Free Trial Screener A/B Test

Udacity Data Analyst Nanodegree
By Trevor Cook

Online experimentation is an effective method to evaluate changes that may affect user experience, and can lead to better decision making within a company. A/B testing is one type of experimentation in which users are randomly diverted to different versions of a web page. By comparing the behavior of the Control group vs. the Experiment group against a set of evaluation metrics, the data can be statistically analyzed to determine whether the experiment is worth launching.

Udacity has launched an experiment that targets students who visit the home page. There are two options that students may choose from prior to accessing Udacity course materials: "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.

Students are randomly diverted to either the Control or Experiment group using a cookie, or anonymous identifier. If the student enrolls in the free trial, they will be tracked by their user ID from that point onward.

Using the data provided for the Free Trial Screener A/B test experiment run by Udacity, I have come up with a null and alternate hypothesis, chosen appropriate evaluation metrics, performed statistical analysis, and come up with a conclusion as to whether the experiment should be launched.

Experiment Design

Metric Choice

Metrics are used to measure whether the Experiment group performs better than the Control group. For each metric, the practical significance level is given as d_min. This boundary refers to the changes, in absolute terms, that would have to be observed for Udacity to consider this meaningful. The metrics that I have chosen for this experiment can be categorized as either Invariant or Evaluation metrics.

Invariant Metrics:

Invariant metrics are those that we expect to see no change across the Experiment and Control groups. Viewing the course overview page and to clicking the "Start free trial" button are events that occur before the free trial screener is triggered. The metrics relating to these events (number of cookies, number of clicks, and click-through-probability) should therefore be consistent between the two test groups.

Number of cookies: That is, number of unique cookies to view the course overview page. (d_min=3000)

Number of clicks: That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). (d_min=240)

Click-through-probability: That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page. (d_min=0.01)

Evaluation Metrics:

Evaluation metrics are more detailed and will be used to evaluate whether the A/B test resulted in any changes, given the overall business objectives of the parties running the experiment. I have chosen gross conversion and net conversion as evaluation metrics because they are recorded after the free trial screener is triggered. The numerator in both metrics is measured in number of user-ids, which are created once the user enrolls in the free trial.

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. (d_min=0.01)

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 unique cookies to click the "Start free trial" button. (d_min=0.0075)

Metrics not used in experiment:

Number of user-ids: That is, number of users who enroll in the free trial. (d_min=50)

Although the number of user-ids could be used as an evaluation metric, it is not an ideal metric for this experiment. The user-id metric is a count rather than a ratio, such as in gross conversion and net conversion. Given that the probability of being diverted to the control and experiment group is 50%, the gross conversion and net conversion metrics are more appropriate as they are normalized by the unit of their denominator being set as cookies. Therefore, I have chosen not to include the number of user-ids as a metric because it could show variability between the control and experiment groups.

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. (d_min=0.01)

Retention is being ignored as an evaluation metric because using it would require a significant amount more of page views to power the experiment. Excluding retention reduces the amount of time to complete the experiment by several months (see section on Sizing below).

Hypothesis:

The hypothesis of this experiment is that the screener will set clearer expectations for students before committing to the course. This will reduce the number of students who leave the free trial due to not dedicating enough time to the lessons, without significantly reducing the number of students who make it past the free trial and eventually complete the course.

In order to launch this experiment, I would need to see the Gross Conversion metric be reduced in the Control group by a practically significant and statistically significant margin. However, the Net Conversion metric would need to not be significantly reduced.

It is important to look at the changes in both the Gross Conversion and Net Conversion from this experiment. Although the purpose of the screener is to filter out frustrated students prior to starting the free trial, Udacity would not want to run this experiment if the result had a negative impact on the number of students who enroll and pay for the course.

