Final project A/B testing – Olga Bradford

## Experiment Overview: Free Trial Screener

At the time of this experiment, Udacity courses currently have two options on the course overview page: "start free trial", and "access course materials". If the student clicks "start free trial", they will be asked to enter their credit card information, and then they will be enrolled in a free trial for the paid version of the course. After 14 days, they will automatically be charged unless they cancel first. If the student clicks "access course materials", they will be able to view the videos and take the quizzes for free, but they will not receive coaching support or a verified certificate, and they will not submit their final project for feedback.

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. If the student indicated 5 or more hours per week, they would be taken through the checkout 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 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 screenshot](https://www.google.com/url?q=https://drive.google.com/a/knowlabs.com/file/d/0ByAfiG8HpNUMakVrS0s4cGN2TjQ/view?usp%3Dsharing&sa=D&ust=1544157880588000) shows what the experiment looks like.

The hypothesis was that this 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. 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.

**Metric Choice**

**List which metrics you will use as invariant metrics and evaluation metrics here:**

Choosing Invariant metrics:

1. Number of cookies (number of cookies to view the course overview page)(dmin=3000)

A good population size metric. Cookies should be independent of the experiment change and should be randomly assigned to each group.

1. Number of clicks (number of unique cookies to click the “start free trial” which happens before the free trial screener is trigger) (dmin=240)

A good population size metric. Click is an event, it should be independent of the experiment change and should be randomly assigned to each group.

1. 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.) (dmin=0.01)

A good invariant metric. It happened before the change triggered and should not be affected by the experiment.

Choosing Evaluation metrics:

1. Gross conversion (that is, 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”) (dmin=0.01)

A good evaluation metric. Since the one of the experiment targets was to find whether change would reduce students with not enough time left the free trail, and the number of cookies complete enrollments divided by the number of clicks would be a good evaluation. The gross conversion should decrease after the change.

The underlying assumption would be the gross conversion in the control group is higher than experiment group Therefore, it can be used as an evaluation metric to check if the experiement makes a significant difference in the enrolment.

1. Retention (that is, 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) (dmin=0.01)

A good evaluation metric. Retention has all the characteristics of an evaluation metric. It is normalized by enrollments and measures the probability to pay given enrollment. This normalized probability makes for a ready comparison between experiment and control groups and is the most direct way of answering the second part of our hypothesis. However, if we take retention as evaluate metrics, it need 2,370,010\*2 = 4,740,020 samples to get the experiment, if we put all the traffic on the experiment, it need 4,740,020/ 40,000 = 118.5 days. It took too long to get the result and should be rejected.

1. Net conversion(that is number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment, divided by the number of cookies to click the “start free trial button”) (dmin=0.0075)

A good evaluation metric. As mentioned above, the number of cookies made a payment divided by the number of clicks would be a good evaluation for whether change would not significantly reduce the number of students continue past the free trial and eventually complete the course. The net conversion should remain the same after the change.

**Measuring Standard Deviation**

**baseline.xls** spreadsheet contains rough estimates of the baseline values for these metrics (again, these numbers have been changed from Udacity's true numbers).

|  |  |  |
| --- | --- | --- |
| Unique cookies to view page per day: | 40000 | 5000 |
| Unique cookies to click "Start free trial" per day: | 3200 | 400 |
|  |  |  |
| Enrollments per day: | 660 |  |
| Click-through-probability on "Start free trial": | 0.08 |  |
| Probability of enrolling, given click: | 0.20625 | 0.02022325 |
| Probability of payment, given enroll: | 0.53 | 0.05494785 |
| Probability of payment, given click | 0.109313 | 0.0156016 |

For each metric you selected as an evaluation metric, estimate its standard deviation analytically.

See baseline.xls for calculations

| **Evaluation metrics** | **standard deviation** |
| --- | --- |
| Gross conversion=0.20625 | SE = sqrt(0.20625\*(1-0.20625)/3200) = 0.00715 (correspond to 3200 clicks & 40000 pageviews).  For 5000 pageviews, we have new\_SE = 0.00715 \* sqrt(40000/5000) = 0.0202 |
| Net conversion=0.109313 | For 40000 views SE=sqrt(0.109313\*(1-0.109313)/3200=0.005516  For 5000 views SE=0.005516\*sqrt(40000/5000)=0.015602 |
| Retention=0.53 | For 40000 views SE=sqrt((0.53\*(1-0.53)/660)=0.019427  For 5000 views=0.019427\*SQRT(40000/5000)=0.05498 |

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?

Both Gross Conversion and Net Conversion using number of cookies as denominator, which is also unit of diversion. Here, the unit of diversion is equal to unit of analysis, which indicate the analytical estimate would be comparable to the empirical variability.

For Retention, the denominator is "Number of users enrolled the courseware" which is not similar as Unit of Diversion. The unit of analysis and the unit of diversion are not the same therefore the analytical and the empirical estimates are different.

