost will be lower by introducing the new screener and the metric Net conversion will indicate how the change affects our revenues.

The results they were looking for to launch the experiment:

To launch the experiment, it will be required that the Gross conversion evaluation metric has a significant decrease, and the Net conversion evaluation metric has a statistically significant increase.

## Measuring Standard Deviation

List the standard deviation of each of the evaluation metrics:

**Gross conversion: se** = sqrt(0.20625*(1-0.20625)/3200) = 0.00715 (correspond to 3200 clicks & 40000 pageviews). For 50000 pageviews, se = 0.00715 * sqrt(40000/5000) = **0.0202**
**Retention: se =** sqrt((0.53*(1-0.53)/660) * sqrt(40000/5000))) = 0.0549
**Net conversion: se** = sqrt(0.1093125*(1-0.1093125)/3200) = 0.0055159 (correspond to 3200 clicks & 40000 pageviews). For 50000 pageviews, se = 0.00715 * sqrt(40000/5000) = **0.0156**

**Gros conversion: 0.0202**
**Retention: 0.0549**
**Net conversion: 0.0156**

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.

For Retention, the denominator is "Number of users enrolled the courseware" which is not similar as Unit of Diversion. The unit of analysis and the unit of diversion are not the same therefore the analytical and the empirical estimates are 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:

I did not use Bonferroni correction as the metrics in the test has high correlation and the Bonferroni correction will be too conservative.

Probability of enrolling, given click:
20.625% base conversion rate, 1% min d.
Samples needed: 25,835

Probability of payment, given click:
10.93125% base conversion rate, 0.75% min d.
Samples needed: 27,413 (chosen)

Therefore, pageview/group = 27413/0.08 = 342662.5
Total pageview = 342662.5*2 = 685325

I will need 685 324 pageviews to power the experiment with these metrics. That is, double (control + experiment groups) of the number of samples required for the more demanding Net conversion metric.

### 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:

With daily traffic of 40000, I'd direct 70% of my traffic (28000) to the experiment, which means it would take us approximately 25 days (685325/28000 = 25) for the experiment.

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

The experiment is not extremely risky as it does not affect existing paying customers. Nevertheless it may have a substantial impact on new enrollments, and diverting 100% of the traffic may thus not be advisable as some bugs may occur.

# 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:

Number of cookies: [.4988, .5012]; observed .5006; PASS

Number of clicks on “Start free trial”: [.4959, .5041]; observed .5005; PASS

Click-through-probability on “Start free trial”: [.0812, .0830]; observed .0822; 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:

Gross conversion: [-.0291, -.0120], statistically significant, practically significant

Net conversion: [-.0116, .0019], not statistically significant, not 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: .0026, statistically significant

Net conversion: .6776, not statistically significant

### 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 a Bonferroni correction because the metrics in the test has high correlation (high variance) and the Bonferroni correction will be too conservative to it. We would only launch if all evaluation metrics show a significant change. In that case, there would be no need to use Bonferroni correction.

## Recommendation

Make a recommendation and briefly describe your reasoning.

The evaluation metrics are Gross conversion and Net conversion.

Gross conversion is negative and practically significant. This is a good outcome because we lower our costs by discouraging trial signups that are unlikely to convert.

But Net conversion unfortunately is statistically and practically insignificant and the confidence interval includes negative numbers. Therefore, there is a risk that the introduction of the trial screener may lead to a decrease in revenue.

We should therefore consider test other designs of the screener before we decide whether to release the feature, or abandon the idea.

## 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.

We can provide a discount for the first month if the student arrange a one-to-one discussion with coach during free trial period; it may reduce early cancellation.

After students experiment the support of coach they would feel supported and have more confidence that they will accomplish the course, thus they would be more willing to pay.

We want to use user-ids as unit of diversion because it is more stable than cookies. We define cancellation rate as number of users who click “Start free trial”and cancel enrollment in 14-days divided by number of users who click “Start free trial” button.

We will use cancellation rate as evaluation because this is about probability of early cancellation, and we want to know whether the probability of early cancel decrease in our experiment.