Null Hypothesis:

1. The Gross Conversion proportion of the Control group is equal to the Gross Conversion proportion of the Experiment group

2. The Net Conversion proportion of the Control group is equal to the Net Conversion proportion of the Experiment group

Alternate Hypothesis:

1. The Gross Conversion proportion of the Control group is *not* equal to the Gross Conversion proportion of the Experiment group

2. The Net Conversion proportion of the Control group is *not* equal to the Net Conversion proportion of the Experiment group

Measuring Standard Deviation

The following standard deviations for the chosen metrics have been calculated analytically using a sample size of 5,000 cookies visiting the “course overview” page.

For a Bernoulli distribution with probability *P* and population *N*, the formula to calculate the standard deviation is:

$$std = \sqrt{\frac{P(1-P)}{N}}$$

Unique cookies to view 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

Estimates of Baseline Values for Metrics

Gross conversion:

Sample size = 5000
Click-through-probability on “Start free trial” = 0.08
N = 5000 * 0.08 = **400**
Probability of enrolling, given click = **0.20625**

$$std = \sqrt{\frac{0.20625(1-0.20625)}{400}}$$

$$\boldsymbol{std = 0.0202}$$

Net conversion:

Sample size = 5000
Click-through-probability on “Start free trial” = 0.08
N = 5000 * 0.08 = **400**
Probability of payment, given click = **0.1093125**

$$std = \sqrt{\frac{0.1093125(1 - 0.1093125)}{400}}$$

$$\boldsymbol{std = 0.0156}$$

For both the Gross Conversion and Net Conversion, the unit of analysis (the metric’s denominator) is measured by the number of unique cookies to click the "Start free trial" button. Since cookies are being used for both the unit of diversion as well as the unit of analysis, an empirical estimate of variability is not necessary to compute for these metrics. However, if the unit of analysis were anything other than cookies, I would want to look at the empirical variability and see how it compares to the analytic variability.

Sizing

Number of Samples vs. Power

The number of page views required to power this experiment can be calculated using an online sample size calculator (see References).

We assume for this calculation that the standard error (SE) is proportional to $1/\sqrt{N}$

Gross conversion:
α = 0.05
β = 0.2
Minimum Detectable Effect (d_min) = 0.01
Baseline Conversion Rate = 0.20625

These inputs yield a required sample size of 25,855 per variation.

Click-through-probability on “Start free trial” = 0.08
Number of variations (assume Control and Experiment are same size): 2

$$25{,}855 * \left(\frac{2}{0.08}\right) = \boldsymbol{646{,}375}$$

Net conversion:
α = 0.05
β = 0.2
Minimum Detectable Effect (d_min) = 0.0075
Baseline Conversion Rate = 0.1093125

These inputs yield a required sample size of 27,413 per variation.

Click-through-probability on “Start free trial” = 0.08
Number of variations (assume Control and Experiment are same size): 2

$$27{,}413 * \left(\frac{2}{0.08}\right) = \mathbf{685{,}325}$$

Retention:
α = 0.05
β = 0.2
Minimum Detectable Effect (d_min) = 0.01
Baseline Conversion Rate = 0.53

These inputs yield a required sample size of 39,087 per variation.

Click-through-probability on “Start free trial” = 0.08
Probability of enrolling, given click = 0.20625
Number of variations (assume Control and Experiment are same size): 2

$$39{,}087 * \left(\frac{2}{0.08 * 0.20625}\right) = \mathbf{4{,}737{,}818}$$

The number of page views required to power this experiment is **685,325**. I have excluded Retention as an evaluation metric due to the amount of time it would take to reach 4,737,818 page views.

Duration vs. Exposure

Using the total number of required page views and the average website traffic per day, I can calculate the duration of the experiment. Depending on how risky the change is, I may choose to limit exposure to the Control and Experiment groups. However, since this is not a risky experiment and Udacity is not running any other experiments during this timeframe, I would recommend diverting all traffic to the experiment. The screening process simply asks users to

indicate their estimated time commitment. This is not sensitive information that could hurt the user in any way.

Number of page views required: 685,325
Average cookies per day: 40,000
Percent of traffic diverted: 100%

$$\frac{685{,}325}{40{,}000} = 17.13$$

It would therefore take **18 days** to run this experiment.

