UDACITY A/B TESTING

Experiment Design

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

Invariant metrics

- Number of cookies: random assigned, as the unit of diversion is cookies, the users are evenly split between the two groups. This is a good invariant metric.
- Number of clicks: happens before the free-trial screener is triggered so likely not affected
- Click-through probability: calculated based on 2 invariant metrics (number of clicks and number of cookies)

Evaluation metrics

- Gross conversion:
  This metric helps test the effect of the free trial screener on number of enrollment in the free trial. After a user clicks and gets exposed to the screener, how many of them will continue to enroll. This proportion is expected to reduce, because number of students who click 'start' is the same but those who enroll will reduce as the screener is intended to screen out potential frustrated students.
- Net conversion:
  This metric helps test the effect of the free trial screener on number of payment after the free trial. After a user passes the screener and decided to enroll, how many will stay until the end of free trial and actually make payment. This proportion is expected to not significantly reduce, because the screener should only screen out potential frustrated students only, and not the students who likely to finish the free trial.

The hypothesis of the experiment

- $H_{o(1)}$: the proportion of students who continue to enroll is the same in control and experiment group. Gross conversion is the same.
  $H_{A(1)}$: the proportion of student who continue to enroll will be different in experiment group as compared to in control group. Gross conversion is different between 2 groups.
- $H_{o(2)}$: the proportion of students remain enrolled past the 14-day boundary (and thus make at least one payment) remain the same in control and experiment group. Net conversion is the same
  $H_{A(2)}$: the proportion of students remain enrolled past the 14-day boundary (and thus make at least one payment) will be different in experiment group. Net conversion is different between 2 groups.

The experiment should only be launched with 2 conditions. The first condition is even if the number of enrollment (gross conversion) in free trial does decrease (reject $H_{o(1)}$), there will be no change in net conversion rate (fail to reject $H_{o(2)}$). The second condition is to look at the confidence interval of both metrics and see if they satisfy the statistical and practical significance.

Other metrics

- Number of user-ids: this is only tracked when the users enroll in the trial. Before they enroll or if they do not enroll, their user-ids are not tracked. If the trial screener reduces the number of users who enroll, number of user-ids cannot be an invariant metric. Although number of user-ids is highly correlated to gross conversion, which measure the number of enrollment, number of user-ids is more prone to randomness when assigned to control and experiment group. As the 2 metrics are correlated, keeping one is redundant and we should prefer the one that is more generalized (which is gross conversion)
- Retention: this proportion is calculated via number of user-ids so it cannot be invariant metric. On the other hand, it tracks the performance via the "number of payment" on "the number of user-ids to complete check-out" which have passed the screener stage and doesn't cover the period of interest (clicking – enrollment, clicking – payment). As the result, I don't want to choose Retention as evaluation metric. Moreover, "the number of user-ids to complete check-out" is unit of analysis while the unit of diversion is unique cookies, the analytic variance will be much different from empirical variance if we use Retention as evaluation metric.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

Standard deviation: analytic estimate

- Gross conversion: 0.0202
- Net conversion: 0.0156

The analytic estimate is expected to be comparable to the empirical variability as in both of the evaluation metrics above, both unit of analysis and unit of diversion are cookies. When unit of analysis is also unit of diversion, variability tends to be lower and closer to analytical estimate

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. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

I didn't use the Bonferroni correction in my analysis. Therefore the significant level for both gross conversion and net conversion metric is 0.05. the number of pageviews needed to give enough power to each of evaluation metric:

Probability of enrolling, given click (gross conversion)

20.625% base conversion rate, 1% min d.
Samples needed: 25,835

Pageviews: 25,835/(3200/40000) = 322938
Number of pageviews needed for both groups: 322938*2 = 645875

Probability of payment, given click (net conversion)
10.93125% base conversion rate, 0.75% min d.
Samples needed: 27,413
Pageviews: 27,413/(3200/40000) = 342663
Number of pageviews needed for both groups: 342663*2 = 685325

To give enough power for both evaluation metrics, the number of pageviews needed should be **685326.**

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. (These should be the answers from the "Choosing Duration and Exposure" quiz.)

I would divert **70%** of the traffic to the experiment, equivalent to the length of **25 days (685325/40000*0.7)** which is reasonable enough. The experiment is not extremely risky because it doesn't affect enrolled students, and it's also simple to implement. However, given that the experiment may possibly stop some potential students who likely to make payment in the future from trying the free trial, expose the whole traffic to the experiment will reduce revenue of the site if the experiment turns out to have a negative effect.

