# *PROJECT REPORT*

*On*

***Design an A/B Test***

*Submitted by:* ***Pankaj NATH***

*As a part of AIRBUS Data Analyst Nanodegree*

## Table of Revision

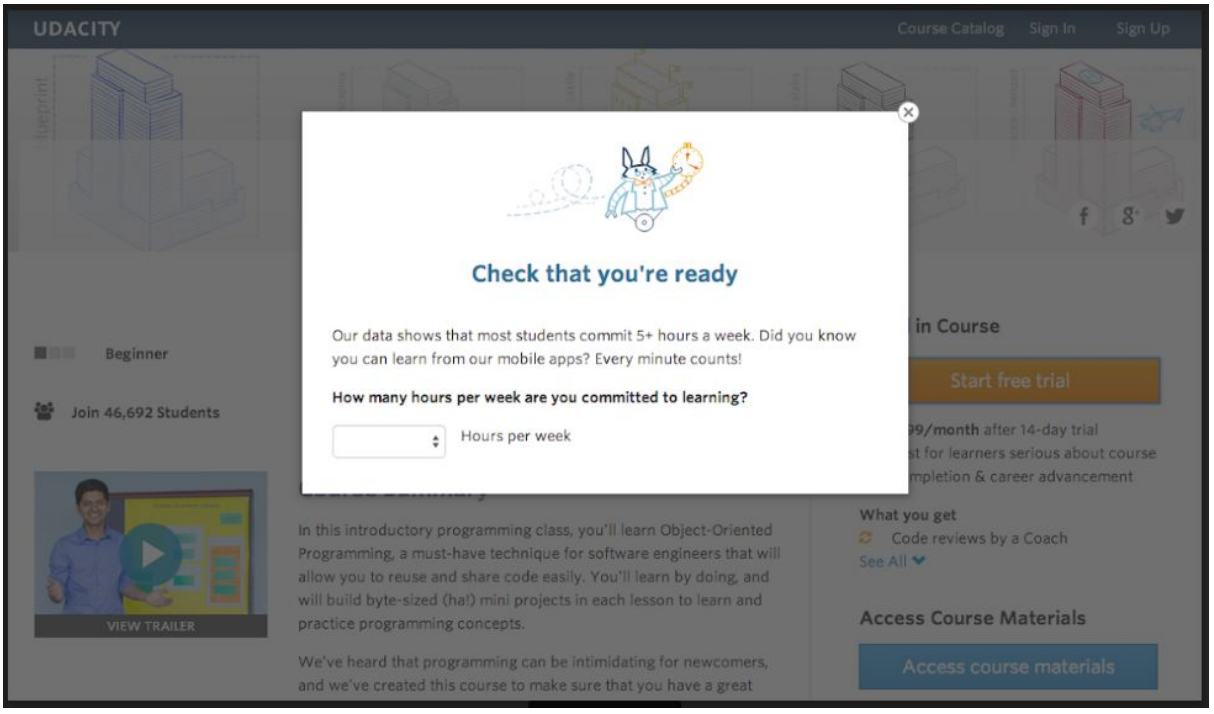
| Issue No. | Issue Date | Reason for Revision |
|---|---|---|
| 1.0 | 02nd of May, 2020 | First submission |
| | | |

# Table of Contents

## 1. 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. Below screenshot 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.

# 2. Metric Choice

## 2.1 Question-1:

List which metrics you will use as invariant metrics and evaluation metrics here.
For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

## 2.2 Answer-1:

For an experiment, two sets of different metrics are needed that measure how an experiment group is better than the control group.[1]

Metrics in the first set are called **Invariant metrics**. These metrics as their name suggests, do not change. They are used to measure the distribution of test subjects in an experiment and control group. It is intended to have similar distribution among these groups for consistency. If an invariant metrics differ in results when applied to experiment and control group then that experiment is not good to proceed and needs to be redone again.

Metrics in the second set are called **Evaluation metrics**. These metrics as their name suggests, evaluate the results of an experiment. If there is a difference in results between experiment group and control group due to the experiment, then these metrics should measure this difference and their significance as well.