Experiment Analysis

Sanity Checks

Before analyzing the data, it is important to perform sanity checks to ensure the experiment was run properly. These checks can be performed by looking at the invariant metrics as I expect them to be the comparable in the Control and Experiment groups.

Given:
Each cookie is assigned to the Control or Experiment with 50% probability
Assume a normal distribution
z-score (95% confidence) = 1.96

Number of cookies:

$$SE = \sqrt{\frac{0.5 * 0.5}{345543 + 344660}}$$
$$SE = 0.000601841$$

$$m = 0.000601841 * 1.96$$
$$m = 0.001179608$$

$$CI = (\mathbf{0.49882}, \mathbf{0.50117})$$

$$\hat{P} = \frac{345543}{345543 + 344660}$$
$$\hat{P} = \mathbf{0.5006}$$

Since $\hat{P}$ is within the confidence interval, the number of cookies between the Control and Experiment are not significantly different.

Number of clicks:

$$SE = \sqrt{\frac{0.5 * 0.5}{28378 + 28325}}$$
$$SE = 0.0020997$$

$$m = 0.0020997 * 1.96$$
$$m = 0.0041155$$

$$CI = (\mathbf{0.49588}, \mathbf{0.50411})$$

$$\hat{P} = \frac{28378}{28378 + 28325}$$
$$\hat{P} = \mathbf{0.500467}$$

Since $\hat{P}$ is within the confidence interval, the number of clicks between the Control and Experiment are not significantly different.

Click-through-probability:

$$\hat{P}pool = \frac{Xcont + Xexp}{Ncont + Nexp} = \frac{28378 + 28325}{345543 + 344660}$$
$$\hat{P}pool = 0.082154$$

$$SEpool = \sqrt{0.082154(1 - 0.082154)\left(\frac{1}{345543} + \frac{1}{344660}\right)}$$
$$SEpool = 0.000661$$

$$m = 0.000661 * 1.96$$
$$m = 0.001295$$

$$CI = (-\mathbf{0.001295}, \mathbf{0.001295})$$

$$\hat{d} = \hat{P}exp - \hat{P}cont = 0.0821 - 0.0821$$
$$\hat{d} = \mathbf{0}$$

Since $\hat{d}$ is within the confidence interval, the difference of proportions for click-through-probability between the Control and Experiment are not significantly different.

Result Analysis

Effect Size Tests

The following calculations give a 95% confidence interval around the difference between the experiment and control groups for the experiment evaluation metrics.

Gross conversion:
α = 0.05
β = 0.2
d_min = 0.01

$$\hat{P}pool = \frac{Xcont + Xexp}{Ncont + Nexp} = \frac{3785 + 3423}{17293 + 17260}$$
$$\hat{P}pool = 0.2086$$

$$SEpool = \sqrt{0.2086(1 - 0.2086)\left(\frac{1}{17293} + \frac{1}{17260}\right)}$$
$$SEpool = 0.00437$$

$$m = 0.00437 * 1.96$$
$$m = 0.0085684$$

$$\hat{d} = \hat{r}exp - \hat{r}cont = \frac{3423}{17260} - \frac{3785}{17293} = 0.1983198 - 0.2188746$$
$$\hat{d} = \mathbf{-0.0205548}$$

$$CI = (\mathbf{-0.0291}, \mathbf{0.0119})$$

The confidence interval does not include d_min: **Practically significant**
The confidence interval does not include zero: **Statistically significant**

Net conversion:
α = 0.05
β = 0.2
d_min = 0.0075

$$\hat{P}pool = \frac{Xcont + Xexp}{Ncont + Nexp} = \frac{2033 + 1945}{17293 + 17260}$$
$$\hat{P}pool = 0.115127$$

$$SEpool = \sqrt{0.115127(1 - 0.115127)\left(\frac{1}{17293} + \frac{1}{17260}\right)}$$
$$SEpool = 0.003434$$

$$m = 0.003434 * 1.96$$
$$m = 0.00673$$

$$\hat{d} = \hat{r}exp - \hat{r}cont = \frac{1945}{17260} - \frac{2033}{17293} = 0.11268 - 0.11756$$
$$\hat{d} = \mathbf{-0.00488}$$

$$CI = (\mathbf{-0.0116}, \mathbf{0.0018})$$

The confidence interval does include d_min: **Not practically significant**
The confidence interval does include zero: **Not statistically significant**

Sign Tests

The following calculations looks at the day-to-day differences between experiment and control groups. Using the online sign test calculator and assuming the probability of being randomly assigned to either group being 0.5, I can calculate whether the chance of observing *X* successes in *N* days is significant.

