A/B Testing Final Project

Dani Chao
danicychao@gmail.com

1 Experiment Design

The experiment is shown in Fig. 1.

1.1 Metric Choice

1.1.1 Invariant metrics

- **Number of cookies:** number of unique cookies to view the course overview page.
- **Number of clicks:** number of unique cookies to click the 'Start free trial' button.
- **Click-through-probability:** number of unique cookies to click the 'Start free trial' button divided by number of unique cookies to view the course overview page.

Since the number of cookies, the number of clicks, and the click-through-probability all happen before the free trial screener is triggered, they are all invariant metrics that will be used for checking if the experiment setups are the same between the control group and experiment group. These three metrics should not have any significant difference between the control and experiment groups.

1.1.2 Evaluation metric(s)

Scenario 1: retention Retention is the number of user-ids to remain enrolled past the 14-day boundary divided by the number of user-ids to complete checkout, and it is the most direct metric used to test the hypothesis in this experiment. If the retention of the experiment group is significantly higher than the retention of the control group, it is verified by the experiment that

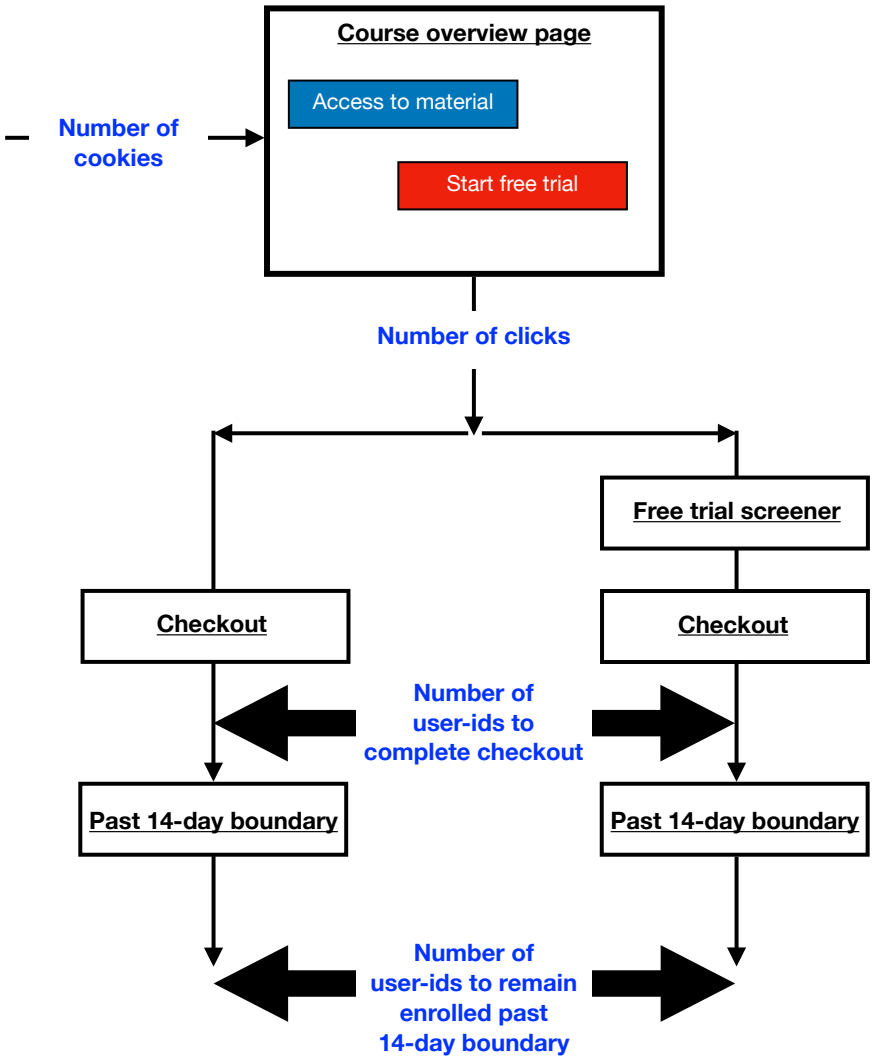

Figure 1: Flowchart of the experiment.

the free trial screener effectively reduces the number of frustrated students leaving the free trial without significantly reducing the number of students continuing the free trial and completing the course. However, the free trial screener might also reduce the number of user-ids to complete checkout, potentially requiring more time to achieve a sufficiently large sample for the experiment group.

