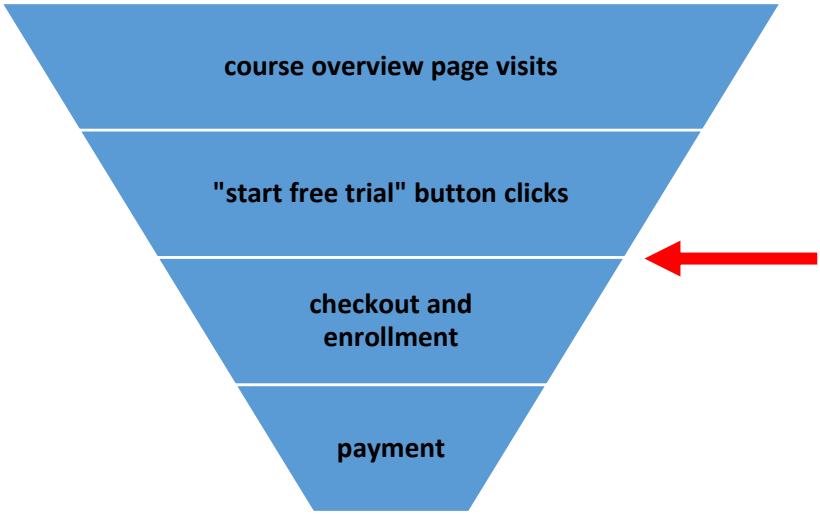
Udacity Free Trial Screener A/B Test Design

Experiment Design

Customer funnel



"In the experiment, Udacity tested a change where if the student clicked "start free trial", they were asked how much time they had available to devote to the course."

Metric Choice

Invariant metrics

- Number of cookies:
  - As shown in the customer funnel and described in the instruction, the experimental change starts after a cookie clicks "start free trial" button. Therefore, metrics before the experiment step should be invariant. Number of cookies represents the total course page visit. It is a population sizing metric and should be invariant.
- Number of clicks:
  - Same as the number of cookies, number of clicks is a population sizing metric, and invariant.
- Click-through-probability(CTP):
  - CTP derives from the ratio of two invariant metrics above, and should thus be an invariant metric.

Evaluation metrics

- Gross conversion:
  - As described in the instruction, "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 student might like to access the course materials for free. At this point, the student would have the option to continue enrolling in the free trial, **or access the course materials for free instead**." This suggests that some students, after answering

the question, will not enroll, thus the fraction of number of user-ids enrolled in the number of cookies clicking “start free trial” button (i.e. gross conversion) is expected to decrease as a result.
    - The hypothesis of this test is that adding the time commitment question after “start free trial” click “might set clearer expectations for students upfront, thus **reducing the 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.” Gross conversion therefore, can be an effective evaluation metric to test the hypothesis. **Decrease in gross conversion** suggests success of the experiment, showing the experiment reduces number of students enrolled by setting a clear expectation of time commitment.
    - The null hypothesis for gross conversion is that it does not decrease.
- Net conversion:
    - Similarly, net conversation is a good evaluation metric to test the hypothesis. Decrease in net conversion would suggest a decrease in final enrollment, thus a decrease in Udacity revenue. For Udacity to grow its business sustainably, we need to make sure the net conversion at least does not decrease, or even better increase. In this case, if gross conversion decreases and net conversion remains unchanged, it is still a success since although the final revenue for Udacity is unchanged, fewer students enroll in free trial, and less free total instruction time would be needed during the free trial. Therefore, the cost for Udacity would likely to decrease, resulting in an increase in profit for Udacity, which is desirable.
    - The null hypothesis for net conversion is that it does not decrease.

To launch the experiment, I require to reject null hypothesis for gross conversion and accept null hypothesis of net conversion. Therefore, I expect to see gross conversion decreases, and net conversion does not decrease.