Gross conversion:

α = 0.05
Number of days: 23
Number of days with change: 19
p-value = **0.0026**

Date	GrossConversion_Cont	GrossConversion_Exp	difference (Cont - Exp)
Sat, Oct 11	0.195050946	0.153061224	-0.041989722
Sun, Oct 12	0.188703466	0.147770701	-0.040932765
Mon, Oct 13	0.183718372	0.164027149	-0.019691223
Tue, Oct 14	0.186602871	0.166868198	-0.019734673
Wed, Oct 15	0.19474313	0.168269231	-0.026473899
Thu, Oct 16	0.167679222	0.163705584	-0.003973639
Fri, Oct 17	0.195187166	0.162820513	-0.032366653
Sat, Oct 18	0.174050633	0.144171779	-0.029878854
Sun, Oct 19	0.189580318	0.172166428	-0.017413891
Mon, Oct 20	0.191637631	0.177906977	-0.013730654
Tue, Oct 21	0.226066897	0.165509259	-0.060557638
Wed, Oct 22	0.193317422	0.15980025	-0.033517173
Thu, Oct 23	0.190977444	0.190031153	-0.000946291
Fri, Oct 24	0.326894502	0.278335725	-0.048558778
Sat, Oct 25	0.254703329	0.189835575	-0.064867753
Sun, Oct 26	0.22740113	0.220779221	-0.006621909
Mon, Oct 27	0.306982872	0.276264591	-0.030718281
Tue, Oct 28	0.20923913	0.220108696	0.010869565
Wed, Oct 29	0.265223275	0.27647868	0.011255405
Thu, Oct 30	0.227520436	0.284340659	0.056820223
Fri, Oct 31	0.246458924	0.252077562	0.005618639
Sat, Nov 1	0.22907489	0.204316547	-0.024758343
Sun, Nov 2	0.297258297	0.251381215	-0.045877082
	0.218874689	0.198319815	
		# Days	23
		# Days with change	19

Since p-value < 0.05, the day-to-day difference between Control and Experiment for gross conversion is statistically significant.

Net conversion:

α = 0.05
Number of days: 23
Number of days with change: 10
p-value = **0.6776**

Date	NetConversion_Cont	NetConversion_Exp	difference (Cont - Exp)
Sat, Oct 11	0.101892285	0.049562682	0.052329603
Sun, Oct 12	0.089858793	0.115923567	-0.026064774
Mon, Oct 13	0.104510451	0.089366516	0.015143935
Tue, Oct 14	0.125598086	0.111245466	0.014352621
Wed, Oct 15	0.07646356	0.112980769	-0.036517209
Thu, Oct 16	0.09963548	0.077411168	0.022224312
Fri, Oct 17	0.101604278	0.056410256	0.045194022
Sat, Oct 18	0.110759494	0.095092025	0.015667469
Sun, Oct 19	0.08683068	0.110473458	-0.023642778
Mon, Oct 20	0.112659698	0.113953488	-0.00129379
Tue, Oct 21	0.121107266	0.082175926	0.038931341
Wed, Oct 22	0.109785203	0.087390762	0.022394441
Thu, Oct 23	0.084210526	0.105919003	-0.021708477
Fri, Oct 24	0.18127786	0.134863702	0.046414159
Sat, Oct 25	0.185238784	0.121076233	0.064162551
Sun, Oct 26	0.146892655	0.145743146	0.00114951
Mon, Oct 27	0.163372859	0.154345006	0.009027853
Tue, Oct 28	0.123641304	0.163043478	-0.039402174
Wed, Oct 29	0.116373478	0.132049519	-0.015676041
Thu, Oct 30	0.102179837	0.092032967	0.010146869
Fri, Oct 31	0.14305949	0.170360111	-0.027300621
Sat, Nov 1	0.136563877	0.143884892	-0.007321015
Sun, Nov 2	0.096681097	0.142265193	-0.045584097
	0.117562019	0.112688297	0.004873723
		# Days	23
		# Days with change	10