Hypothesis testing of retention:

$$H_0 : R_{\text{exp.}} = R_{\text{con.}} \quad \text{vs.} \quad H_1 : R_{\text{exp.}} \neq R_{\text{con.}}, \tag{1}$$

where $R_{\text{exp.}}$ and $R_{\text{con.}}$ are the retention of the experiment and control groups, respectively.

Scenario 2: gross conversion and net conversion An alternative way to verify whether the free trial screener is effective or not is to conduct two hypothesis tests on gross conversion and net conversion, which are defined as the following:

- **Gross conversion:** 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.
- **Net conversion:** number of user-ids to remain enrolled past the 14-day boundary divided by the number of unique cookies to click the 'Start free trial' button.

The hypothesis of this experiment is that the free trial screener might reduce the number of frustrated students leaving the free trial without significantly reducing the number of students continuing the free trial and completing the course, and this equivalently assumes that:

A. The students having indicated fewer than 5 hours per week in the experiment group might bail out from enrolling in the free trial after the free trial screener is triggered.

B. The students with enough time might proceed to enroll and remain enrolled past the 14-day boundary regardless of the group assignment.

If both of the following are true:

a. The gross conversion of the experiment group is significantly lower than the gross conversion of the control group.

b. There is no significant difference in net conversion between the experiment and control groups.

then the experiment provides evidence that the free trial screener effectively reduces the number of frustrated students leaving the free trial, without significantly reducing the number of students continuing the free trial and completing the course.

Hypothesis tests of gross conversion and net conversion:

$$H_0^1 : GC_{\text{exp.}} = GC_{\text{con.}} \quad \text{vs.} \quad H_1^1 : GC_{\text{exp.}} \neq GC_{\text{con.}} \tag{2}$$

$$H_0^2 : NC_{\text{exp.}} = NC_{\text{con.}} \quad \text{vs.} \quad H_1^2 : NC_{\text{exp.}} \neq NC_{\text{con.}}, \tag{3}$$

where $GC_{\text{exp.}}$ is the gross conversion of the experiment group, $GC_{\text{con.}}$ is the gross conversion of the control group, $NC_{\text{exp.}}$ is the net conversion of the experiment group, and $NC_{\text{con.}}$ is the net conversion of the control group.

1.2 Measuring Standard Deviation

- **The analytical standard deviation of retention:** 0.01942741.
- **The analytical standard deviation of gross conversion:** 0.007152599.
- **The analytical standard deviation of net conversion:** 0.005515979.

While the standard deviations of the three metrics are all very likely to be underestimated by their analytical standard deviation, it is strongly recommended to use empirical variability to estimate the standard deviation for retention (scenario 1) and net conversion (scenario 2). Since it takes 14 days to measure retention and net conversion, this 14-day delayed feedback might correlate with early behaviors or cohort effects. The fact that the numerator and denominator of net conversion are in different units of diversion also introduces potential noise and correlation due to multi-device behavior (1 user = many cookies) and dropped/unconverted cookies (not all clickers enroll).

When the time is enough, we only have to use retention to launch the experiment. As a result, it might be worth doing an empirical estimate for retention. When we decide to use both gross conversion and net conversion to launch the experiment, it might still be worth doing an empirical estimate for net conversion.

