P7-A/B Testing
By: Tedros Hagos

Introduction:
Udacity courses currently have two options on the home 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.

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. In this project, I will investigate if the introduction of this experiment helps to increase the retention rate of students who at least started the free trial.

Experiment Design
Here are the list of metrics used in throughout this project:
Number of cookies, number of user-ids, number of clicks, click-through-probability, gross conversion, retention and net conversion

Metric Choice

- **Invariant metrics**: are those metrics which are not affected by the introduction of the screener.
- Number of cookies and Number of clicks
- **Evaluation metrics**: Gross conversion, Retention and Net conversion

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.

- **Number of cookies**: (number of unique cookies to view the course overview) is an invariant metric. The user see the page before the experimental button: “start a free trial”, so the number of cookies does not change as the event takes place before seeing the new button.
- **Number of user-ids:** (number of users who enroll in the free trial) number of user-ids are affected by the experiment (minimum time devotion message) button so it is not a good invariant metric. Therefore, it is an evaluation metric. But it is not used in our studies, as Number of user-

Id is simply a count and the Gross conversion is a fraction that includes the Number of User-Id which provides a better way to track the effect of the experiment.

- **Number of clicks:** (number of unique cookies to click the "Start free trial" button) is an invariant matric, because visitors have not seen the number of hours devotion message and thus it does not have any effect on weather visitors have to click the button or not.
- **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) is an invariant matric, because the clicks happen before the user sees the time devotion message (experiment), but since it is derived from 2 invariant metrics: number of clicks and number of cookies, it can be well represented by those 2 metrics and thus we can safely omit in our study as an invariant metric.
- **Gross conversion: (**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) is not an invariant metric, because the number of user-ids to enroll in the free trial is affected by the experiment (if user does not have time to devote more than 5 hours, is more likely that s/he will not proceed to the next level) and thus affects the number of students to enroll. But it is a good measure of evaluation matric which is a good indicator of the impact of our experiment and how it improves the attrition rate of enrollment and course completion.
- **Retention:** (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) is not an invariant metrics, as it is directly affected by the experiment. But it is an evaluation matric and thus it could be a good indicator of the experiment was successful in fulfilling its intended output.
- **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) is not a good invariant matric as the number of students to enroll are affected by the experiment after clicking the button and getting the time devotion message. However, it is an evaluation matric and it can be a good evaluation feature to check the effectiveness of our experiment.

  To launch the experiment we need to decrease the number of students who might withdraw during the 14 free trial days. At the same time, we need the number of students who remained enrolled after the 14 days of free trial increases or remain the same. In other words, we will launch only if Gross conversion decrease, but Net conversion stays the same or increases.

Measuring Standard Deviation

Based on the provided information and sample size of 5000 cookies who visits the course overview page, I will compute the Standard Deviations of:

- Gross Conversion:

  Given: click-through-probability=0.08, Probability of enrolling, given click:=0.20625
  Number of Clicks (N) given 5000 views: 0.08*5000= 400

  Std. Dev= SQRT(p*(1-p)/N)= SQRT(0.20625* (1- 0.20625)/400)= 0.020230604

- Retention:
  Given: probability of enrolling=0.20625, Probability of payment, given enroll (P)=0.53
  Number of Clicks (N) given 5000 views: 0.08*5000= 400 (computed above)
  Number of Enrollments given 400 clicks (N)=400*0.20625= 82.5
  Std.Dev= SQRT(P*(1-P)/N)= SQRT(0.53*(1-0.53)/82.5)= 0.054949012

- Net Conversion:
  Given: P= 0.1093125, N= Number of Clicks (N) given 5000 views=400 (The same as gross converstio
  $\sqrt{\frac{p(1-p)}{N}}$ =
  Std. Dev= SQRT(p*(1-p)/N)= SQRT(0.1093125* (1- 0.1093125)/400)= 0.015601545

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the 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.

Gross Conversion and Net conversion have both number of cookies as their denominator, and thus their analytic estimate would be comparable to their empirical variability.
In the case of retention, our denominator is Number of enrolment, so our units of analysis and variability are not comparable.

Sizing

Number of Samples vs. Power

The Bonferroni correction is a multiple-comparison correction used when several dependent or independent statistical tests are being performed simultaneously. In our case, we are performing only one test, so it is not needed.