Metrics not in use

- Number of user-ids:
    - As described in the instruction, the unit of diversion is a cookie, and only if the student enrolls in the free trial, user-id is used to track students from that point (enrollment) forward. Thus, the number of uses-ids represents the number of enrollments, which is expected to change in the test. It is thus not an invariant metric.
    - I did not choose it as an evaluation metric although it is expected to change and is used as numerator in gross conversion. Decrease in user-ids alone does not guarantee success because it is affected by the size of experiment on different days, which may skew the result. To be able to compare the results on different days, we need to use fraction rather than the absolute value of number of user-ids enrolled.
- Retention:

- Retention seems to be a direct metric to indicate the success of the experiment. Increase in retention would show fewer students leave free trials. However, later in the number of pageviews needed to power the experiment appropriately, I calculated the number need for retention metric and found that we would need **6,062,182** total pageviews, and even with 100% of site traffic, it still takes **152** days. It is because retention is measured from enrollment, whereas conversion metrics are measured from the click. It takes too long to run the experiment with this metric, so I decided not to use it as an evaluation metric.

Measuring Standard Deviation

N = number of sample cookies (pageview) X Click-through-probability = 5,000 X 0.08 = 400

- Gross conversion
  p = Probability of enrolling, given click = 0.20625

$$SD = \sqrt{p \times \frac{1-p}{N}} = \sqrt{0.20625 \times \frac{(1-0.20625)}{400}} = 0.0202$$

- Net conversion
  p = Probability of payment, given click = 0.1093125

$$SD = \sqrt{p \times \frac{1-p}{N}} = \sqrt{0.1093125 \times \frac{(1-0.1093125)}{400}} = 0.0156$$

Both gross conversion and net conversion use the number of cookies as denominator, which is also our unit of diversion. unit of analysis = unit of diversion. Therefore, analytical estimate is comparably to empirical estimate of variance.

Sizing

Number of Samples vs. Power

I did not use Bonferroni correction in my analysis. The metrics in the test has high correlation (covariant) and the Bonferroni correction will be too conservative to it.

alpha = 0.05
beta = 0.2
http://www.evanmiller.org/ab-testing/sample-size.html

- Gross conversion
  - Baseline conversion rate:20.625%
  - Minimum Detectable Effect: 1%
  - Number of subjects are needed for an A/B test: 25835
  - Number of pageviews (cookies) = number of button clicks / CTP = 25835/0.08 =322938

    - To run the experiment, we need the same number of pageviews in both control and experiment groups, the total number of pageviews needed to power the experiment is 322938 X 2 = **645876**
- Net conversion
    - Baseline conversion rate:10.93125%
    - Minimum Detectable Effect: 0.75%
    - Number of subjects are needed for an A/B test: 27413
    - number of pageviews (cookies) = number of button clicks / CTP = 27413/0.08 = 342663
    - To run the experiment, we need the same number of pageviews in both control and experiment groups, ths total number of pageviews needed to power the experiment is 342663 X 2 = **685626**

To make sure both evaluation metrics have the statistical power, we need at least **685626** pageviews.

Duration vs. Exposure

I decided to use **100%** of traffic to run this experiment because 1) this experiment is not highly risky for Udacity. It adds one more step on the customer funnel, which does not change the infrastructure and content profoundly 2) The chance of anyone gets hurt is extremely low because the operation does not affect existing paying customers nor highly motivated students, and we are not dealing with sensitive data in this experiment, no private information or personal data required. 3) given than same amount of cookies would be in both experiment group and control group, half of the cookies would not be affected anyway. 4) the more fraction we use, the shorter duration we need to achieve the minimal size for power. Running the experiment with shorter time would reduce operation expenses, and allows other experiments to be conducted soon at Udacity.

With 100% of traffic, we can run the experiment for as short as **20 days**.

Experiment Analysis

Sanity Checks

- Number of cookies

Total number of cookies in control groups = N(con_cookie) = 345543
Total number of cookies in experiment groups N(exp_cookie) = 346660
Total number of cookies = N(total_cookie) = N(con_cookie) + N(exp_cookie) = 691086
Expected probability of a cookie assigned in the control group p = 0.5
Standard deviation of all cookies

$$SD = \sqrt{p \times \frac{1-p}{N}} = \sqrt{0.5 \times \frac{(1-0.5)}{691086}} = 0.0006$$

With 95% confidence interval, z score is 1.96
Margin of error m = z score X SD = 0.0006 X 1.96 = 0.0012
Lower bound = p – m = 0.5 – 0.0012 = 0.4988
Upper bound = p + m = 0.5 + 0.0012 = 0.5012

Observed fraction of cookies in control groups = N(con_cookie)/N(total_cookie) = 345543/691086 = 0.5006

Observed fraction of cookies in control group 0.5006 is in 95% confidence interval [0.4988, 0.5012]. Thus number of cookies passes sanity check.