1.3 Sizing

1.3.1 Scenario 1: using retention only

If we only use retention to launch the experiment, we will not need the Bonferroni correction. To achieve the usual 5% significance level and 80% statistical power for retention, we will need at least 39,051 user-ids completing the checkout and enrolling for each group. Based on the probability of enrolling given click and click-through-probability listed in the baseline values, we will need $\frac{39{,}051}{0.20625} \sim 189,339$ clicks and $\frac{189{,}339}{0.08} \sim 2,366,738$ pageviews for each group. In total, we will need 4,733,476 pageviews to achieve 5% significance level and 80% statistical power for retention.

Given the daily pageviews listed in the baseline values, we will need $\frac{4{,}733{,}476}{40{,}000} \sim 119$ days with full traffic to achieve 5% significance level and 80% statistical power for retention. Therefore, the experiment will take too long if we use retention as the evaluation metric. The facts that it is too risky to use full traffic and that there is a potential decrease of the enrolling probability given click due to the triggered screener could make the situation even worse.

1.3.2 Scenario 2: using gross conversion and net conversion

Now, we have to take the alternative way using gross conversion and net conversion to verify whether the free trial screener is effective or not. Although there are two hypothesis tests, the scenario here is not a Bonferroni test. Unlike a Bonferroni test, where one rejects null hypothesis if any single test's p-value is below the Bonferroni-corrected significance level, the scenario of using gross conversion and net conversion requires both hypothesis tests to pass, rejecting the null hypothesis in Eq. 2 and not rejecting the null hypothesis in Eq. 3. In other words, the Bonferroni test is an 'OR' condition, but our scenario here is an 'AND' condition. Therefore, we will not use the Bonferroni correction.

We will need at least 25,231 and 27,978 unique cookies to click the 'Start free trial' button in each group for gross conversion and net conversion, respectively. To reach 27,978 clicks for each group, we will need $\frac{27{,}978}{0.08} \sim 349,725$ pageviews for each group. In total, we will need 699,450 pageviews to achieve 5% significance level for the experiment.

If we use the full traffic to run the experiment with the two hypothesis tests, we will need $699,450/40,000 \sim 18$ days to achieve 5% significance level. Since taking full traffic for an experiment is not realistic, I would suggest to use 75% of the traffic to run the experiment, which will take ∼1 month to collect enough pageviews to achieve 5% significance level. However, we will need extra 14 days to collect all the feedback.

2 Experiment Analysis

2.1 Sanity Checks

We will check if the pageviews and click-through-probabilities are equivalent between the control and experiment groups.

2.1.1 Pageviews

The distribution of pageviews between the control and experiment groups should be a Bernoulli distribution with the fair probability $\frac{1}{2}$, Ber($\frac{1}{2}$). In the Bernoulli distribution of the pageviews, we denote the pageviews of the experiment group as 1 and the pageviews of the control group as 0. Given that the pageviews of the experiment and control groups are 211,362 and 212,163, respectively, the standard deviation of pageviews across the experiment and control groups is

$$\sqrt{\frac{0.5 \times 0.5}{211,362 + 212,163}} \sim 0.0007682994, \tag{4}$$

and the ratio of the experiment group's pageviews is

$$\frac{211,362}{211,362 + 212,163} \sim 0.4990544. \tag{5}$$

Since the interval

$$[0.5 - 1.96 \times 0.0007682994, 0.5 + 1.96 \times 0.0007682994] = [0.4984941, 0.5015059] \tag{6}$$

contains the value 0.4990544, the pageviews are equivalent between the experiment and control groups (with 95% confidence).

2.1.2 Click-through-probability (CTR)

To verify whether the CTRs are equivalent between the control and experiment groups, we need to calculate its pooled mean, its pooled standard error, and its difference between the control and experiment groups. The pooled mean of the CTR is

$$\frac{17,293 + 17,260}{212,163 + 211,362} \sim 0.08158432, \tag{7}$$

the pooled standard error of the CTR is

$$\sqrt{0.0816 \times (1 - 0.0816) \times (\frac{1}{212,163} + \frac{1}{211,362})} \sim 0.0008413027, \tag{8}$$

and the difference of the CTRs is

$$\frac{17,260}{211,362} - \frac{17,293}{212,163} \sim 0.00015276. \tag{9}$$

Since the interval

$$[-1.96 \times 0.0008413027, 1.96 \times 0.0008413027] = [-0.001648953, 0.001648953] \tag{10}$$

contains the value 0.00015276, the CTRs are also equivalent between the control and experiment groups (with 95% confidence).

