| Author | Project Title | Date |
|---|---|---|
| Dewi Octavia | Design an A/B Test | 14th May 2017 |

# Experiment Overview

In this project, A/B Test is deployed on actual Udacity experiment called Free Trial Screener with a mock-up data. The experiment tested a change on having a pop-up time commitment question to enrollees after they opted for a 14 days free trial course. If 5 or more hours per week was indicated, they would be taken through the check-out process as usual. If they indicated fewer than 5 hours per week, a message would appear indicating that Udacity courses usually require a greater time commitment for successful completion, and suggesting that the enrollees might like to access the course materials for free which do not come with coaching support or a verified completion certificate. At this point, enrollees would have the option to continue enrolling in the free trial, or access free course material instead.

The hypothesis was that this might set clearer expectations for students upfront hence reducing number of frustrated students who left the free trial because they didn't have enough time—without significantly reducing the number of students to continue past the free trial and eventually complete the course. If this hypothesis held true, Udacity could improve the overall student experience and improve coaches' capacity to support students who are likely to complete the course.

The unit of diversion is a cookie, although if the student enrolls in the free trial, they are tracked by user-id from that point forward. The same user-id cannot enroll in the free trial twice. For users that do not enroll, their user-id is not tracked in the experiment, even if they were signed in when they visited the course overview page.

# 1. Experiment Design

## 1.1 Metric Choice

The metric chosen as invariant metrics are number of cookies, number of clicks and click-through-probability.
The chosen evaluation metrics are gross conversion, retention and net conversion.

The reasons behind metrics choice are as follows:

- Number of cookies: number of unique cookies to view the course overview page
  It is an invariant metric because viewing overview page happens before the experiment hence it is not expected to change.

- Number of user-ids: Number of users who enroll in the free trial
  It is not a good variant metric because number of users enroll in free trial is dependent on the experiment, number of user-ids in control group is expected to be different to experiment group. Number of user-ids alone is also not a good evaluation metric because the number of enrolment could fluctuate on a particular day and skew the analysis. It is more useful to have the number of user-ids as ratio to number of unique cookies who clicked.

- Number of clicks: number of unique cookies to click the “Start Free Trial” button
  This is also an invariant metric for the same reason as number of cookies. Clicking “Start Free Trial” button happens right before the experiment hence it should not change.

- Click-through-probability: number of unique cookies to click “Start Free Trial” button divided by number of unique cookies to view the course overview page.
  This metric is combination of two metrics earlier, number of cookies and number of clicks. Click-through-probability can be defined as invariant metric as both its numerator and denominator are not expected to change in this free trial screener experiment. Again, “Start Free Trial” click happens before the experiment therefore it should not change.

- Gross conversion: number of user-ids to complete check-out and enroll in the free trial divided by number of unique cookies to click the “Start Free Trial” button.
  This is chosen as evaluation metric because gross conversion is a measurement that is directly dependent of the impact of experiment. It is expected to be different between control and experiment groups. In the trial screener experiment, users in experiment group get to respond to the time commitment heads-up, either the user committed to course and continue on the enrolment step, or user choose to access the course materials for free instead. On the other hand, the user in control group is not affected by the experiment and will be taken through to the checkout and enrolment process. In this experiment, the gross conversion in experiment group is expected to be lower than control group. Lower gross conversion can also indicate lower chance of losing potential student hence lower coaches’ capacity (hence cost).

- Retention: 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.
  Retention is a good evaluation metric for this experiment. This measurement allow us to confirm the commitment of experiment group who enroll in the free trial course, in other words, experiment group is expected to have higher retention and less likely to leave the free trial than control group who was not questioned about time commitment.

- Net Conversion: 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.
  Net conversion is also a good evaluation metric. It is expected that the user in experiment group who is aware of the course time commitment to stay on the course after the free trial period hence net conversion in experiment group might be higher than the control who has no clue about the time commitment.

The hypothesis of this experiment was to reduce number of students leaving the free trial due to insufficient time as well as not significantly reduce number of students to continue past the trial and eventually complete the course.

For launching the experiment, the gross conversion in experiment group is expected to be lower than control group, which could indicate lower chance of losing potential student and coaches’ capacity hence lower cost. Net conversion is expected not to decrease and retention rate to increase. The experiment is considered a success if we are able to reduce the number of less motivated enrollees and increase those who remain.

Retention is measured by numbers of user-ids to complete checkout and stay on past 14 days. These two steps are the later stages in customer flow hence it suspected to require large number of cookies visiting overview page to properly power the experiment. This will be confirmed in sizing stage.