- Number of clicks

Total number of clicks in control groups = N(con_click) = 28378
Total number of clicks in experiment groups N(exp_click) = 28325
Total number of clicks= N(total_click) = N(con_click) + N(exp_click) = 56703
Standard deviation of all cookies

$$SD = \sqrt{p \times \frac{1-p}{N}} = \sqrt{0.5 \times \frac{(1-0.5)}{56703}} = 0.0021$$

With 95% confidence interval, z score is 1.96
Margin of error m = z score X SD = 0.0021X 1.96 = 0.0041
Lower bound = p – m = 0.5 – 0.0041 = 0.4959
Upper bound = p + m = 0.5 + 0.0041= 0.5041
Observed fraction of cookies in control groups = N(con_click)/N(total_click) = 28378/56703= 0.5005

Observed fraction of cookies in control group 0.5005 is in 95% confidence interval [0.4959, 0.5041]. Thus number of clicks passes sanity check.

- Click-through-probability(CTP)

CTP(con) = N(con_click)/N(con_cookie) = 28378/345543 = 0.0821
CTP(exp) = N(exp_click)/N(exp_cookie) = 28325/346660 = 0.0822
p(pool) = (N(con_click) + N(exp_click)) / (N(con_cookie) + N(exp_cookie)) = (28378+28325)/(34554+346660) = 0.0822
Standard deviation

$$SD = \sqrt{p \times (1-\mathrm{p}) \times \left(\frac{1}{N(con\_cookie)} + \frac{1}{N(\exp\_cookie}\right)}$$

$$= \sqrt{0.0082 \times (1-0.0082) \times (\frac{1}{345543} + \frac{1}{346660})} = 0.00066$$

Expected difference in CTP p(exp) = 0 when there is no difference in control and experiment groups.
With 95% confidence interval, z score is 1.96.
Margin of error m = z score X SD = 0.00066X 1.96 = 0.0013
Lower bound = p(exp) – m = 0 – 0.0013 = -0.0013
Upper bound = p(exp) + m = 0 + 0.0013= 0.0013
Observed difference CTP(diff) = CTP(exp) - CTP(con) = 0.0822 – 0.0821 = 0.0001

Observed difference in CTP 0.0001 is in 95% confidence interval [-0.0013, 0.0013]. Thus CTP passes sanity check.

So far, all invariant metrics pass sanity check, and we can proceed to result analysis.

Result Analysis

Effect Size Tests

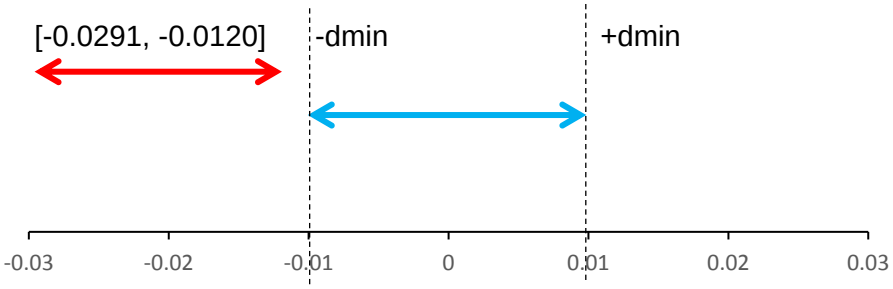
- Gross conversion
  N(con_click) = 17293
  N(con_enroll) = 3785
  N(exp_click) = 17260
  N(exp_enroll) = 3423
  p(pool) = (N(con_enroll) + N(exp_enroll)) / (N(con_click)+ N(exp_click)) = (3785+3423) /(17293+17260) = 0.2086

$$SD = \sqrt{p \times (1-\mathrm{p}) \times \left(\frac{1}{N(con\_click)} + \frac{1}{N(\exp\_click)}\right)}$$

$$= \sqrt{0.2086 \times (1-0.2086) \times \left(\frac{1}{17293} + \frac{1}{17260}\right)} = 0.0044$$

