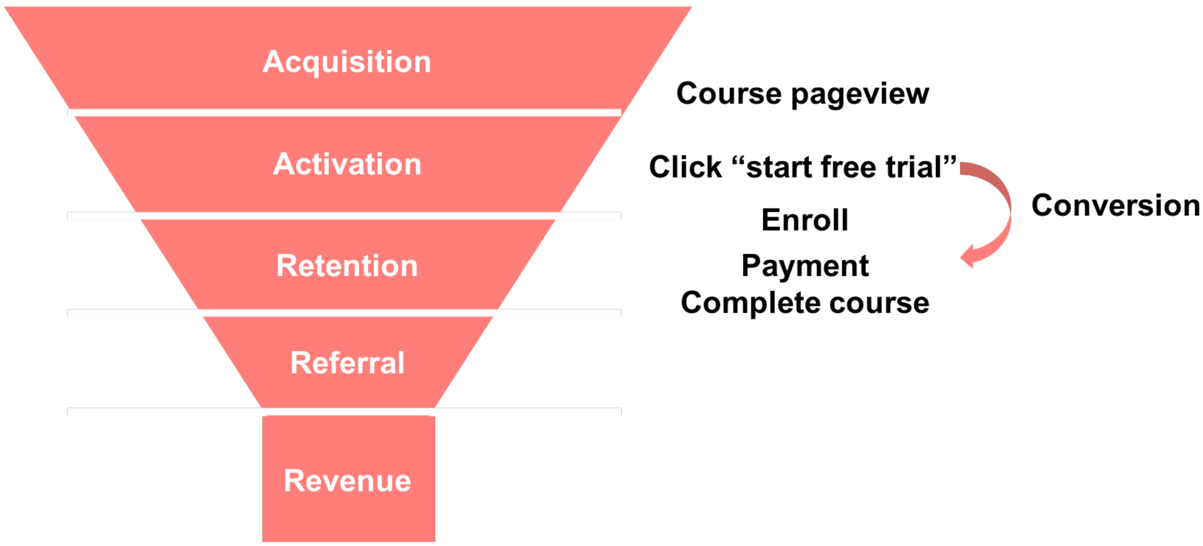
AB Test Final Project Instructions

Experiment Overview: Free Trial Screener



Metric Choice

Evaluation metrics

Gross conversion

The hypothesis is by adding question on time commitment, students spend less time will choose free materials over free trial and less students will quit because of frustration. So, the number of user-ids to complete check out and enroll the free trial is expected to decrease if hypothesis tested true.

Net conversion

The net conversion rate is not expected to decrease based on the prediction that students spend less time study tend to quit after free not but not students spend enough time, so the net conversion should stay the same.

Invariant metrics

Number of cookies

The experiment only affect behavior after "start free trial" button clicks, so number of cookies before that will not changes.

Number of clicks

Same as above, the clicks happen upstream of experimental changes.

click-through-probability (CTP)

CTP is calculated from invariant metrics above.

Metrics not used
Number of user-ids
The number is not directly associated with our goal, and is reflected in gross conversion.

Retention
While retention is a direct metrics in the funnel model, it takes a long time to run the experiment: we would need over 6 million total pageviews to achieve efficient power, which takes over 150 days even with 100% traffic. Retention evaluation will be a low priority in this AB test design [1].

Measuring Variability

Unique cookies to view course overview page per day:	40000
Unique cookies to click "Start free trial" per day:	3200
Enrollments 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

This spreadsheet contains rough estimates of the baseline values for these metrics (again, these numbers have been changed from Udacity's true numbers).
For each metric you selected as an evaluation metric, estimate its standard deviation analytically. Do you expect the analytic estimates to be accurate? That is, for which metrics, if any, would you want to collect an empirical estimate of the variability if you had time?

N = number of cookies (pageview) * CTP = 3200
Gross conversion
p = Probability of enrolling, given click = 0.20625

$$\mathrm{sd} = \sqrt{\mathrm{p} * \frac{(1-\mathrm{p})}{\mathrm{N}}} = \sqrt{0.20625 * \frac{(1-0.20625)}{3200}} = 0.0072$$

Net conversion
p = Probability of payment, given click = 0.1093125
sd = 0.0055

Both gross and net conversion are both unit of analysis and unit of diversion, thus the estimate is comparable to empirical estimation of variance.