To compute the number of page-views that we need to conduct our experiment, first we have to calculate the sample size for each evaluation matric. Based on the following prior information:
α=5%, β=20% are given for all metrics.
Besides to the above given parameters, the Minimum Detectable Effect (d_{min}) and Baseline conversion rates (BCR) are given for each evaluation metric as follows:

- Gross conversion: d_{min} =1%, BCR=20.625,
- Retention: d_{min} =1%, BCR=53%
- Net Conversion: d_{min} =0.75%, BCR= 10.93129%

Using the sample size online calculator and the above given parameters, the sample size (per group) for each evaluation metric is found to be as follows:
Gross conversion=25,835, Retention=39,115 and Net Conversion=27,413.
We are also given that, in order to make 40,000 page-views are required to make 3200 clicks and 660 enrollments. If we write this in the form of ratios, we get the following:
Ration of clicks to page views: 3200/40,000= 0.08
Ratio of enrollment to page-views: 660/40,000= 0.0165

So the number of page-views for each metric will be as follows:

- Gross conversion=25,835/0.08=31,732.5 page views for each control and experiment groups
  For both groups=31,732.5*2= 634,625 page-views

- Retention=39,115/0.0165= 2,370,606, for both groups: 4,741,212.12 page-views
- Net-Conversion: 27413/0.08= 342,662.50, for both groups: 685,325 page views

Duration vs. Exposure

Given a daily page-view of 40,000 (for both groups), I would take 100% of traffic to perform this experiment which will result in a daily 40,000 page views. In this case if we take retention (our greatest sample size and page-view need), it would take around 119 days (4 months) which is too long and not advisable both in terms of time and cost (inefficient use of coaching resources). Let us check how long it takes with Net-conversion (our next largest sample size and page view). In this case it will need 17 days, which is by far better than the time it takes for retention. So I will drop the evaluation metric retention and continue my experiment with only Gross Conversion and Net Conversion.

I don't see any potential risk in performing this experiment. Moreover, it is simply a popup message that reminds users about the time commitment they need. None of the users could suffer physical harm as a result of the experiment, nor is sensitive data being collected, therefore a 100% exposure is a safe.

Experiment Analysis

Sanity Checks

In this section, we will check if all our invariant metrics pass the sanity check, if not we have to look back and investigate what went wrong. I will assume number of cookies and clicks are evenly distributed between our control and experimental group. I will also assume equal probability of 0.5 and 95% confidence Interval, I will compute the Standard deviation, margin of error, confidence interval and actual observed value for each invariant metric.

1. **Cookies:**
   Total Number of page views for Experiment Group (N_E)=344,660
   Total Number of page Views for Control Group (N_C) = 345,543
   Total Number of page views: 344,660+345,543=690,203
   Standard Deviation: Std.Dev = $\sqrt{\frac{P(1-P)}{N}} = \sqrt{\frac{0.5(1-0.5)}{690,203}} = 0.000601841$

   Assuming Normal Distribution:

   Z score for 95% CI= 1.96

   Margin of Error (ME) = Z score * Std.Dev = 1.96 * 0.000601841 = 0.001179608
   Confidence Interval (CI) =(0.4988,0.5012)
   Observed Value=0.500639667
   Observed value lies within the confidence Interval value, so it has passed the sanity check

2. **Number of Clicks**:
   Total Number of Clicks for Experiment Group (N_E)=28325
   Total Number of Clicks of control Group (N_C)=28378
   Total Number of clicks (both gropus)= 56703
   Standard Deviation: Std.Dev = $\sqrt{\frac{P(1-P)}{N}} = \sqrt{\frac{0.5(1-0.5)}{56703}} = 0.002099747$
   Margin of Error (ME) = Z score * Std.Dev = 1.96 * 0.002099747= 0.004115504
   Confidence Interval (CI) =(0.495884496, 0.504115504)

Observed Value=28378/5603 = 0.500467347
Observed value lies within the confidence Interval value, so it has passed sanity check

All the invariant metrics pass the Sanity check, so it seems plausible to continue with our experiment.

Result Analysis

Effect Size Tests

With our invariant metrics (Gross conversion and Net conversion), I will check if each of them are statically and practically significant (at 95% confidence interval). A metric is statistically significant if the confidence interval does not include 0 and it is practically significant if the confidence interval does not include the practical significance boundary.

1. **Gross conversion:**

To compute the pooled probability of enrolment and Std. Deviation, we consider the number of clicks and page views which happened before the 14 days trial period, so that it gives us an accurate assessment of enrolment.

Total Number of Clicks for Experiment Group (N_E) = 17260
Total Number of Clicks of control Group (N_C) = 17293
Total Number of clicks both groups (N)= 34553

Total Number of Enrollments for Experiment Group (E_E)=3423
Total Number of Enrollments of control Group (E_C)=3785
Total Number of Enrollments both groups (E)= 7208

$\hat{\hat{P}}$ =E/N =7208/34553 = 0.208607067

Std. Dev= $\sqrt{\frac{P(1-P)}{N/(NE*NC}} = \sqrt{\frac{0.208607067(1-0.208607067)}{34553/(17260*17293)}}$ = 0.004371675
Margin of Error (ME) = Z score * Std.Dev = 1.96 * 0.004371675= 0.008568484

Observed Difference (D)= $E_E/N_E - E_C/N_C$ = 3423/17260-3785/17293 = -0.020554875
Confidence Interval = (-0.029123358 , -0.011986391)
0 (Zero) is not included in the confidence interval, so Gross conversion is statistically significant. Moreover, it is practically significant as it does not include the practical significance boundary.