Since both the pageviews and CTRs are equivalent between the experiment and control groups, we pass the sanity check.

2.2 Result Analysis

2.2.1 Effect size tests

We first calculate the 95% confidence intervals around the difference of gross conversion and net conversion between the experiment and control groups for the tests of Eqs. 2 and 3.

Gross conversion The difference of gross conversion between the experiment and control groups is $\frac{3,423}{17,260} - \frac{3,785}{17,293} \sim -0.02055487$. The pooled mean of gross conversion is

$$\frac{3,423 + 3,785}{17,260 + 17,293} \sim 0.2086071, \tag{11}$$

and the pooled standard deviation of gross conversion is

$$\sqrt{0.2086071 \times (1 - 0.2086071) \times (\frac{1}{17,260} + \frac{1}{17,293})} \sim 0.004371676. \tag{12}$$

Therefore, the 95% confidence interval around the difference of gross conversion is

$$[-0.02055487 - 1.96 \times 0.004371676, -0.02055487 + 1.96 \times 0.004371676] = [-0.02912335, -0.01198639]. \tag{13}$$

Since 0 does not lie in the interval in Eq. 13 and the absolute difference, $|-0.02055487| = 0.02055487$, is larger than the practical significance boundary, 0.01, we reject the null hypothesis H_0^1 in Eq. 2. It is verified that the experiment group has significantly lower gross conversion than the control group both statistically and practically.

Net conversion The difference of net conversion between the experiment and control groups is $\frac{1,945}{17,260} - \frac{2,033}{17,293} \sim -0.004873723$. The pooled mean of net conversion is

$$\frac{1,945 + 2,033}{17,260 + 17,293} \sim 0.1151275, \tag{14}$$

and the pooled standard deviation of net conversion is

$$\sqrt{0.1151275 \times (1 - 0.1151275) \times (\frac{1}{17,260} + \frac{1}{17,293})} \sim 0.003434134. \tag{15}$$

Therefore, the 95% confidence interval around the difference of net conversion is

$$[-0.004873723 - 1.96 \times 0.003434134, -0.004873723 + 1.96 \times 0.003434134] = [-0.01160463, 0.00185718] \tag{16}$$

Since 0 lies in the interval in Eq. 16, the net conversions are equivalent between the experiment and control groups (with 95% confidence). Therefore, we do not reject the null hypothesis H_0^2 in Eq. 3. It is verified that there is no significant difference in the net conversion between the experiment and control groups.

2.2.2 Sign tests

Gross conversion Using the day-by-day data, we denote

$$X = \begin{cases} 1, & \text{if } GC_{\text{con.}} > GC_{\text{exp.}} \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

The distribution of X is a Bernoulli distribution, $X \sim \text{Ber}(p)$, and $p = \frac{1}{2}$ if $GC_{\text{con.}}$ and $GC_{\text{exp.}}$ are equally likely to be larger than the other. The hypothesis of the sign test:

$$H_0 : p = 0.5 \quad \text{vs.} \quad H_1 : p \neq 0.5. \tag{18}$$

Among the total 23 days with full data, there are 19 days that $GC_{\text{con.}} > GC_{\text{exp.}}$. Therefore, the test statistic

$$T = \frac{\frac{19}{23} - \frac{1}{2}}{\sqrt{\frac{\frac{19}{23} \times (1 - \frac{19}{23})}{23}}} \sim 4.125897, \tag{19}$$

and the corresponding p-value is ~ 0.000037. Since the p-value 0.000037 is much smaller than 5%, we reject the null hypothesis, $H_0 : p = 0.5$. Therefore, the result that the experiment group has lower gross conversion is statistically significant.