For our experiment, I selected the metrics as per below table:

| Invariant Metrics | Evaluation Metrics | Un-used Metrics |
|---|---|---|
| Number of cookies<br>Number of clicks<br>Click-through-probability | Gross conversion<br>Retention<br>Net conversion | Number of user-ids |

The justification for selection of metrics is provided below:

- **Number of cookies:** That is, number of unique cookies to view the course overview page. ($d_{min}$=3000)

  This is the unit of diversion for this experiment to approximate the unique pageviews. This diversion towards experiment and control groups is on a random basis and should not vary much between both groups. For this reason it is a good candidate for an invariant metric.

- **Number of user-ids:** That is, number of users who enroll in the free trial. ($d_{min}$=50)

  As per our experiment set-up, user-ids were only tracked after enrolling into free trial, otherwise they are not tracked. For this reason its distribution would not be consistent in the experimental group and it would be difficult to compare this to the control group. So this cannot be an invariant metric.

  If we select this as an evaluation metric then it will not match with our unit of diversion (count of unique cookies). For this reason this cannot be an evaluation metric as well.

- **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)

  At this point of the funnel, the change has not happened and all users go through this step of clicking the “Start free trial” button. For this reason this metric cannot be an evaluation metric and can only be an invariant metric which should not change between experiment and control group for an even distribution.

- **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)

  This metric is a ratio of two previously identified invariant metrics, i.e. number of unique cookies clicking “Start free trial” button and number of unique cookies visiting the course overview page. For this reason it can be an invariant metric and should not vary between experiment and control groups.

- **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)

  The expectation from our experiment is that users who cannot devote the minimum hours per week will not enroll and complete the checkout. With this expectation we should see a decrease in gross conversion. For this reason it can be an evaluation metric.

- **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)

  As discussed above, since non-interested users will be filtered out in our experiment group, we expect that more users will continue within the experiment user group and will make the next payment as well. Due to this reason retention will be higher in the experiment group compared to the control group. For this it can be an evaluation metric.

- **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)

  We expect in our experiment that more users will remain enrolled past the free period in the experiment group compared to the control group. But the number of unique cookies clicking the “Start free trial” button will be ideally the same in both groups.

For this reason this can be an evaluation metric. We should see a positive trend for this metric in our experiment group.

# 3. Measuring Variability

## 3.1 Question-2:

List the standard deviation of each of your evaluation metrics.

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

## 3.2 Answer-2:

When the unit of diversion is equal to the unit of analysis, the analytical estimate of standard deviation tends to be equal to empirical estimate. Since for two of our selected evaluation metric (Gross conversion and Net conversion) the unit of diversion and the unit of analysis is equal i.e., cookies, both analytical and empirical standard deviation would be nearly equal. But for one remaining evaluation metric (Retention), the unit of diversion (cookies) is not equal to the unit of analysis (user-ids), we would not have the same analytical and empirical estimates of standard deviation. For this reason, we will calculate the standard deviation for all three evaluation metrics analytically.

For a binomial distribution with a probability of ***p*** and population ***N***, analytical standard deviation is given by $SD = \sqrt{\frac{p(1-p)}{N}}$

For given 5000 cookies per day, the fraction from the baseline data[2] is $\frac{5000}{40000}$ = 0.125.

Therefore, unique number of cookies to click "Start free trial" per day = 0.125 * 3200 = 400.

and Enrollments per day = 0.125 * 660 = 82.5.

