P7: A/B TESTING

Experiment Design

Metric Choice:

I will begin by listing out my choice of invariant and evaluation metrics:

Invariant Metrics: Number of Cookies, Number of Clicks, Click-through-probability.
Evaluation Metrics: Gross Conversion, Net Conversion, Retention.

For each metric, I will proceed to spell out the reasoning for my choices.

- **Number of Cookies:** The number of cookies is a good invariant metric, because the assignment of cookies happens before the experiment; and this is the unit of diversion. This metric is thus independent, and for this reason, Number of Cookies will not make for a good evaluation metric.

- **Number of User-IDs:** Udacity moves to user-id tracking only after a student enrolls in a free trial. The data captured under this metric is influenced by the experiment. Number of User-IDs is not a good invariant metric.

  Moreover, Number of User-Ids will not make a good evaluation metric either. Because thought the number of users assigned to the experiment and control groups might vary – making for a potential evaluation metric, this metric is not normalized.

- **Number of Clicks:** Here, this metric represents the number of unique cookies to click the "Start Free Trial" button. This still happens before the experiment – which only kicks off after the user sees the message. This metric is independent and a good invariant. These reasons do not make it a fit evaluation metric.

- **Click-through-probability:** The click-through-probability will make for a good invariant metric because the clicks come in before the experiment happens. Click-through-probability is thus independent, and not a fit evaluation metric.

- **Gross Conversion:** Gross Conversion makes for a good evaluation metric. By definition, it is the number of user-ids to complete checkout and enroll in a free trial divided by the number of unique cookies to click on the "Start Free Trial" button. It seems to me that the experiment (displaying the

time commitment) will affect the number of user-ids that complete checkout, and this effect should be well captured by Gross Conversion.

It will not make for a good invariant, because the experiment might affect it.

- **Retention:** Again, this should make for a good evaluation metric for similar reasons as stated above. It measures the number of people who complete the checkout, and make payment toward the regular fee divided by the number of user-ids to complete checkout. This is a good evaluation metric, because the experiment affects the number of users who go through with the checkout process, and this will filter down to the people who stay back in the course even after the free trial expires.

  The lack of independence from the experiment makes Retention a bad invariant metric.

  SIDENOTE: Retention is a probability measure, it measures the probability that a user pays for the course.

- **Net Conversion:** Net conversion has a definition similar to Gross Conversion. It measures the number of successful paid enrollments divided by the number of unique cookies to click on "Start Free Trial" buttons. It makes for a good evaluation metric.

  It will not make for a good invariant metric because of its lack of independence from the experiment.

The hypothesis set was that the experiment *"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."*

Given my metrics, I will look for the following results:

- An **increase** in the probability that a user in the trial pays or Retention.
- A **decrease** in the number of user-ids completing the free trial checkout with respect to unique cookie clicks or the Gross Conversion Rate,
- **Without a decrease** in the number of user-ids enrolling in the course with respect to unique cookie clicks or the Net Conversion Rate.

Measuring Standard Deviation

Table of Baseline Values

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	0.109313

Above is a table of the baseline values observed for this experiment. These values are for the 40,000 page views happening every day. Since I only want to look at a slice of this – 5000 page views – I should compute metrics accordingly.

Given my choice of evaluation metrics, I will have to impute the following two values to compute the standard deviations of my evaluation metrics:

- Number of Unique Cookies that click:

  I can look at the probability that any person – who views the page – clicks, and scale it up by the number of page views.

  Number of Unique Cookies that click = 5000 * 0.08
  = **400**

- Number of User-IDs that enroll:

  I can first look at the probability that any person – who views the page – clicks, scale it up by the number of page views. And, then look at the probability that these people enroll.

  Number of User-IDs that enroll = 5000 * 0.08 * 0.20625
  = **82.5**

Now I can go ahead and calculate the standard deviation for my evaluation metrics.