Sizing

Choosing Number of Samples given Power

Using the analytic estimates of variance, how many pageviews **total** (across both groups) would you need to collect to adequately power the experiment? Use an alpha of 0.05 and a beta of 0.2. Make sure you have enough power for **each** metric.
Using tools by https://www.evanmiller.org/ab-testing/sample-size.html

Gross conversion

Baseline conversion rate: 20.625%
Minimum Detectable Effect: 1%
Sample size: 25835
Number of cookies (pageview) = number of clicks / CTP * 2 groups = 25835/0.08 *2 = 645875

Net conversion
Baseline conversion rate: 10.93125%
Minimum Detectable Effect: 0.75 %
Sample size: 27413
Number of cookies (pageview) = number of clicks / CTP * 2 groups = 27413/0.08 *2 = 685325

Thus, we need at lease 685325 cookies to achieve enough power.

Choosing Duration vs. Exposure

What percentage of Udacity's traffic would you divert to this experiment (assuming there were no other experiments you wanted to run simultaneously)? Is the change risky enough that you wouldn't want to run on all traffic?
Given the percentage you chose, how long would the experiment take to run, using the analytic estimates of variance? If the answer is longer than a few weeks, then this is unreasonably long, and you should reconsider an earlier decision.

Although it's common to start with a small percent (1%, 10%) of traffic and expand to a bigger percentage in AB test, we might start with 50% or greater because:
1/ the experiment is not risky, it is unlikely to annoy users or cause any negative effect
2/ there are not technical difficulties that might hurt user experience
3/ with more traffic daily can reduce time of running, and speed up the test cycle iteration
Days required for test = 685325 * 50% / 40000 = 9

Analysis

Sanity Checks

Start by checking whether your invariant metrics are equivalent between the two groups. If the invariant metric is a simple count that should be randomly split between the 2 groups, you can use a binomial test as demonstrated in Lesson 5. Otherwise, you will need to construct a confidence interval for a difference in proportions using a similar strategy as in Lesson 1, then check whether the difference between group values falls within that confidence level.
If your sanity checks fail, look at the day by day data and see if you can offer any insight into what is causing the problem.

95% confidence interval, z score = 1.96
Probability of a cookie in the control group p = 0.5

Number of cookies
Total number of cookies in control = N(cookies_ctl) = 345543
Total number of cookies in experiment = N(cookies_exp) = 344660
Total number of cookies = N(cookies_total) = 690203
$\mathrm{sd} = \sqrt{\mathrm{p} * \frac{(1-\mathrm{p})}{\mathrm{N}}} = \sqrt{0.5 * \frac{(1-0.5)}{690203}} = 0.0006$

margin of error m = z score * sd = 0.0006 * 1.96 = 0.0012
lower bound = p – m = 0.4988
upper bound = p + m = 0.5012
Observed fraction of cookies in control groups = N(cookies_ctl)/N(cookies_total) = 345543/690203= 0.5006
The value 0.5006 is in 95% confidence interval [lower= 0.4988, upper= 0.5012], thus number of cookies passes sanity check.

(Apply the same formula above in ab_test_design.R)

Number of clicks
Total number of clicks in control = 28378
Total number of clicks in experiment = 28325
Total number of cookies = 56703
sd = 0.0021
margin of error m = 0.0041
lower bound = 0.4959
upper bound = 0.5041
Observed fraction of clicks in control = 0.5005
The value 0.5005 is in 95% confidence interval, thus number of clicks passes sanity check.