**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.)

Alpha=0.05, beta=0.2

For gross conversion:

<http://www.evanmiller.org/ab-testing/sample-size.html>

Baseline conversion rate 20.63%

Minimum detectable effect 1%

Alpha=5%, 1-betta=80%

| **Metric** | **Value** | **Std** | **Std /5000** | **Sample size** | **Click throw Probability** | **Pageviews** |
| --- | --- | --- | --- | --- | --- | --- |
| Gross conversion  Probability of enrolling, given click | 0.2063 | 0.0072 | 0.0202 | 25835 | 0.0800 | 645875 |
| Retention | 0.5300 | 0.0194 | 0.0549 | 39115 | 0.0165 | 4741212 |
| Net conversion  Probability of payment given click | 0.1093 | 0.0055 | 0.0156 | 27413 | 0.0800 | 685325 |

Therefore for highest sample size (27413), pageview/group=27413/0.08=432662.5

Total pageview=342662.5\*2=685325

Double pageview because we need total pageview for both experiments and control group.

No, I did not use Bonferroni correction during my analysis phase.

We use just 2 metrics which are highly correlate and expect that both our metrics will have statistical significance, so it will be quite conservative to use Bonferroni correction.

**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.)

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

Number of pageviews: 685325 (from previous calculations)

Fraction of traffic exposed:

What fraction of Udacity traffic would you divert to this experiment? 1

Length of experiment:

Given this, how many days will Udacity need to run the experiment: 18

I'll launch this experiment on all traffic available. This experiment is not risky for users, they are not limited in their actions, also for business it should be quite safe Another moment, that it will take too much time to run this experiment on some fraction of traffic.

I'll not use Retention as metric, because it requires 4,7 millions pageviews to achieve level of statistical significance, and experiment will take 119 days. That's too long. Without this metric experiment can be finished in **18** days.

Experiment Analysis

**Data\_all.xls - All calculations are in this file.**

**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.)

For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. **Do not proceed to the rest of the analysis unless all sanity checks pass.**

Number of cookies

From data\_all.xls file calculate total number of cookies

SUM of Control group pageviews: 345543

SUM of Experiment group pageviews: 344660

Total of both: 690203

Probability calculation = control/(control+experiment)

Probability of cookie in control or experiment group: 0.5

STDEV=sqrt(p\*(1-p)\*(1/control+1/experiment)

STDEV=sqrt(0.5\*(1-0.5)\*(1/345543+1/344660) = 0.0006018

Margin of error (m) = STDEV\*1.96=0.0011796 (95% confidence interval)

Confidence interval = [0.5-m, 0.5+m] = [0.4998,0.5012]

Observed value=344660/690203=0.5006

Number of clicks

From data\_all.xls file calculate total number of clicks

SUM of Control group clicks: 28378

SUM of Experiment group clicks 28325

Total clicks: 56703

Probability calculation = control/(control+experiment)

Probability of clicks in control or experiment group: 0.5

STDEV=sqrt(0.5\*(1-0.5)\*(1/28378+1/28325)=0.0021

Margin or error (m) = SE \*1.96 = 0.0041

Confidence Interval = [0.4959, 0.5041]

Observed value = 28378/56703 = 0.50046

Results of Sanity check: both observed values are within confidence interval, so both PASS sanity check.

**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.)

Gross Conversion (from table data\_all.xls) :

|  |  |  |
| --- | --- | --- |
|  | Control Group | Experiment |
| Clicks | 17293 | 17260 |
| Enrolment | 3785 | 3423 |
| Payments | 2033 | 1945 |
| Gross Conversion | 0.2188746892 | 0.1983198146 |
| Net Conversion | 0.1175620193 | 0.1126882966 |

Gross conversion = enrollment/clicks - control probability of enrollment given click

Net conversion = payment/clicks – control probability of payment given click

Gross conversion:

Pooled probability=(sum.clicks.control+sum.clicks.experiment)/(sum.pagev.control+sum.pagev.experi)=0.2086

SE=sqrt(pooled p\*(1-pooled p)\*(1/sum.clicks.control + 1/sum.clicks.experiment))=0.004371675385

95% confidence interval

Margin of error (m) = SE \* 1.96 =0.00856848375

D hat = p experiment (probability of enroll given click) – p control = 0.198-0.218= -0.02055

Confidence Interval = [-0.0291, -0.0120]

Outcome: statistically significant (CI doesn’t contain zero)

Practically significant (CI doesn’t contain d\_min value (d\_min=0.01)

Net Conversion:

Pooled probability=(payments\_in\_control\_group + payments\_in\_experiment\_group) /

(clicks\_in\_control\_group + clicks\_in\_experiment\_group)

Pooled probability = 0.117562019

SE=sqrt(pooled p\*(1-pooledp)\*(1/sum.payment.control+1/sum.payment.experiment))= 0.003434134

Margin of error(m)=1.96\*SE= 0.0067309

D hat = = p experiment (probability of payment given click) – p control = -0.00487372

Confidence Interval =[-0.0116045, 0.001857055]

Outcome: statistically not significant (CI contains zero)

Practically not significant (CI doesn’t contain d\_min value (dmin=0.0075)

**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:

By comparing corresponding day by day data points of enrollments between control and experiment group (in Excel, see data\_all.xls), I found out that there are only 4 days when the data in the experiment group exceeds those in the control group.