I will estimate the standard deviation analytically by assuming a binomial distribution. Making this assumption makes sense because:

1. *The number of observations is fixed.*
2. *Each observation is independent.*
3. *Each observation represents one of two outcomes ("success" or "failure").*
4. *The probability of "success" is the same for each outcome.*

I can use the make the estimation in the following way:

$$\sqrt{p \cdot (1-p) \cdot \frac{1}{n}}$$

- Retention:

  $\sqrt{0.53 \cdot (1-0.53) \cdot 1 \,/\, 82 \cdot 5}$ = **0.0549**

- Gross Conversion:

  $\sqrt{0.20625 \cdot (1-0.20635) \cdot 1 \,/\, 400}$ = **0.02023**

- Net Conversion:

  $\sqrt{0.109313 \cdot (1-0.109313) \cdot 1 \,/\, 400}$ = **0.0156**

In the case of Net and Gross Conversion rates, the Unit of Diversion is the same as the Unit of Analysis (Cookies), thus the analytical and empirical estimates of the variability will roughly be the same.

However, the Unit of Analysis in the case of Retention is different from the Unit of Diversion, thus the empirical variability will be higher than the analytical variability

Sizing

Number of Samples vs. Power

Since I am only analyzing three metrics, I will not use Bonferroni correction.

Calculating Page Views:

With $\alpha = 0.05$ and $\beta = 0.2$, we need the following page views to have enough power to detect a practical significance $d_{min} = 0.01$:

- Retention: **39,115** clicks per group.
- Gross Conversion: **25,835** clicks per group.
- Net Conversion: **27,413** clicks per group.

I arrived at the above values using this online calculator. To arrive at the actual page views required:

- Retention: 39,115 * 1/660 * 40,000 = **2370606**
- Gross Conversion: 25,835 * 1/3,200 * 40,000 = **322937.5**
- Net Conversion: 27,4133 * 1/3,200 * 40,000 = **342662.5**

Since the online calculator throws up the number of page views per group and not the number of views for both – control and experiment – groups, I will scale the metric appropriately.

Total number of page views required = 2370606 * 2 = **4741212**.

Duration vs. Exposure

No student is in harm's way for the purposes of this experiment. No additional personally identifiable information is collected. No changes to the backend infrastructure is made for the experiment. Given these considerations, I believe we can divert all of Udacity's traffic to the experiment – provided the experiment is concluded in a short timeframe (If the new feature does in fact cause a loss in revenue, I wouldn't want the feature up as a part of the experiment for an extended period of time.).

Now let us calculate the duration our experiment should run for to gain the requisite page views:

Duration: 4741212/40000 = 118.5303 or about **119 days**

This seems like too long a duration for the analysis to run. I think it will be prudent to drop Retention as an evaluation metric and look only at the Gross and Net Conversion instead.

Duration: 342662.5 * 2 / 40000 = 17133125 or about **18 days**

This looks more doable. The experiment will run for a little more than two weeks if all of the traffic is diverted.

Experiment Analysis

Sanity Checks

Number of Cookies:

To run this metric through a sanity check, we will have to determine the total number of clicks in both groups.

	Number of Cookies
Control	345543
Experiment	344660

Next, we can proceed to calculate the Confidence Interval.

SD = $\sqrt{0.5 \cdot (1 - 0.5)/345543 + 344660}$ = **0.000602**

m = 1.96 * 0.000602 = **0.00118**

CI = (0.5 – 0.00118, 0.5 + 0.00118) = **(0.49882, 0.50118)**

Next, we calculate p = 345543 / (345543 + 344660) = **0.50064**

Number of Clicks:

	Number of Clicks
Control	28378
Experiment	28375

SD = $\sqrt{0.5 \cdot (1 - 0.5)/28378 + 28375}$ = **0.0021**

m = 1.96 * 0.000602 = **0.004116**

CI = (0.5 – 0.00118, 0.5 + 0.00118) = **(0.495884, 0.504116)**

Next, we calculate p = 28378/ (28378+ 28375) = **0.500467**

Click-through-probability:

Click-through-probability is composed of two of the invariants examined above – number of unique cookies to click by number of unique cookies to view the page. Both these metrics passed the Sanity Check. It is obvious that, by extension, click-through-probability should pass the check as well.