Standard deviation for Gross Conversion = $\sqrt{\frac{0.20625\ (1-0.20625)}{400}}$ = 0.0202

Standard deviation for Retention = $\sqrt{\frac{0.53\ (1-0.53)}{82.5}}$ = 0.0549

Standard deviation for Net Conversion = $\sqrt{\frac{0.1093125\ (1-0.1093125)}{82.5}}$ = 0.0156

To summarize, below table gives the standard deviation for the evaluation metrics selected for this experiment:

| Evaluation Metric | Calculated Standard Deviation |
|---|---|
| Gross Conversion | 0.0202 |
| Retention | 0.0549 |
| Net Conversion | 0.0156 |

# 4. Sizing

## 4.1 Number of Samples vs. Power

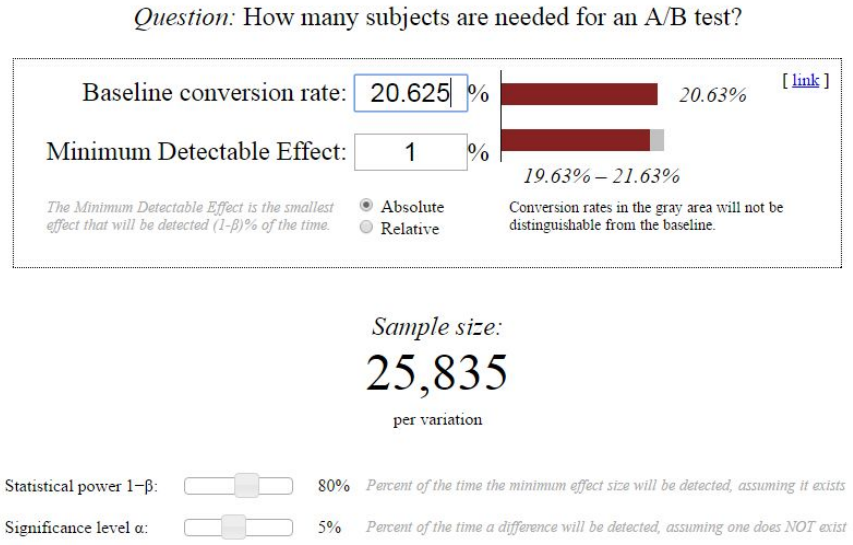
### 4.1.1 Question-3:

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power your experiment appropriately.

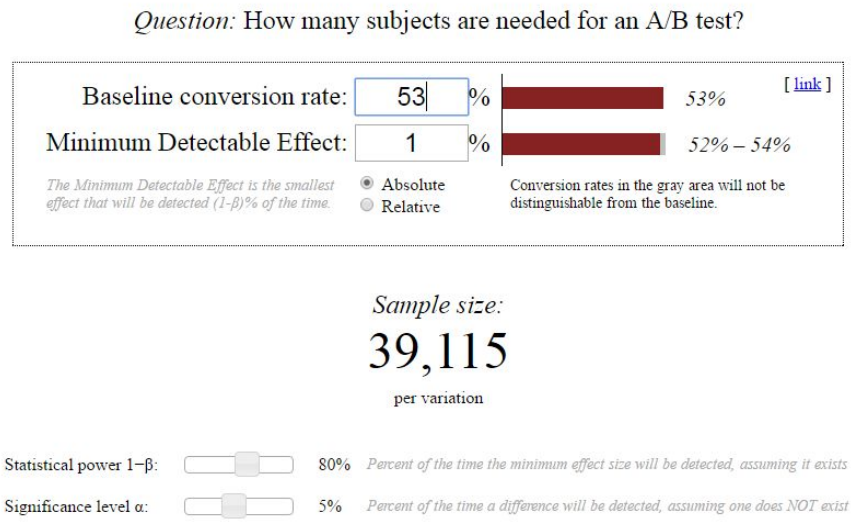
### 4.1.2 Answer-3:

The three evaluation metrics are related to each other and using Bonferroni correction is a conservative approach. For this reason I choose not to use Bonferroni correction during my analysis phase.