## 1.2 Measuring Standard Deviation

Given a sample size of 5000 cookies visiting the course overview page, make an analytical estimate of standard deviation of evaluation metrics. Baseline values are as follows:

| | |
|---|---|
| Unique cookies to view the page per day | 40000 |
| Unique cookies to click "Start Free Trial" per day | 3200 |
| Enrolments per day | 660 |
| Click-through-probability on "Start Free Trial" | 0.08 |
| Probability of enrolling, given click | 0.20625 |
| Probability of payment, given enroll | 0.53 |
| Probability of payment, given click | 0.1093125 |

Calculation of standard deviations:

$$Gross\ conversion = \frac{\#\ enroll}{\#\ cookies\ to\ click}$$

| | |
|---|---|
| p, probability of enrolling, given click | 0.20625 |
| N, number of click $= \frac{3200\times5000}{4000}$ | 400 |
| SD = $\sqrt{p(1-p)/N}$ | **0.0202** |

$$Retention = \frac{\#\ payment}{\#\ enroll}$$

| | |
|---|---|
| p, probability of payment, given enroll | 0.53 |
| N, number of enroll $= \frac{660\times5000}{4000}$ | 82.5 |
| SD = $\sqrt{p(1-p)/N}$ | **0.0549** |

$$Net\ conversion = \frac{\#\ payment}{\#\ cookies\ to\ click}$$

| | |
|---|---|
| p, probability of payment, given click | 0.1093125 |
| N, number of click $= \frac{3200\times5000}{4000}$ | 400 |
| SD = $\sqrt{p(1-p)/N}$ | **0.0156** |

**Standard deviation results:**

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

Both gross and net conversions have denominator of number of unique cookies who click the "Start Free Trial", which is also unit of diversion. As the unit of analysis equals unit of diversion, the analytical estimate is comparable to empirical estimate.

Retention has denominator of number of user-ids who enroll which is different from the unit of diversion, cookie. As the unit of analysis is different to unit of diversion, analytical and empirical estimates are different therefore collecting empirical estimate of the variability would be desirable if time allows.

## 1.3 Sizing

### 1.3.1 Number of Samples vs. Power

The number of unique cookies who click required will be calculated using online sample size calculator and number of pageviews will be calculated subsequently.

Total number of pageviews required to adequately power the experiment, with alpha of 0.05 and beta of 0.2:

**Gross Conversion:**

| | |
|---|---|
| Baseline conversion rate (%) | 20.625 |
| Minimum detectable effect (%) | 1 |
| Number of unique cookies who click required | 25835 per group |
| Number of pageviews = $\frac{25835\times2}{0.08}$ | 645875 |

**Net Conversion:**

| | |
|---|---|
| Baseline conversion rate (%) | 10.93125 |
| Minimum detectable effect (%) | 0.75 |
| Number of unique cookies who click required | 27413 per group |
| Number of pageviews = $\frac{27413\times2}{0.08}$ | 685325 |

**Retention:**

| | |
|---|---|
| Baseline conversion rate (%) | 53 |
| Minimum detectable effect (%) | 1 |
| Number of unique cookies who click required | 39115 per group |
| Number of pageviews = $\frac{39115\times2}{0.0165}$ | 4741213 |

**Final Results of Number of Pageviews Calculations**

| | |
|---|---|
| Gross Conversion | 645875 |
| Net Conversion | 685325 |
| Retention | 4741213 |

As suspected, in order to power the retention adequately, large number of pageviews is required. Based on this calculation, it is likely that retention is abandoned as evaluation metric. Its experiment duration will be calculated next.

Therefore, at this stage the number of pageviews required is **685325** to properly power gross and net conversions.

### 1.3.2 Duration vs. Exposure

The risk of this experiment is assessed to see whether it exceeds minimal risk and further review is required before proceeding. In this experiment, users are required to simply respond to time commitment to the course hence there is no harm expected to user physically or psychologically. The data collected was number of unique cookies and user-ids once the user enrols in the course. Cookies is an anonymous data and does not allow re-identification of an individual. User-ids are sensitive data as they are personally identifiable however number of user-ids is not. Tally of user-ids does not expose any personal details. Besides data of user-ids exists before this experiment therefore it is not a new data. In data collection process, no other sensitive data is collected such as personal health information, political views or sexual preferences. Therefore this experiment has low risk and it does not exceed minimal risk boundary.

Assuming there were no other experiments running simultaneously to this free trial screener experiment, entire Udacity traffic will be diverted to this experiment as the experiment is low risk and does not touch on any ethical issues. One possible failure in this experiment is issue in infrastructure setup or click counter setup however it will not affect other Udacity activities.