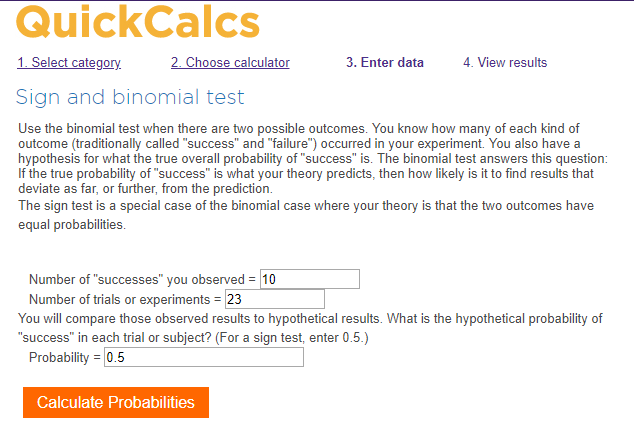
 With number of “successes” equals to 4 and number of trials equals to 23, I run the sign test by assuming the probability of “success” in each trial is 0.5, so the two-tail P value I get is 0.0026, which is the chance of observing either 4 or fewer successes, or 19 or more successes, in 23 trials. Since 0.0026 is less than the alpha (0.5), the change is statistically significant.

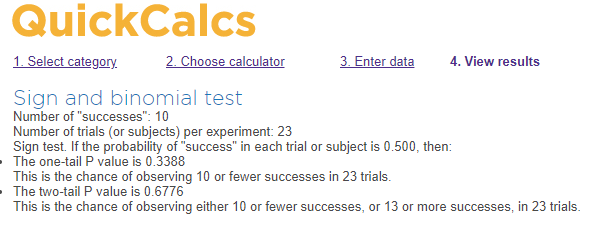
Net Conversion:

Following the same process as before, I compare the corresponding day-by-day data points of payments between the control and experiment group. The two-tail P value is 0.6776, which is the chance of observing either 10 or fewer successes, or 13 or more successes, in 23 trials. This P value is larger than the alpha (0.5), so the change is not significantly significant.

To calculate Two tailed P-value I used

<https://www.graphpad.com/quickcalcs/binomial2/>





Total Number of days: 23

|  |  |  |  |
| --- | --- | --- | --- |
|  | Days with positive change | Two tailed P-value | Significant (p < 0.025) |
| Gross Conversion | 4 | 0.0026 | YES |
| Net Conversion | 10 | 0.6776 | NO |

For Gross Conversion:

P-Value of the test is 0.0026. Since the probability to pass the test for each day is 1-0.0026 = 0.9974 which is greater than 97.5% (Due to using Bonferroni, Alpha is 0.05/2 = 0.025), this result is not happen by chance and pass the sign test.

For Net Conversion:

P-Value of the test is 0.6776. Therefore, the probability to get pass result for each day is 1-0.6776 = 0.3224 which is lower than 97.5% (Due to using Bonferroni, Alpha is 0.05/2 = 0.025). Therefore, not pass the sign test meaning that Net Conversion tends to not reduce due to apply this change.

**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.

I apply Bonferroni because these metrics might dependent since they are in the same conversion funnel. Reducing in enrolled user as a result of Gross Conversion reduction may also reduce Net Conversion.

The **result of sign test, effect size test are in alignment with hypothesis** for both metrics.

* Gross Conversion - Statistically and significantly reduction and confirmed with Sign Test.
* Net Conversion - Not statistically and significantly changes and confirmed with Sign Test.

**Recommendation**

Make a recommendation and briefly describe your reasoning.

I will recommend launching this change to all users as the testing result help confirm that this change will reduce number of users enroll in free trial and also not affect the users who enroll longer than 14 days.

Follow-Up Experiment

* Improve revenue by increasing the number of students passing free trial period by improving the expectations of contribution by adding duration that the students might take to complete the course using the estimated 5 hours/week.
  + Hypothesis - by adding duration need to complete course it will increase gross conversion and net conversion
  + Unit of diversion is user-id
  + Evaluation metrics are gross conversion, retention and net conversion
* Another experimental might be adding total hours required to completed course to improve the number of students passing free trial period. Hypothesis, unit of diversion and evaluation metrics are the same as above experiment.

**Good links for theory and calculations:**

* <http://www.had2know.com/academics/normal-distribution-table-z-scores.html>
* <https://cran.r-project.org/web/packages/tigerstats/vignettes/qnorm.html>
* <http://ncalculators.com/math-worksheets/calculate-standard-error.htm>
* <https://en.wikipedia.org/wiki/Standard_deviation>
* <http://www.had2know.com/academics/normal-distribution-table-z-scores.html>
* <http://www.evanmiller.org/ab-testing/sample-size.html>
* <http://graphpad.com/quickcalcs/binomial1.cfm>
* <http://www.stat.berkeley.edu/~mgoldman/Section0402.pdf>
* <http://www.utdallas.edu/~herve/Abdi-Bonferroni2007-pretty.pdf>