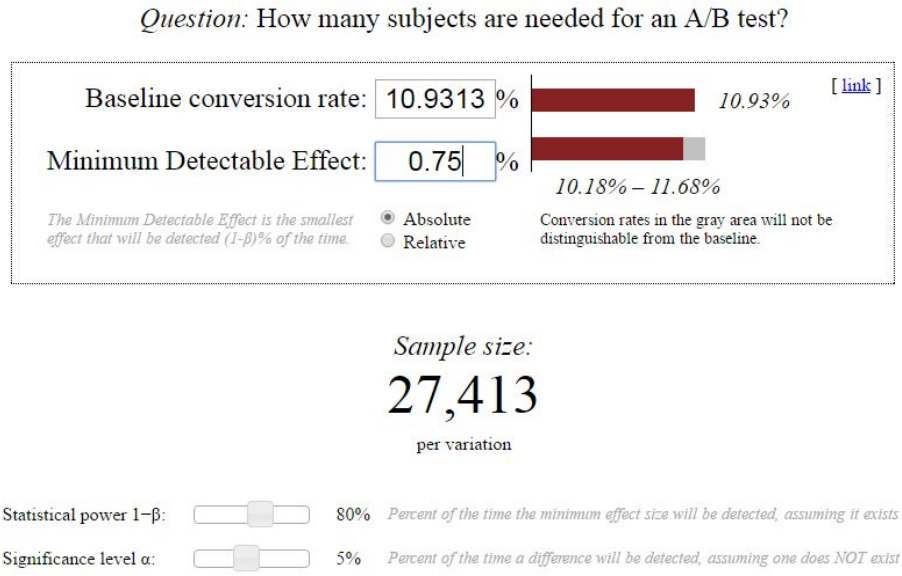
Using the online sample size calculator tool[3] with an alpha of 0.05 and a beta of 0.2, first a suitable sample size for each of the evaluation metric is calculated as below:
Sample size required for Gross Conversion metric:




Sample size required for Retention metric:



Sample size required for Net Conversion metric:




From the baseline data[2], the ratio of unique cookies to click "select free trial" per day to unique cookies to view course overview page per day is $\frac{3200}{40000}$ = 0.08.

Similarly, the ratio of enrollments per day to unique cookies to view course overview page per day is $\frac{660}{40000}$ = 0.0165.

Also note that the sample size we calculated above is for the one group say experiment group. But for our experiment we need a control group as well. For this reason, for a complete experiment we would need twice the size of the sample calculated above.

Below table summarizes the total sample size and pageviews needed for each of the evaluation metric:

| Evaluation Metric | Sample Size | Pageviews Needed |
|---|---|---|
| Gross Conversion | 25,835 * 2 = 51,670 | 51,670 / 0.08 = 6,45,875 |
| Retention | 39,115 * 2 = 78,230 | 78,230 / 0.0165 = 47,41,212.12 |
| Net Conversion | 27,413 * 2 = 54,826 | 54,826 / 0.08 = 6,85,325 |

To measure all three evaluation metrics, we would need at least a total of **47,41,212** pageviews.

## 4.2 Duration vs. Exposure

### 4.2.1 Question-4:

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment.

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

### 4.2.2 Answer-4:

With the given baseline per day pageview of 40,000 and 100% traffic diversion towards this experiment, we would need 47,41,212 / 40,000 ≅ 119 days. This time period is too long to continue any experiment. For this reason, I propose to drop Retention as an evaluation metric for this experiment.
In this new scenario, we should have at least 6,85,325 pageviews to measure our two evaluation metrics. At the rate of 100% diversion, we would need 6,85,325 / 40,000 ≅ 18 days.

Although this experiment is not too risky to divert 100% of the traffic, a business continuity sense will always say not to do that. With this thought I propose to divert only 50% of the traffic and we would be able to get 6,85,352 page views in 6,85,352 / (40,000 * 0.5) ≅ 35 days. The 5 weeks period is neither too-short nor too-long.

## 5. Experiment Analysis

The data for you to analyze is here. This data contains the raw information needed to compute the above metrics, broken down day by day. Note that there are two sheets within the spreadsheet - one for the experiment group, and one for the control group.
The meaning of each column is:

- **Pageviews:** Number of unique cookies to view the course overview page that day.
- **Clicks:** Number of unique cookies to click the course overview page that day.
- **Enrollments:** Number of user-ids to enroll in the free trial that day.
- **Payments:** Number of user-ids who who enrolled on that day to remain enrolled for 14 days and thus make a payment. (Note that the date for this column is the start date, that is, the date of enrollment, rather than the date of the payment. The payment happened 14 days later. Because of this, the enrollments and payments are tracked for 14 fewer days than the other columns.