CTP
CTP_ctl = N(clicks_ctl)/N(cookies_ctl) = 28378/345543 = 0.0821
CTP_exp = N(clicks_exp)/N(cookies_exp) = 28325/346660 = 0.0822
p = (N(clicks_ctl) + N(clicks_exp)) / (N(cookies_ctl) + N(cookies_exp)) = (28378+28325)/(345543+344660) = 0.0822

$$\mathrm{sd} = \sqrt{\mathrm{p} * (1 - \mathrm{p}) * (\frac{1}{\mathrm{N(cookies\_ctl)}} + \frac{1}{\mathrm{N}(cookies\_exp)})} = 0.00066$$

margin of error m = 0.0013
lower bound = -0.0013
upper bound = 0.0013
Observed difference of CTP = CTP_exp – CTP_ctl = 0.001
The value 0.001 is in 95% confidence interval, thus CTP passes sanity check.

Check for Practical and Statistical Significance

Next, for your evaluation metrics, calculate a confidence interval for the difference between the experiment and control groups, and check whether each metric is statistically and/or practically significance. A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident there was a change), and it is practically significant if the confidence interval does not include the practical significance boundary (that is, you can be confident there is a change that matters to the business.)
If you have chosen multiple evaluation metrics, you will need to decide whether to use the Bonferroni correction. When deciding, keep in mind the results you are looking for in order to launch the experiment. Will the fact that you have multiple metrics make those results more likely to occur by chance than the alpha level of 0.05?

Gross conversion (Sat, Oct 11 to Sun, Nov 2)
p = (N(enroll_ctl) + N(enroll_exp)) / (N(clicks_ctl) + N(clicks_exp)) = (3785+3423)/(17293+17260) = 0.2086

$$sd = \sqrt{p * (1 - p) * (\frac{1}{N(\text{clicks\_ctl})} + \frac{1}{N(clicks\_exp)})} = 0.0044$$

margin of error m = 0.0086
p(enroll_ctl) = N(enroll_ctl)/N(clicks_ctl) = 0.2189
p(enroll_exp) = N(enroll_exp)/N(clicks_exp) = 0.1983
p(diff) = p(enroll_exp) - p(enroll_ctl) = -0.0206
lower bound = -0.0291
upper bound = -0.0120
dmin = 0.01
As the confidence interval does not include 0 and does not include dim boundary, the decrease in gross conversion is both statistically and practically significant. We can safely reject the null hypothesis that there is no difference between the 2 groups.

Net conversion (Sat, Oct 11 to Sun, Nov 2)
p = (N(pay_ctl) + N(pay_exp)) / (N(clicks_ctl) + N(clicks_exp)) = (2033+1945)/(17293+17260) = 0.1151

$$sd = \sqrt{p * (1 - p) * (\frac{1}{N(\text{clicks\_ctl})} + \frac{1}{N(clicks\_exp)})} = 0.0034$$

margin of error m = 0.0067
p(pay_ctl) = N(pay_ctl)/N(clicks_ctl) = 0.1176
p(pay_exp) = N(pay_exp)/N(clicks_exp) = 0.1127
p(diff) = p(pay_exp) - p(pay_ctl) = -0.0049
lower bound = -0.0116
upper bound = 0.0019
dmin = 0.0075
As the confidence interval include 0 but include dim boundary, the decrease in gross conversion is not statistically nor practically significant. We can accept the null hypothesis that there is no difference between the 2 groups.

Run Sign Tests

For each evaluation metric, do a sign test using the day-by-day breakdown. If the sign test does not agree with the confidence interval for the difference, see if you can figure out why.
Using tools by https://www.graphpad.com/quickcalcs/binomial1.cfm

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 two-tail P value is 0.0026
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 two-tail P value is 0.6776

The decrease in gross conversion is statistically significant while the change in net conversion is not.