If we direct the whole Udacity traffic to this experiment, and using gross and net conversion metrics, the duration of this experiment will be:

$$experiment\ duration = \frac{685325\ pageviews}{40000\ pageviews/day} \approx 18\ days$$

This experiment will take about 18 days or 3 weeks to run, which is reasonable.

The experiment duration is also calculated if retention was considered:

$$retention\ experiment\ duration = \frac{4741213\ pageviews}{0.8 \times 40000\ pageviews/day} \approx 119\ days$$

It is confirmed from the calculation above that running experiment with retention as evaluation metric will take a long time, about 119 days or approximately 4.5 months. This is not feasible hence retention is abandoned.

# 2. Experiment Analysis

## 2.1 Sanity Checks

**Experiment Results Summary:**

| | **Control** | **Experiment** |
|---|---|---|
| Total Pageviews | 345543 | 344660 |
| Total Clicks | 28378 | 28325 |

For each invariant metrics chosen, sanity check is performed with a 95% confidence level as shown below:

**Number of cookies**

| | |
|---|---|
| p, probability of cookie assigned into control group | 0.5 |
| SD = $\sqrt{(0.5)(0.5)(1/(345543 + 344660)}$ | 0.0006 |
| m, margin of error with z score= 1.96 | 0.0012 |
| CI lower bound | **0.4988** |
| CI upper bound | **0.5012** |
| Observed = $345543/(345543 + 344660)$ | **0.5006** |
| Observed value is within confidence interval ∴ **pass sanity check** | |

**Number of clicks**

| | |
|---|---|
| p, probability of cookie assigned into control group | 0.5 |
| SD = $\sqrt{(0.5)(0.5)(1/(28378 + 28325)}$ | 0.0021 |
| m, margin of error with z score= 1.96 | 0.0041 |
| CI lower bound | **0.4959** |
| CI upper bound | **0.5041** |
| Observed = $28378/(28378 + 28325)$ | **0.5005** |
| Observed value is within confidence interval ∴ **pass sanity check** | |

**Click-through-probability (CTP)**

| | |
|---|---|
| CTP control = $28378/345543$ | 0.0821 |
| SD = $\sqrt{(0.0821)(0.9179)(1/345543)}$ | 0.0005 |
| m, margin of error with z score= 1.96 | 0.0009 |
| CI lower bound | **0.0812** |
| CI upper bound | **0.0830** |
| Observed = CTP experiment = $28325/344660$ | **0.0822** |
| Observed value is within confidence interval ∴ **pass sanity check** | |

## 2.2 Results Analysis

### 2.2.1 Effect Size Tests

**Experiment Results Summary**

| | Control | Experiment |
|---|---|---|
| Total Clicks | 17293 | 17260 |
| Total Enrollments | 3785 | 3423 |
| Total Payments | 2033 | 1945 |

For each evaluation metrics chosen, confidence interval for the difference between the experiment and control groups are calculated as follows:

$$Gross\ conversion = \frac{\#\ enroll}{\#\ cookies\ to\ click}$$

| | |
|---|---|
| $\hat{p}_{pool}$ = $(3785 + 3423)/(17293 + 17260)$ | 0.2086 |
| $SE_{pool}$ = $\sqrt{0.2086(0.7914)\left((\frac{1}{17293}) + (\frac{1}{17260})\right)}$ | 0.0044 |
| $\hat{d}$ = $\left(\frac{3423}{17260}\right) - \left(\frac{3785}{17293}\right)$ | -0.0206 |
| m = 1.96 x $SE_{pool}$ | 0.0086 |
| CI lower bound | **-0.0291** |
| CI upper bound | **-0.0120** |
| $d_{min}$ | 0.01 |
| $\lvert m \rvert < \lvert \hat{d} \rvert$ and CI not include 0 ∴ **statistically significant** | |
| CI [-0.0291, -0.0120] not include $d_{min}$ of 0.01 ∴ **practically significant** | |

$$Net\ conversion = \frac{\#\ payment}{\#\ cookies\ to\ click}$$

| | |
|---|---|
| $\hat{p}_{pool}$ = $(2033 + 1945)/(17293 + 17260)$ | 0.1151 |
| $SE_{pool}$ = $\sqrt{0.1151(0.8846)\left((\frac{1}{17293}) + (\frac{1}{17260})\right)}$ | 0.0034 |
| $\hat{d}$ = $\left(\frac{1945}{17260}\right) - \left(\frac{2033}{17293}\right)$ | -0.0049 |
| m = 1.96 x $SE_{pool}$ | 0.0067 |
| CI lower bound | **-0.0116** |
| CI upper bound | **0.0019** |
| $d_{min}$ | 0.0075 |
| $\lvert m\rvert > \lvert\hat{d}\rvert$ and CI includes 0 ∴ **not statistically significant** | |
| CI [-0.0116, 0.0019] includes $d_{min}$ of +/- 0.0075 ∴ **not practically significant** | |

### 2.2.2 Sign Tests

For each evaluation metric, sign test was performed by using 23 days data breakdown. P-values are obtained from GraphPad's sign and binomial test calculator.
Successes are defined as when there is change between control and experiment groups.