### 5.1 Sanity Check

#### 5.1.1 Question-5:

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.
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.**

#### 5.1.2 Answer-5:

Sanity check is performed with measurement of invariant metrics where we assume that these metrics will not vary between experiment and control groups because the diversion is assumed to be done evenly thanks to random allocation (probability, $p$ = 0.5).

If sanity check fails then there is no point in analysing the experiment result because it would mean that our initial assumption itself did not hold true.

For a confidence interval (CI) of 95%, the z critical value (two tailed) from z-table[4] is 1.96 at $\alpha$ = 0.05.

Number of Cookies

Expected value for Number of unique cookies in the control group to the total count of unique cookies visiting the course overview page is 0.5.

The observed rate =
SUM(Control!B2:B38)/(SUM(Control!B2:B38)+SUM(Experiment!B2:B38)) = 0.50064.

The standard error, SE =
SQRT((0.5*(1-0.5))/(SUM(Experiment!B2:B38)+SUM(Control!B2:B38))) = 0.000602.

Margin of error, ME = z * SE = 1.96 * 0.000602 = 0.00118.

Therefore, Upper bound = SE + ME = 0.50064 + 0.00118 = 0.50118,
and Lower bound = SE - ME = 0.50064 - 0.00118 = 0.49882.

Since the expected value 0.5 is within the bounds [0.49882, 0.50118], sanity check for **Number of cookies metric** is **Passed**.

Number of Clicks

Expected value for Number of unique cookies in the control group clicking the “Start free trial” button to the total count of unique cookies clicking the “Start free trial” button is 0.5.

The observed rate =
SUM(Control!C2:C38)/(SUM(Control!C2:C38)+SUM(Experiment!C2:C38)) = 0.500467.

The standard error, SE =
SQRT((0.5*(1-0.5))/(SUM(Experiment!C2:C38)+SUM(Control!C2:C38))) = 0.0021.

Margin of error, ME = z * SE = 1.96 * 0.0021 = 0.004116.

Therefore, Upper bound = SE + ME = 0.500467 + 0.004116 = 0.504116,
and Lower bound = SE - ME = 0.500467 - 0.004116 = 0.495884.

Since the expected value 0.5 is within the bounds [0.495884, 0.504116], sanity check for **Number of clicks metric** is **Passed**.

Click-through-probability

We expect that the click-through-probability in both the groups should be near to equal. So let's observe this value in the experiment group and then check if it is matching with the control group or not?

The expected value from the experiment group =
SUM(Experiment!C2:C38)/SUM(Experiment!B2:B38) = 0.082182.

The observed value in the control group, SE =
SUM(Control!C2:C38)/SUM(Control!B2:B38) = 0.082126.

The standard error from the control group, SE = SQRT((0.082126*(1-0.082126))/SUM(Control!B2:B38)) = 0.000467.

Margin of error, ME = z * SE = 1.96 * 0.000467 = 0.000915.

Therefore, Upper bound = SE + ME = 0.082126 + 0.000915 = 0.083041,
and Lower bound = SE - ME = 0.082126 - 0.000915 = 0.08121.

Since the expected value from the experiment group, 0.082182 is within the bounds [0.08121, 0.083041], sanity check for **Click-through-probability** metric is **Passed**.

Below table summarizes the results of sanity check:

| Metrics | Expected Value | Observed Value | Upper Bound | Lower Bound | Check Status |
|---|---|---|---|---|---|
| **Number of Cookies** | 0.5 | 0.5006 | 0.4988 | 0.5012 | Pass |
| **Number of Clicks** | 0.5 | 0.5005 | 0.4959 | 0.5041 | Pass |
| **Click-through-probability** | 0.0822 | 0.0821 | 0.0812 | 0.0830 | Pass |

*NOTE: Values in the table above are rounded-up to 4-decimal places.

## 5.2 Result Analysis

### 5.2.1 Effect Size Test, Question-6:

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.