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. (These should be the answers from the "Sanity Checks" quiz.)

	control pageviews	experiment pageviews	control clicks	experiment click
p-hat	0.5006	0.4994	0.5005	0.4995
SE $\sqrt{\frac{0.5*0.5}{no\ of\ pageviews}}$	0.0006		0.0021	
Margin of error (m)	0.0012		0.0041	
Upper bound (control group): $0.5 + \mathrm{m}$	0.5012		0.5041	
Lower bound (control group): $0.5 - m$	0.4988		0.4959	

Number of cookies: **[0.4988, 0.5012]** with observed value **0.5006** => PASS sanity check
Number of clicks**: [0.4959, 0.5041]** with observed value **0.5005** => PASS sanity check

	CTP control	CTP experiment
p hat	0.0821	0.0822
difference (d)		0.0001
p-hat pool $\frac{X1+X2}{N1+N2}$		0.0822
SE pool $\sqrt{\frac{p\ hat\ pool*(1-p\ hat\ pool)}{\frac{1}{N1}+\frac{1}{N2}}}$		0.0007
margin of error (m) $1.96*SE$		0.0013
upper bound $+m$		0.0013
lower bound $-m$		-0.0013

Click-through-probability: **[-0.0013, 0.0013]** with observed value **0.0001** fall within the Confidence Interval range => PASS sanity check
As all three invariant metrics have passed the sanity check, we can continue to perform the Result Analysis (Effect size test and sign test)

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. (These should be the answers from the "Effect Size Tests" quiz.)

	enrollment	payments
control (X cont)	3785	2033
experiment (X exp)	3423	1945
N cont	17293	<= the number of clicks up to day 24th
N exp	17260	<= the number of clicks up to day 24th

	Gross conversion	Net conversion
p pool	0.2086	0.1151
SE pool	0.0044	0.0034
p control (gross conversion)	0.2189	0.1176
p experiment (net conversion)	0.1983	0.1127
difference	-0.0206	-0.0049
margin	0.0086	0.0067
Upper bound	-0.0120	0.0019
Lower bound	-0.0291	-0.0116

- Gross conversion Confidence Interval: **[-0.0291, -0.0120].** The confidence interval does not include 0, the metric is statistically significant with 95% confidence. The confidence interval does not include the practical significance boundary (d-min = -0.01), so the metric is practically significant
- Net conversion Confidence Interval: **[-00116, 0.0019].** The confidence interval includes 0, the metric is NOT statistically significant. The confidence interval includes the practical significance boundary (d-min = -0.0075), so the metric is NOT practically significant

(Note: d is calculated by (p-hat exp – p-hat control). Therefore, d carries negative value. Bonferroni correction is not used in this calculation, alpha = 0.05 is used for both Confidence Interval to determine margin of error)

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. (These should be the answers from the "Sign Tests" quiz.)

Gross conversion

- Number of "successes": 19
  Number of trials (or subjects) per experiment: 23
- Sign test. If the probability of "success" in each trial or subject is 0.500, then:
  The one-tail P value is 0.0013. This is the chance of observing 19 or more successes in 23 trials.
  The two-tail P value is **0.0026**. This is the chance of observing either 19 or more successes, or 4 or fewer successes, in 23 trials.

That the two-tail P value is much smaller than alpha level indicates that the result is statistically significant which mean that gross conversion in experiment group is different from that of control group (reject $H_{o(1)}$)

Net conversion:

- Number of "successes": 13
  Number of trials (or subjects) per experiment: 23
- Sign test. If the probability of "success" in each trial or subject is 0.500, then:
  The one-tail P value is 0.3388. This is the chance of observing 13 or more successes in 23 trials.
  The two-tail P value is **0.6776.** This is the chance of observing either 13 or more successes, or 10 or fewer successes, in 23 trials.

The two-tail P value is much larger than alpha level, which indicates that the result is statistically insignificant, which also means that net conversion in experiment group is not significantly different from that of control group (fail to reject $H_{o(2)}$)

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.

There are no discrepancy between the hypothesis test and the sign test.

I did not use the Bonferroni correction in my analysis because:

(1) The purpose of Bonferroni correction is to reduce the likelihood of Type I error (or false positive) when the null hypothesis (H_o) is true, but is rejected. However by doing that, this method is more prone to Type II error when the null hypothesis is false, but erroneously fails to be rejected. Consequently, applying Bonferroni will make it harder to detect the true effect (reject the null). Considering our experiment purpose, we want to test whether the net conversion is indeed NOT different between two groups (do not reject H_o). In such case, we don't want to fall into Type II error of failure to reject H_o, which may mislead us to launch the experiment when in fact net conversion rate is actually different between 2 groups