Net conversion Similarly, we denote

$$X = \begin{cases} 1, & \text{if } NC_{\text{exp.}} > NC_{\text{con.}} \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

$X \sim \text{Ber}(p)$, and the hypothesis of the sign test:

$$H_0 : p = 0.5 \quad \text{vs.} \quad H_1 : p \neq 0.5. \tag{21}$$

Among the total 23 days with full data, there are 10 days that $NC_{\text{exp.}} > NC_{\text{con.}}$. Therefore, the test statistic

$$T = \frac{\frac{10}{23} - \frac{1}{2}}{\sqrt{\frac{\frac{10}{23} \times (1 - \frac{10}{23})}{23}}} \sim -0.6309334, \tag{22}$$

and the corresponding p-value is ~ 0.5280841. Since the p-value 0.5280841 is much greater than 5%, we fail to reject the null hypothesis, $H_0 : p = 0.5$. Therefore, the net conversions of the experiment and control groups are equally likely to be larger than the other.

2.2.3 Summary

The effect size hypothesis tests and the sign tests agree with each other, and both provide evidence to reject the null hypothesis in Eq. 2 and to not reject the null hypothesis in Eq. 3. As a result, this experiment provides evidence that the free trial screener effectively reduces the number of frustrated students leaving the free trial, without significantly reducing the number of students continuing the free trial and completing the course. However, we only have 423,525 pageviews over the 23 days when all feedback is collected, which is below 699,450 pageviews required for sufficient statistical power.

2.3 Recommendation

Although the experiment so far has shown that the free trial screener could help Udacity reduce the number of frustrated students, and further improving the overall student experience and coaches' capacity to support students who are likely to complete the course, I suggest that **we should wait until we collect all the feedback over the 37 days covering the whole experiment period**. According to the spreadsheet provided by the project instruction, there are 690,203 pageviews in total over the 37 days, and we have only collected the delayed feedback of the number of payments and enrollments for 23 days. After we collect the number of payments and enrollments for the remaining 14 days, the experiment will have enough (statistical) power for us to make the decision if we should launch the new feature of the free trial screener.

I note that the free trial screener is actually a passive strategy in terms of business. Although the free trial screener successfully reduces the number of frustrated students, Udacity also loses the opportunity to make profits from the people who click the 'Start free trial' button but do not proceed to enroll because of being triggered by the free trial screener. Udacity should also come up with a follow-up experiment to actively encourage the people who enroll to have steady learning progress and to help them grow sense of achievement. In short, Udacity should also have active strategies to improve the retention.

3 Follow-Up Experiment

- **Experiment:** we send a daily reminder with suggested progress to the students who complete the checkout and enroll in the free trial.
- **Hypothesis:** this might set a consistent pace for students who enroll, helping them manage their time and progress steadily, thus reducing the number of students who leave the free trial due to falling behind. If this hypothesis holds, Udacity could further improve the overall student experience.
- **Unit of diversion:** a user-id. Since students are tracked by user-ids after they enroll, we should use user-id as the unit of diversion to record if they remain enrolled past the 14-day boundary.
- **Metric choice:**
  - **Number of checkout ids:** number of user-ids to complete checkout and enroll in the free trial.
  - **Number of remaining ids:** number of user-ids to remain enrolled past the 14-day boundary.
  - **Retention:** number of user-ids remain enrolled past the 14-day boundary divided by number of user-ids to complete checkout.

Since completing checkout and enrolling happens before the daily reminder is sent, the number of checkout ids should be used as the invariant metric. We use retention as the evaluation metric. If the retention of experiment group is significantly higher than the retention of control group, it will be verified that sending the daily reminder effectively reduces the number of students leaving the free trial due to falling behind.