### 5.2.2 Answer-6:

Using the measures from evaluation metrics we identify if our experiment is statistically and practically significant or not. If the bounds of CI does not include zero (0) then it is said to be statistically significant. For the results to be practically significant for the business, the minimum detectable effect ($d_{min}$) shall be included inside the bounds of CI. We will now check these for both of our evaluation metrics.

For a confidence interval (CI) of 95%, the z critical value (two tailed) from z-table[4] is 1.96 at $\alpha$ = 0.05.

Gross Conversion

The pooled probability, $p$ = (SUM(Experiment!D2:D24)+SUM(Control!D2:D24)) / (SUM(Experiment!C2:C24)+SUM(Control!C2:C24)) = 0.208607.

then pooled SE = SQRT($p$*(1-$p$)*((1/SUM(Experiment!C2:C24))+(1/SUM(Control!C2:C24)))) = 0.004372.

Margin of error, ME = z * SE = 1.96 * 0.004372 = 0.008568.

The expectation from our experiment is that users who cannot devote the minimum hours per week will not enroll and complete the checkout. With this expectation we should see a decrease in gross conversion. In other words the difference between experiment and control group would be negative, or ($d_{min}$= -0.01). Lets calculate the observed difference, d =

(SUM(Experiment!D2:D24)/SUM(Experiment!C2:C24))-(SUM(Control!D2:D24)/SUM(Control!C2:C24)) = -0.02055.

Therefore, Upper bound = d + ME = -0.02055 + 0.008568 = -0.01199,
and Lower bound = d - ME = -0.02055 - 0.008568 = -0.02912.

Rounding up to 2-decimal places, 95% CI of [-0.03, -0.01] does not include zero (0) and includes $d_{min}$= -0.01, hence Gross conversion metric is both statistically and practically significant.

Net Conversion

The pooled probability, *p* = (SUM(Experiment!E2:E24)+SUM(Control!E2:E24)) / (SUM(Experiment!C2:C24)+SUM(Control!C2:C24)) = 0.115127.

then pooled SE = SQRT(*p**(1-*p*)*((1/SUM(Experiment!C2:C24))+(1/SUM(Control!C2:C24)))) = 0.003434.

Margin of error, ME = z * SE = 1.96 * 0.003434 = 0.006731.

The observed difference, d = (SUM(Experiment!E2:E24)/SUM(Experiment!C2:C24))-(SUM(Control!E2:E24)/SUM(Control!C2:C24)) = -0.00487.

Therefore, Upper bound = d + ME = -0.00487 + 0.006731 = 0.001857,
and Lower bound = d - ME = -0.00487 - 0.006731 = -0.0116.

Rounding up to 4-decimal places, 95% CI of [-0.0116, 0.0019] does include zero (0) and does not include $d_{min}$= 0.0075, hence Net conversion metric is both statistically and practically insignificant.

Table below summarizes the result from effect size testing:

| Evaluation Metrics | Minimum Difference, $d_{min}$ | Observed Value | Upper Bound | Lower Bound | Statistical Significance | Practical Significance |
|---|---|---|---|---|---|---|
| Gross Conversion* | -0.01 | -0.02 | -0.01 | -0.03 | **Yes**, since 0 is out-of-bound | **Yes**, since $d_{min}$ is inside bound |
| Net Conversion** | 0.0075 | -0.0049 | 0.0019 | -0.0116 | **No**, since 0 is inside bound | **No**, since $d_{min}$ is out-of-bound |

*NOTE: Values rounded-up to 2-decimal places.

**NOTE: Values rounded-up to 4-decimal places.

### 5.2.3 Sign Test, Question-7:

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.

### 5.2.4 Answer-7:

In the sign test, we compare the difference between experiment and control groups on a day-by-day basis for the entire duration of our experiment. Each count of positive delta is considered as a success and negative delta as a failure. Below table summarizes the count of differences on day-by-day basis:

| Evaluation Metrics | Total Count of days | Total Count of Success | Total Count of Failure |
|---|---|---|---|
| Gross Conversion | 23 | 4 | 19 |
| Net Conversion | 23 | 10 | 13 |

*NOTE: For more details, refer to the Excel file "*Sign Test.xlsx*" submitted with this report.