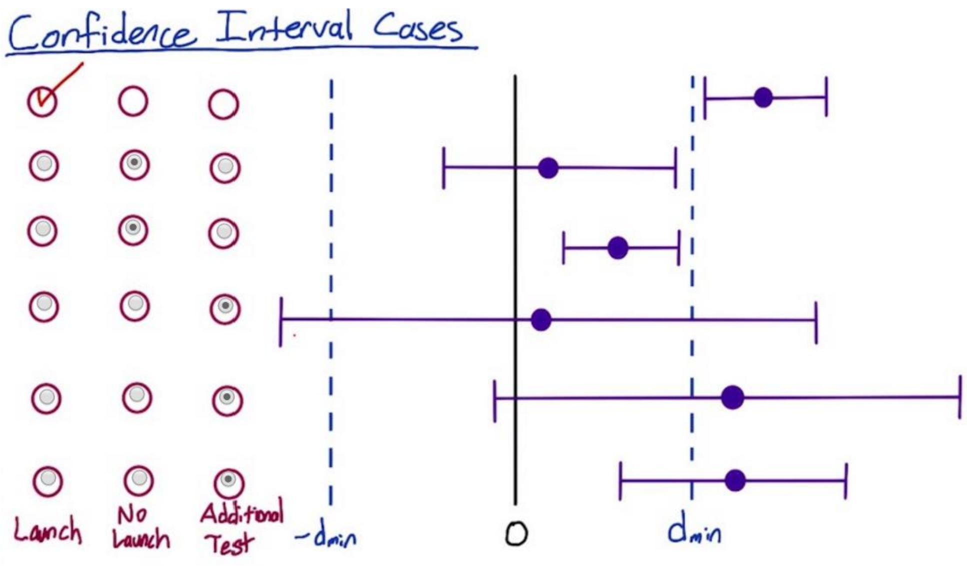
(2) We have to assume the metrics in our test should be independent to apply the Bonferroni correction. If they are not independent, the method will become too conservative. Consider our experiment, there are more or less positive dependencies between Gross Conversion and Net Conversion (the direction of one metric suggests a correlated metric to be in the same direction). That being said, we expect an "AND" scenario in our test: *both* evaluation metrics should be relevant to match the hypothesis and meet significant in their p-value; this is reasonable since the metrics are expected to move together. Both the "AND" condition and Bonferroni correction are conservative in the same way, hence combining them ends up making it too hard to detect true positives. In fact, even if we don't use the Bonferroni in this analysis, we encounter a rare event that although Gross Conversion is significant with 95% confidence, the Net Conversion is not.

Given that we don't want high Type II error in this particular test and that an "AND" scenario is used as well as the independency assumption doesn't hold, I would say that Bonferroni correction is not necessary and preferable in this experiment.

Recommendation

Make a recommendation and briefly describe your reasoning.

I recommend to hold the launch and perform additional test because: Although the effect size test and sign test prove the hypothesis that there is a significant change in gross conversion and insignificant change in net conversion, which seems to be the experiment's expected result; there are negative values smaller than –d-min in the confidence interval for net conversion, which poses a risk that the introduction of the trial screener may actually lead to a decrease in number of payment. If we look at the Confidence Interval in comparison with d-min, we see that the metric is both statistically and practically insignificant and slightly fall in the situation number 4 below, where we should consider an additional test with greater power



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.

To reduce the number of frustrated students who cancel the course early, we may want to look at the source of such frustration. Possibly, there are 4 major ones for online classroom:

1. <u>Communication and instruction</u>: Lack of intermediate feedbacks from instructor. Students don't know how to contact personal coach to address concerns in person. Form of communication is not highly effective, mostly through discussion forum, and the answers on forum may also vague.
2. <u>Ambiguous instruction and course content</u>: The course content is confusing, too much materials require extensive reading and note taking. Instructions for the assignments are vague. Sometimes students did not understand what the instructor's expectations were.
3. <u>Student skill levels do not match the requirement for the course:</u> If a beginner enrolls in an intermediate course without any realistic expectation, it would be a problem. They will feel discouraged by all the tedious study they have to do by themselves.
4. <u>Technical problems without technical support</u>

There are several solutions to above problems. For example, we can assign each student to a personalized coach to provide constant support, or a "contact your coach" button shown in each lecture helps student keep in touch with their contact point. To deal with technical problem, we can provide students with technical handbook. To reduce ambiguous course content, we can attach document including notes and reading materials so busy students can have a summary on hand and be clear about the course instruction when they listen to course videos. To match student skill level, a trial screener asking simple skill-based question then redirect students to take pre-requisites before they are prepared for the course may help.

From my own experience, I believe that Udacity can reduce frustrated students if the courses are designed more thorough and the information is provided consistently and clearly. The videos provide good information, but too much of information that students feel overflown. After finishing all the videos, they don't really remember much from the lectures. Some beginning courses of the program do provide notes, but most subsequent courses are lacking of notes. As the lecture goes on quickly, a provided note will help student keep track with essential knowledge from the class without ploughing through the videos over and over again. The note can include: expectation in the course, lecture in a comprehensive format and related reading materials

I expect to carry on an experiment where a note is included in every lesson of the course students take during the trial period.

- The **null hypothesis** is that there is no change in Retention Rate (number of user-ids to remain enrolled past the 14-day boundary and make payment divided by number of user-ids to complete checkout)
- The **alternate hypothesis** is that Retention Rate will be different between experiment and control group.
- **Unit of diversion** is user-ids as the change is tracked after a trial account is created.
- **The invariant metric** will be Number of user-ids as it's also the diversion metric
- **The evaluation metric** will be Retention Rate.
- As the experiment is low risk, we can divert the whole traffic towards the experiment, given that no experiment is employed at the same time.

If our expectation hold true, including note will help reduce students' frustration, and hence increase the Retention Rate in experiment group.

REFERENCE

Udacity forum: https://discussions.udacity.com/t/when-to-use-bonferroni-correction/37713/10
What's wrong with Bonferroni adjustments:
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1112991/
Student experience in online course:
http://www.ctdlc.org/resourcedocs/evaluation/studentexperience.pdf