Make a Recommendation

Finally, make a recommendation. Would you launch this experiment, not launch it, dig deeper, run a follow-up experiment, or is it a judgment call? If you would dig deeper, explain what area you would investigate. If you would run follow-up experiments, briefly describe that experiment. If it is a judgment call, explain what factors would be relevant to the decision.

		state of nature	
		H0 true	H1 true
	accept H0	True negative	False negative (fail to reject, type II error)
action	reject H0	False positive (fail to reject, type I error)	True positive

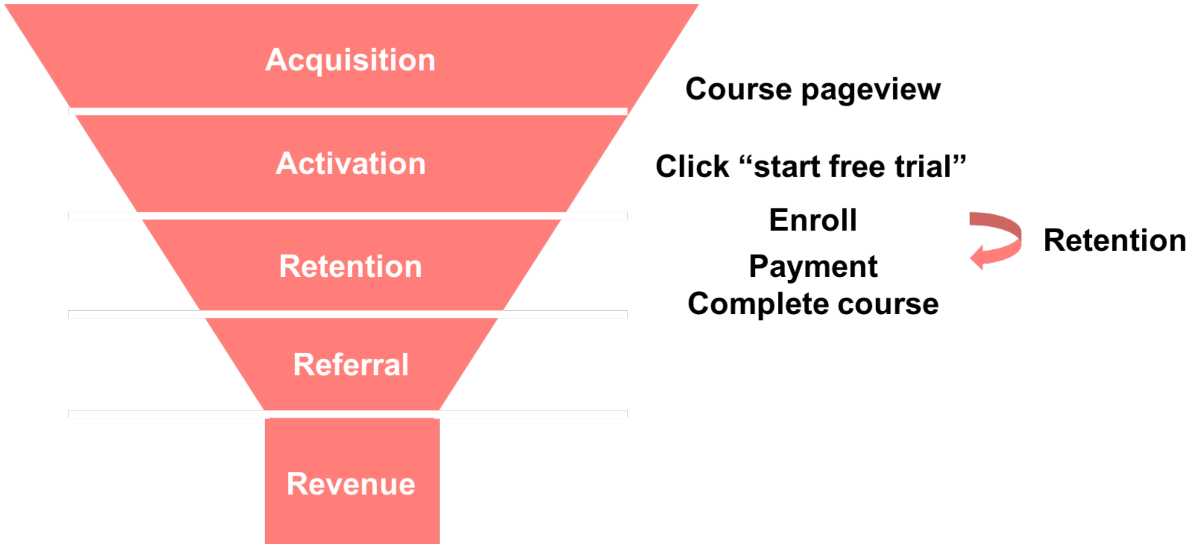
Bonferroni correction was not applied as there were not many evaluation metrics in this case. Multiple testing correction reduce type I errors at the cost of type II errors. In the case here, net conversion is a more important metrics as it is more related to the downstream of the funnel. If we falsely accepted H0 that there is no change in net conversion, we will risk the reality that it might decrease, the new version will hurt the business. So, the correction is not practical in net conversion.

Udacity should not launch the new version.

With a focus on net conversion, although we accepted H0 that there is no change in experiment and control groups, the 95% confidence interval include practical significant boundary on the negative side, suggesting that net conversion might decrease and hurt business. We should further check 99% intervals.
This could be explained if students spend less time change their behavior after enrollment and spend more time study and eventually complete course. The new version will drive away such students and end up with decreased conversion.

Follow-Up Experiment: How to Reduce Early Cancellations

We want to study the behavior and features of early cancellations by survey to get users' demographical and phycological information. We start with the assumption that student cancel early is because of financial reason that they feel subscription cost is too high. We want to offer students selected financial reason a 10% discount for the first payment cycle.



Survey design
We want to first get students’ age, education level, family education level, job income, etc. in addition to the survey specific for this retention goal:

financial	Fee is too high
	I need more time to finish (longer time cost more)
knowledge	Classes are of poor quality
	Classes are too difficult, I need more prerequisite knowledge

Hypothesis
H0: no change in retention
H1: increased retention rate
Students received discount are more likely to complete first payment.

Evaluation metrics
Retention: number of user-ids remained enrolled past 14-day boundary (and thus make at least one payment) divided by the number of user-ids to complete checkout.
(other downstream metrics should be monitored as well, as cancelation might happen later when fee returned to normal)

References
1/ Hacking Growth: How Today's Fastest-Growing Companies Drive Breakout Success; Morgan Brown, Sean Ellis;
https://books.google.com/books/about/Hacking_Growth.html?id=LcS_DAAAQBAJ