  With 95% confidence interval, z score is 1.96
  Margin of error m = z score X SD = 1.96 * 0.0044 = 0.0086

  p(con_enroll) = N(con_enroll)/ N(con_click) = 3785/17293 = 0.2189
  p(exp_enroll) = N(exp_enroll)/ N(exp_click) = 3423/17260 = 0.1983
  p(diff) = p(exp_enroll) - p(con_enroll) = 0.1983 – 0.2189 = -0.0206
  Lower bound = p(diff) – m = -0.0206 – 0.0086 = -0.0291
  Upper bound = p(diff) + m = -0.0206 + 0.0086 = -0.0120
  dmin = 0.01



  Since confidence interval [-0.0291, -0.0120] does not include 0, it is statistically significant. It also does not include dmin boundary, so the decrease in gross conversion is also practically significant. We can reject the null hypothesis of gross conversion, and therefore gross conversion decreases.

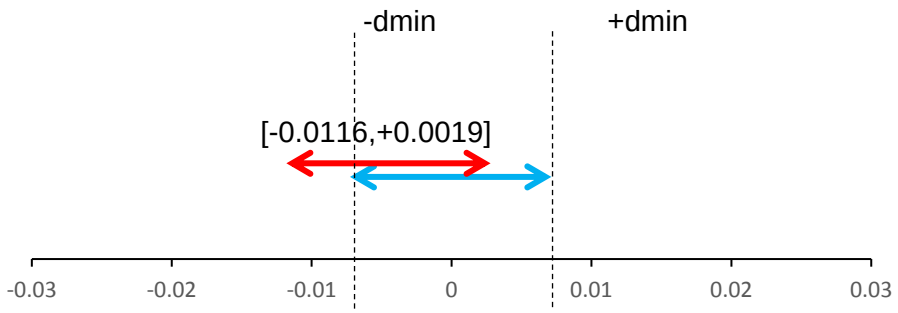
- Net conversion
  X
  N(con_click) = 17293
  N(con_pay) = 2033
  N(exp_click) = 17260
  N(exp_pay) = 1945
  p(pool) = (N(con_pay) + N(exp_pay)) / (N(con_click)+ N(exp_click)) = (2033+1945) /(17293+17260) = 0.1151

$$SD = \sqrt{p \times (1 - \mathrm{p}) \times \left(\frac{1}{N(con\_click)} + \frac{1}{N(\exp\_click)}\right)}$$

$$= \sqrt{0.1151 \times (1 - 0.1151) \times (\frac{1}{17293} + \frac{1}{17260})} = 0.0034$$

With 95% confidence interval, z score is 1.96
Margin of error m = z score X SD = 1.96 * 0.0034 = 0.0067

p(con_pay) = N(con_pay)/ N(con_click) = 2033/17293 = 0.1176
p(exp_pay) = N(exp_pay)/ N(exp_click) = 1945/17260 = 0.1127
p(diff) = p(exp_pay) - p(con_pay) = 0.1127 – 0.1196= -0.0049
Lower bound = p(diff) – m = -0.0049 – 0.0067 = -0.0116
Upper bound = p(diff) + m = -0.0049 + 0.0067 = 0.0019
dmin = 0.0075



Since confidence interval [-0.0116, +0.0019] includes 0, it is not statistically significant. It also includes –dmin boundary, so the decrease in net conversion is also not practically significant. We can accept the null hypothesis of net conversion, which is net conversion does not decrease.

Sign Tests

To check day by day data, since there are only 23 days, the data distribution is binomial. I used http://graphpad.com/quickcalcs/binomial1.cfm to calculate the p-value in sign test.

- Gross conversion
  I defined success as negative value of the difference between experiment group and control group.
  Number of "successes" you observed: 19
  Number of trials or experiments: 23
  Hypothetical probability of "success" in each trial: 0.5
  The two-tail P value: 0.0026

- Net conversion
  I defined success as negative value of the difference between experiment group and control group.
  Number of "successes" you observed: 13
  Number of trials or experiments: 23
  Hypothetical probability of "success" in each trial: 0.5
  The two-tail P value: 0.6776

  Therefore, the decrease in gross conversion is statistically significant and the change in net conversion is not.