Since p-value > 0.05, the day-to-day difference between Control and Experiment for net conversion is not statistically significant.

Summary

There are no discrepancies between the effect size tests and the sign tests. The gross conversion metric is statistically significant while the net conversion metric is not statistically significant for both tests.

The Bonferroni correction is an adjustment made to α when several tests, such as metrics, are being performed on a single dataset. As the number of tests increases, the risk of rejecting the null hypothesis when it is true (Type 1 error) increases. For example, in an experiment using 20 metrics and $\alpha = 0.05$, it would not be unusual for one of these metrics to be statistically significant from pure chance. The Bonferroni correction is useful in situations when the decision to launch an experiment is based on *any* metric would match our expectations. In these situations, the alpha level is corrected to decrease Type 1 error.

In Udacity's screener experiment, we are looking for both metrics to match our expectations before deciding to launch. Udacity is looking to filter out frustrated students, while also maintaining the number of students who enroll and pay for the course. In an experiment when

we are looking for *all* the metrics to meet our expectations using several metrics, there is an increased risk of failing to reject the null hypothesis when the null is not true (Type 2 error). Since this is the case in our experiment, I recommend not using the Bonferroni correction.

Recommendation

After analyzing the results of this experiment, we can reject null hypothesis #1. The data collected from the free trial screener show that the difference between the Gross Conversion proportion of the Control and Experiment groups are practically and statistically significant. At the same time, however, the Net Conversion confidence interval does include the negative boundary of the practical significance level. This means that it is possible for this experiment to have caused a result in the number of enrollments that negatively impacts the business. Since one of the business objectives is to not decrease the number of students who pay for the course, I would recommend not launching this experiment. Although the screener performs well in reducing the number of frustrated students who enroll in the course, it is not worth the risk if it may impact the company's revenue.

Follow-Up Experiment

Goal: To reduce the number of frustrated students who cancel early in the course.

Description: After 7 days of being enrolled in the free trial, a "Weekly Progress Dashboard" will appear at the Udacity homepage before the user continues to the classroom. The dashboard will include an overview of their learning progress from the first week of the user's free trial. Their time spent learning will be in the form of a progress bar, which compares their total classroom time to the weekly time commitment they had initially indicated during the screener.

If a student has spent over 5 hours learning during the past week, they will be congratulated and prompted to continue to the classroom. If their total learning time is under 5 hours, they will have the option to opt out of the free trial and continue learning using the free class materials.

Hypothesis: The hypothesis of this experiment is that it will reduce the number of frustrated students who cancel early in the nanodegree program. The weekly Progress Dashboard will give students an idea of how much time they will need to dedicate to the program each week by being able to compare their past week's commitment to what is expected from them.

Metrics:

Invariant Metrics:

Number of user-ids: The number of users who enroll in the free trial.

Evaluation Metrics:

Cancellation rate: The number of users to cancel after one month of being enrolled divided by the number of users to enroll after completing the free trial.

Unit of Diversion: The unit of diversion is a user-ID as this experiment occurs after users sign up for the free trial.

References:

Sample size calculator:

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

Sign test calculator:

http://graphpad.com/quickcalcs/binomial1.cfm

Udacity forums:

https://discussions.udacity.com/t/empircal-vs-analytical/194893

https://discussions.udacity.com/t/project-r-function-for-experiment-size/198641

https://discussions.udacity.com/t/practical-significance-standard-deviation-empirical-analytical-effect-size-test/163672