Next we use the results from above table and online Sign Testing tool[5] for calculating the *p*-value.

The *p*-value for Gross Conversion is 0.0026.

Sign and binomial test
Number of "successes": 4
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 4 or fewer successes in 23 trials.
- The two-tail P value is 0.0026
  This is the chance of observing either 4 or fewer successes, or 19 or more successes, in 23 trials.

The *p*-value for Net Conversion is 0.6776.

Sign and binomial test
Number of "successes": 10
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 10 or fewer successes in 23 trials.
- The two-tail P value is 0.6776
  This is the chance of observing either 10 or fewer successes, or 13 or more successes, in 23 trials.

If *p*-value is less than $\alpha$ = 0.05, then it is said to be a significant change. This is summarized in table below:

| Evaluation Metric | *p*-value | Statistical Significance at $\alpha$ = 0.05 |
|---|---|---|
| **Gross Conversion** | 0.0026 | **Yes**, since $p < \alpha$ |
| **Net Conversion** | 0.6776 | **No**, since $p > \alpha$ |

### 5.2.5 Summary, Question-8:

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.

### 5.2.6 Answer-8:

The Bonferroni correction adjusts probability (p) values because of the increased risk of a type-I error when making multiple statistical tests. The routine use of this test has been criticised as deleterious to sound statistical judgment, testing the wrong hypothesis, and reducing the chance of a type-I error but at the expense of a type-II error; yet it remains popular.[6]

In our experiment we did not use Bonferroni correction because it gave a conservatrive approach for related metrics. Both our evaluation metrics were related to each other and we wanted to have significant results for both for launching the experiment.

With the effect size hypothesis tests we identified that Gross Conversion metric is both statistically and practically significant at 95% CI whereas Net Conversion metric is neither statistically nor practically significant for the same 95% CI.

The sign test seconded the outcome of effect size test that at 95% CI, Gross Conversion is significant but Net Conversion is insignificant.

## 5.3 Recommendation

### 5.3.1 Question-9:

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.

### 5.3.2 Answer-9:

Gross conversion is found to be both statistically and practically significant. This suggests that our experiment behaved as expected and due to the screener, users who were not able to devote the minimum hours per week refrained from enrolling. But on the other side Net conversion was not statistically nor practically significant. This was not as expected. It could be because of the short period when this test was launched and pageviews data could not be sufficient. Also we were not able to measure Retention metric.

For this reason I recommend not to launch this experiment without further investigation and confirmation that there were no other side effects.

# 6. Follow-up Experiment

## 6.1 Question-10:

If you wanted to reduce the number of frustrated students who cancel early in the course, what experiment would you try? Give a brief description of the change you would make,

what your hypothesis would be about the effect of the change, what metrics you would want to measure, and what unit of diversion you would use. Include an explanation of each of your choices.

## 6.2 Answer-10:

From the previous experiment, Gross conversion showed a favourable outcome but Net conversion wasn't significant to launch the experiment. As recommended above, the launch of the experiment is not proposed.

Next an add-on experiment is proposed to reduce the number of users who cancel early. To design an experiment for this we should know what could be the reasons for early cancellations. I assume that the user is dedicating only a minimum 5 hours and it is not sufficient. Also missing or lack of necessary previous knowledge of the subject could be another reason.

Description

On the course overview page, we will keep the screener asking for a range of hours commitment from the user and also showing some of the pre-requisite skills that are mandatory for the corresponding course. So the screener would be specific and dedicated to each course but not a generalized one. This will help in self-assessment before proceeding further. It may help in avoiding frustrated users who may enroll keeping just the minimum hours in mind but missing the aspect of skills and knowledge necessary to succeed in the course.

After one week of enrollment, a report to the user can be sent where the time spent on the course and progress made during the first week will be compared with average results of other users who successfully completed this course. This will help by motivating and encouraging the users who are not performing better. Some other tips and guidelines can be included in this report to support such users. This will also increase the users' belief in Udacity's support and improve their faith to carry on with the course and refrain from cancelling early.