**Gross Conversion**

| | |
|---|---|
| Number of successes | 19 |
| Number of trials | 23 |
| Probability for sign test | 0.5 |
| Two-tail P value | **0.0026** |
| $\alpha$ | 0.05 |
| p-value < $\alpha$ ∴ sign test agrees with hypothesis test, that is result is unlikely to happen by chance and the change is **statistically significant** which agrees with the effect size test. | |

**Net Conversion**

| | |
|---|---|
| Number of successes | 13 |
| Number of trials | 23 |
| Probability for sign test | 0.5 |
| Two-tail P value | **0.6776** |
| $\alpha$ | 0.05 |
| p-value > $\alpha$ ∴ sign test rejects hypothesis test, that is result could happen by chance and the change is **not statistically significant** which agrees with the effect size test. | |

### 2.2.3 Summary

When we have multiple metrics to consider, the risk of rejecting null by pure chance (Type I error) is high. Bonferroni correction does not assume that metric is independent and gives smaller overall alpha than overall alpha with independence assumption. Smaller overall alpha results in larger z-score hence larger margin of error. Having wider margin of error means it is more likely for metric to have no significant difference hence less chance/risk to reject null hypothesis when there is no true difference (or when null is true). Therefore Bonferroni correction is useful when we have multiple metrics and only need any (or at least one) of the metrics to match the expectations, in other words we can tolerate higher probability of false negatives, which is the event where we fail to reject null when there is true difference.

However, for launching the experiment, we need to strictly meet all two conditions that are gross conversion to decrease and net conversion unchanged. In other words, we need all our metrics to match the expectations and if one of the metric shows a false positive, we would not launch our experiment. Type I error which Bonferroni correction is good at catching is not our main concern. Our risk is Type II error where we accept null when there is significant difference on some of the metrics and costing the launch. Based on this reason, Bonferroni correction is not so useful in this experiment.

The results from effect size hypothesis test agree with sign test results which stated gross conversion metric is both statistically and practically significant, whereas net conversion is neither statistically nor practically significant.

## 2.3 Recommendation

The analysis results above show that only gross conversion is practically significant and there is no statistically significant change in net conversion. In effect size test, the confidence interval includes the negative of practical significance boundary which could mean that the number of paying students decreased by an amount that would matter to the business. Therefore, it is recommended *NOT* to launch the experiment. However a follow-up experiment could be trialled as discussed in next section.

# 3. Follow-up Experiment

Time commitment to the course does not seem to guarantee students to stay on past the free trial period. Early cancellations can be caused by students who previously agreed on time commitment, getting discouraged after finding their pre-requisite skill level is not up to par or having difficulty recalling previous studies or skills.

The proposed follow-up experiment is to have a snippet video of each lessons followed by beginner level quiz corresponding to the lesson after user completes enrolment and before the free trial course. The videos and quizzes would help student to understand further what is expected in the course and be more prepared before diving into the full lesson. The quizzes will have “View Answer” option available, which could also help as refresher mini course (hence build confidence) and not turning away student who gets stuck in the quiz.

The experiment group will be taken through to short videos and quizzes after checkout and enrolment process whereas the control group will bypass the videos and quizzes step after enrolling and can start the course right after.

The hypothesis of this experiment was that this might better prepare and motivate students therefore reducing number of early cancellations.

Unit of diversion will be user-id which is more stable than a cookie that is group assignment of user-id will not change.

Metrics for this follow-up experiment would be number of user-ids as invariant metric, we do not expect number of user-ids to change because the follow-up experiment happens after enrolment process. Retention will be the evaluation metric as it is a measurement that is directly dependent on the experiment. The retention in experiment group is expected to be higher than that of control group.

## 4. Resources

- http://www.evanmiller.org/ab-testing/sample-size.html
- http://graphpad.com/quickcalcs/binomial2/
- http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/type-i-and-type-ii-error/
- https://www.utdallas.edu/~herve/Abdi-Bonferroni2007-pretty.pdf