	Click-through-probability
Control	0.08212581357
Experiment	0.08218244067

P_{pool} = (28375 + 28378)/(345543 + 344660) = **0.08215409**

SE_{pool} = $\sqrt{\mathbf{0}.08227 \cdot (\mathbf{1} - \mathbf{0}.08227)/345543 + 344660}$ = **0.000661**

m = 1.96 * 0.000661 = **0.001296**

CI = (0 – 0.001296, 0.5 + 0.001296) = **(-0.001296, 0.001296)**

d = 0.08218244067 – 0.08212581357 = **0.0000566271**

Since the observed difference (d) falls within the Confidence Interval, this metric passes the sanity check.

Result Analysis

Effect Size Tests

- Gross Conversion:

  Pooled Probability ($\hat{p}$) = (3785 + 3423)/(17293 + 17260) = **0.2086**
  Pooled Difference ($\hat{d}$) = (3423/17260) – (3785/17293) = **-0.02055**

  Standard Error (SE) = $\sqrt{0.2086 \cdot (1 - 0.2086) \cdot \left(\frac{1}{17293} + \frac{1}{17260}\right)}$ = **0.004372**

  Margin of Error (m) = 0.004372 * 1.96 = **0.00857**

  Confidence Interval (CI) = **(-0.0291, – 0.0120)**

Since the Confidence Interval does not include 0, the result is statistically significant. The confidence interval does not include the practical significance threshold 0.1 either, which means the result is practically significant.

- Net Conversion:

  Pooled Probability ($\hat{p}$) = (2033 + 1945)/(17293 + 17260) = **0.1151**
  Pooled Difference ($\hat{d}$) = (1945/17260) – (2033/17293) = **-0.00487**

  Standard Error (SE) = $\sqrt{0.1151 \cdot (1 - 0.1151) \cdot \left(\frac{1}{17293} + \frac{1}{17260}\right)}$ = **0.003434**

  Margin of Error (m) = 0.003434 * 1.96 = **0.00673**

  Confidence Interval (CI) = **(-0.0116, 0.0019)**

  Since the Confidence Interval includes 0, the result is not statistically or practically significant.

Sign Tests

I will use an Online Calculator to compute p values. To do my sign test, I will look at the difference between the observations of the control group from the experiment group for each of the 23 days, and make a note of positive changes or successes.

- Gross Conversion:

  Successes: 4

  With 4 successes out of 23, we get a p-value of 0.0026 which is lesser than the α level of 0.05. Thus, the result is statistically significant.

- Net Conversion:

  Successes: 10

  With 10 successes out of 23, we get a p-value of 0.6776 which is greater than the α level of 0.05. Thus, the result is not statistically significant.

Summary

Bonferroni Correction is designed to reduce False Positives or Type-1 error.

I did not use Bonferroni correction because in this case, I was looking at two metrics, and I needed both the metrics to be significant to launch the experiment. I was not running the risk of launching the experiment by incorrectly rejecting the null hypothesis even though it was actually true (a False Positive).

Recommendation

To understand the impact of this change, I looked at the Gross Conversion and Net Conversion.

Gross Conversion showed a statistically and practically significant decrease in the numbers in the experiment group from the control group. This was encouraging, as we know the feature will reduce the number of people who enroll and drop out due to the time commitment required. However, Net Conversion did not show a statistically or practically significant change. The Confidence Interval of Net Conversion included the negative of the practical significance boundary. In this case, we cannot be sure whether the new feature will cause a loss in revenues.

Since we do not want to reduce the number of "paid students who complete the course", this is not acceptable risk to launch the experiment.
I would not launch this experiment.

Follow-Up Experiment

From personal experience, one aspect that Udacity can change is its project review apparatus by allowing students and reviewers to engage in a back and forth exchange. Though there are many other forums for students to address their concerns, the reviewer gets to know and engage with a student's project. Enabling a conversation between them will help students perform better. I think this will keep them enthused, and keep them from dropping out.

My null hypothesis is that enabling students and reviewers engage in a conversation will improve retention rates.

My Unit of Diversion will be User-ID to provide a consistent experience for every user.

My Invariant Metric will also be User-ID, same as my Unit of Diversion.

The Evaluation Metric will be Retention. However, I will redefine retention to the number of students who successfully complete their first project divided by the number of students who enroll for the free trial.

References:

http://www.stat.yale.edu/Courses/1997-98/101/binom.htm
http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/type-i-and-type-ii-error/