Summary

Bonferroni correction was not used because gross conversion and net conversion are highly dependent on each other and the Bonferroni correction will be too conservative. Bonferroni controls for type I errors (false positive) at the cost of type II errors (false negative), and is applied when we expect ANY metric meet expectations. In order to launch, I require ALL metrics (gross conversion and net conversion) to match expectations (a decrease in gross conversion and no decrease in the net conversion). We are in the situation where multiple metrics need to be all matching what we expect in order to launch.

Gross conversion H0: not decrease H1: decrease (reject null)		Reality	
		H0 is true	H0 is false
Experiment	accept H0	True negative	False negative (fail to reject, type II)
	reject H0	False positive (fail to accept, type I)	**True positive (expected)**

Net conversion H0: not decrease (accept null) H1: decrease		Reality	
		H0 is true	H0 is false
Experiment	accept H0	**True negative (expected)**	False negative (fail to reject, type II)

	reject H0	False positive (fail to accept, type I)	True positive

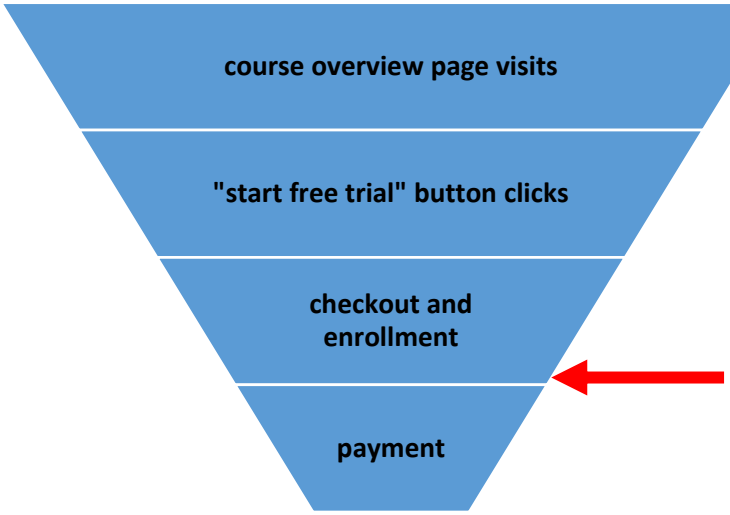
Bonferroni correction reduces type I error but increase type II error. In my metrics, it is fine for gross conversion since we expect true positive and such correction makes false positive less, thus making the result more conservative. But it is bad for net conversion since we expect true negative, and correction increases false negative, thus making the result less reliable. Net conversion is crucial for Udacity to maintain its revenue and we need to make sure we reject null properly to decide whether we want to launch the experiment or not.

Recommendation

Udacity should not launch the experiment.

Gross conversion shows a decrease in the experiment group with both statistical and practical significance, suggesting the number of users enrolled in free trial decreased. This meets my expectation. For net conversion, however, the 95 percent confidence interval [-0.0116, 0.0019] contains the practical significance boundary on the negative side (-0.075), suggesting that it is possible that net conversion may go down by an amount that would matter to the business. Since it is crucial for Udacity to maintain its business, I would choose to be a little conservative and argue it would be too risky to launch the experiment.

Follow-Up Experiment



When students choose to cancel free trial, they will be offered 10% of their first payment if their reason of cancellation is financial.

- Experiment Design

When students choose to cancel free trial, they will be asked the reason of cancellation and select one from a list of reasons including “The fee is too high” and “I don’t have time”. If the reason is financial such as “The fee is too high”, the student will be offered 10% off for their first payment. If students accept the offer, they will be charged the first payment at 10% discount. If they decline the offer, they will be the same as the control group.

- Hypothesis

H0: retention is not changed
H1: retention is increased
If students are provided 10% discount when they choose to cancel, they will be more likely to process the first payment.

- Unit of Diversion: user-id

Once enrolled, users will be tracked by user-ids, which means only users can see the experiment and make payments. We need to track individual students. Cookies or events could change over time.

- 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.