# P7 A/B Testing

By: Darren Liu

# Experiment Design

## Metric Choice

Number of cookies and number of clicks are chosen as the invariant metrics. Invariant metrics serve as “sanity checks”, metrics that one would expect to be the same in both the control group and experimental group. Checking these metrics allow us to confirm that the conditions of the experiment are valid. If the tests fail, the experiment should be halted until the reasons for the discrepancy are found and fixed.

There is a high chance other metrics will not be invariant. Number of user-ids should be lower for the experimental group since visitors are likely discouraged to enroll. Gross conversion, retention, and net conversion are all ratios of user-ids and are affected for the same reason. Click-through-probability is valid, but given it is actually just a ratio of our two invariant metrics, to test it as well would be redundant.

Gross conversion and net conversion are chosen as the evaluation metrics. Evaluation metrics should measure rate of success between control and experiment groups on the hypothesis of interest. In this case, whether a given change will impact user decision to enroll and remain enrolled. Gross conversion is expected to decrease or stay the same, since the tested feature is meant as a deterrent for students unprepared for the workload. Net conversion is expected to increase or stay the same, since the feature should have filtered out students that are more prone to quitting midway.

A decrease in gross conversion without a decrease in net conversion would be viewed as a positive outcome and given consideration for launch. Although an outsized decrease in gross conversion and a relatively smaller decrease in net conversion might theoretically justify a launch, net conversion must not decrease in any significant manner as it will conflict with existing business goals.

Number of pageviews and clicks are invariant and not considered. User-ids are raw counts and can be not be compared fairly if pageviews are not equal. Click-through-probability is not valid since the feature being tested comes after the click. Retention rate is a valid evaluation metric but is found to have unrealistic sample size requirements.

## Measuring Standard Deviation

Gross conversion - 0.0202

Net conversion - 0.0156

The unit of analysis and unit of diversion are the same, eliminating concern of difference in sampling independence that would otherwise significantly impact variance. Therefore, it is safe to assume analytic variability to represent empirical variability.

## Sizing

### Number of Samples vs. Power

Bonferroni correction is not used. 685,325 pageviews are needed for proper experiment setup.

### Duration vs. Exposure

100% of traffic are diverted, requiring 18 days to complete.

All traffic are included because the experiment is assessed as unrisky. Some examples that may give rise to concern include system failure, damage of company or user reputation, or exposure of sensitive information. None of these cases would reasonably apply to this experiment.

# Experiment Analysis

## Sanity Checks

Number of cookies:

Interval - (0.4988, 0.5012)

Observed - 0.5006

Passes - True

Number of clicks:

Interval - (0.4959, 0.5041)

Observed - 0.5005

Passes - True

## Result Analysis

### Effect Size Tests

Gross conversion:

Interval - (0.0291, -0.012)

Statistical significance - True

Practical significance - True

Net conversion:

Interval - (-0.0116, 0.0019)

Statistical significance - False

Practical significance - False

### Sign Tests

Gross conversion:

P-value - 0.0026

Statistically significant - True

Net conversion:

P-value - 0.6776

Statistically significant - False

### Summary

The Bonferroni correction is not used. The use of Bonferroni would essentially lower alpha and increase beta. Its use is advocated by the desire to avoid adverse consequence from observation of ANY false positives. That is not the case here, since the positive outcome of any one metric would not warrant launch considerations, both would have to be satisfied. In other words, false negatives is actually the main concern. The use of Bonferroni would necessarily increase beta, and increase our risk of false experiment.

The results of the effect size tests and sign tests were consistent. Gross conversion is found to be lower in the experimental group than the control group. The difference in net conversion between the two groups are found to be negligible.

## Recommendation

Based on the findings, the recommendation is to launch the tested feature. This is based on our rejection of the first null hypothesis, that gross conversion equals zero, with a negative confidence interval, as well as the failure to reject the second null hypothesis, that net conversion equals zero. As a result of the tested feature, less people signed up for the free trial yet the same number of people made first payment. This shows the feature effectively eliminating users that were not prepared to become dedicated paying customers, without adverse effect to paying customer count. The business benefit is less resources expended on the customers of the 14 day trial and a better retention rate.

# Follow-Up Experiment

In an attempt to lower early cancellation, it might be interesting to introduce a feature where if the total time played on a video exceeds 3 times its length, a pop-up reminds the student there are forum information available. If possible, even filter the exact threads pertaining to the topic of the video. The null hypothesis would be that the student retention difference is equal to zero. It would be a two tail test to test for both positive or negative impacts. User-id would be the unit of diversion, as each user-id is an unique paying student and makes sense as our denominator.

User-id also makes sense as the invariant metric, same number of samples are expected from the control and experiment group. The evaluation metric would be the retention rate 30 days after first payment. It is reasonable to believe students that will end subscription in 30 days will be much less than students that ended free trial. Therefore, sample size problems should not exist as it did before, although we will not know for sure without some analytical estimates.