Hypothesis or Effect of the change

The students receiving the report after the first week of enrollment will be encouraged to carry on and this will reduce the number of early cancellation.

Unit of diversion

Upto enrollment the user experience is the same and after enrollment the diversion to experiment and control group will happen based on **user-id** and evenly on random basis.

Invariant Metric

Since the diversion will happen randomly and user-id will be the common property for both the experiment and control groups, **count of user-ids** would be an ideal invariant metric.

It is expected that the count of user-ids would remain identical between the two groups.

Evaluation Metric

**Retention rate** i.e., 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. This would be an ideal evaluation metric to compare the experiment and control groups.

It is expected that the retention rate would remain higher in the experiment group compared to the control group.

It is also recommended to run the experiment for a longer period of time to cover different demographics of users and it will also support in capturing higher volume of data for analysis. If a statistically and practically significant and positive change is observed in retention rate then this experiment can be launched by Udacity.

## 7. References

[1] A/B Testing Metrics: Blog - http://napitupulu-jon.appspot.com/posts/metrics-abtesting-udacity.html

[2] Baseline values: https://www.google.com/url?q=https://docs.google.com/a/knowlabs.com/spreadsheets/d/1MYNUtC47Pg8hdoCjOXaHqF-thheGpUshrFA21BAJnNc/edit%23gid%3D0&sa=D&ust=1587929703046000

[3] Online Sample Size Calculator: https://www.evanmiller.org/ab-testing/sample-size.html

[4] Z critical value calculator: https://statscalculator.com/zcriticalvaluecalculator?x1=0.05

[5] Sign and Binomial test: https://www.graphpad.com/quickcalcs/binomial1/

[6] When to use the Bonferroni correction: https://onlinelibrary.wiley.com/doi/10.1111/opo.12131

## 8. Rubric Checklist

1. Metric Choice: In Page-5, the table and contents below the table.
   - 1.1. Have good invariant and evaluation metrics been selected for the experiment?
   - 1.2. Has a well-reasoned justification of the choice of metrics been made?
   - 1.3. For which results would we wish to launch the experiment?: Expected results from metrics are underlined in the report.
2. Variability: Chapter 3.2.
   - 2.1. Have the standard deviation for all evaluation metrics been correctly calculated?
   - 2.2. Has reasoning been made whether each analytic standard deviation is likely to be accurate?
3. Sizing
   - 3.1. Does the number of pageviews correctly take into account the planned analysis?: Chapter 4.1.2.
   - 3.2. Has an appropriate level of exposure for the experiment been chosen based on the risk?: Chapter 4.2.2.
   - 3.3. Does the duration of the experiment correctly take the exposure chosen into account?: Chapter 4.2.2.
4. Sanity Checks: Chapter 5.1.2.
   - 4.1. Have sanity checks been performed correctly?
   - 4.2. Have the results of sanity checks been analyzed?
5. Effect Size Tests: Chapter 5.2.2.
   - 5.1. Have confidence intervals been calculated for the difference in all evaluation metrics?
   - 5.2. Have statistical and practical significance been correctly evaluated?
6. Sign Tests: Chapter 5.2.4.
   - 6.1. Has a sign test p-value been reported for each evaluation metric with indications whether the sign test is statistically significant?
7. Results Summary: Chapter 5.2.6.
   - 7.1. Has the choice whether to use the Bonferroni correction been justified?
   - 7.2. Have all discrepancies between the effect size tests and the sign tests been analyzed?: No discrepancies were found. effect size tests and sign tests results were inline to each other.
8. Recommendation: Chapter 5.3.2.
   - 8.1. Has a well-reasoned recommendation been made based on the results of the experiment?
9. Follow-Up Experiment: Chapter 5.3.2.
   - 9.1. Has a plausible experiment for the purpose given been made with a clearly stated hypothesis?
   - 9.2. Have good metrics to evaluate the proposed experiment been selected with good reasoning to support them?
   - 9.3. Has a well-reasoned unit of diversion for the experiment been selected?