2. **Net conversion:**

Total Number of Clicks for Experiment Group (N_E) = 17260
Total Number of Clicks of control Group (N_C) = 17293
Total Number of clicks both groups (N)= 34553

Total Number of Payments for Experiment Group (P_E)=1945
Total Number of Payments of control Group (P_C)=2033
Total Number of Payments both groups (Pay)= 3978

$\hat{\hat{P}}$ =Pay/N =39788/34553 = 0.115127485

Std. Dev= $\sqrt{\frac{P(1-P)}{N/(NE*NC}} = \sqrt{\frac{0.115127485(1-0.1151274857)}{34553/(17260*17293)}}$ = 0.003434134
Margin of Error (ME) = Z score * Std.Dev = 1.96 * 0.003434134 = 0.006730902

Observed Difference (D)= $P_E/N_E - P_C/N_C$ = 1945/17260- 2033/17293 = -0.004873723
Confidence Interval = (-0.011604624, 0.001857179)
0 (Zero) is included in the confidence interval, so Net conversion is not statistically significant.
Moreover, it is not practically significant since its d_{min} = (-)0.0075 is also within the boundary.

Sign Tests

For each day, we compute the difference between experimental and control group for each of our evaluation metrics and the result is recorded. We consider our experiment as success if the difference is positive, failure otherwise. So we have two possible outcomes: success or failure. That means we can safely assume the binomial sign test. We count number of successes for each metrics and compute the 2 tailed P-value. This way, we got 4 successes (out of 23 trials) for Gross-conversion and two-tailed p-value of 0.0026, which is much smaller than ($\alpha/2$=0.025). This indicates that at 95% confidence interval, Gross conversion is statistically significant. On the other hand, Net conversion has 10 of 23 successes and a two-tailed p-value of 0.6776. This indicates that at 95% confidence interval, Net conversion is not statistically significant.

Summary

The Bonferroni correction is used to reduce the chances of obtaining false-positive results (type I errors) when multiple pair wise tests are performed on a single set of data to identify at least one significant result. Our hypothesis requires the significance of both metrics (we look for a decrease in gross conversion and for a no decrease in the net conversion) if we have to decide on launching the new design in the Udacity site. If our aim was launching the new website upon getting at least one significant metrics, then the use of Bonferroni correction would come in effect. Therefore, I did not use the Bonferroni correction in this analysis.

Both the sign test and effect size tests indicate that, Gross conversion is both statistically and practically significant, but Net conversion is not. Since both tests go inline, further tests are not needed.

Recommendation

Based on the analysis we discovered that Gross conversion is statistically and practically significant. This means that the “minimum time devotion message” is effective at eliminating frustrated students who were enrolled in the free trial. However, Net conversion is not significant which means the “minimum time devotion message” reduces the number of students who would have paid at least one payment if they remained enrolled past the free trial period. The confidence interval for net conversion [-0.0116 , 0.0019] contains the negative of practically significant value -0.0075. The net conversion might have diminished by an amount that matters to the business, that is why, giving those numbers, we cannot take the risk to launch.

Follow-Up Experiment

I believe that, students who are performing well in the program and on schedule are most likely to continue. But those who are lagged behind schedule (due to short of time or time management) should be notified before it is too late so that they can fix and continue their studies.

As a follow up, if student is failing behind schedule, I would suggest a popup message appear to encourage his accomplishment so far and how to push more. For instance if a student fail to cover the first week's portion during the 14 days trails period, a popup message would appear with links leading to story of students who struggled at the beginning but eventually make it by sharing their techniques in overcoming such difficulties. This popup message would continue whenever (but occasionally, like once in a week)a student fails behind schedule even after the 14 days trial and reminds student to push more, while encouraging his/her work done so far. Along with the reminder popup message, story of successful graduates in the same course (program) could be shared: be it on how they managed their time or how completing the course helped them to successfully get a better job and lead a better life. Parallel to this Udacity's job placement rate (the rate at which successful graduates of Udacity students get a job with their newly acquired skills) stories of other motivated students can be shared on the student's home page. This will encourage students who are committing 5 or more hours and also on schedule during their studies, who may not get the popup messages reminder. This way, enrolled students will be motivated and helps them to see the bright light at the end of the tunnel and continue to work hard to complete the course.

The presence of the filtering screen that reminds minimum 5 hours commitment at the beginning of the page is essential, as it filters out those students who are not ready to commit some time for study who would otherwise frustrate early and drop anyway. Besides to this, I believe that, students may get discouraged once they are behind schedule (sometimes, one might not be aware is lagged behind schedule) and thus the popup message reminds the student that he is lagged behind and the suggested links will help on how to cope in such tight schedule, which I believe would help students from frustrating throughout the course.

This way the best interest of Udacity will not be only getting a one-time payment of enrolled students, but also students will complete the course.

Therefore, our main goal is to minimize early cancelation and encourage students to complete the course. That is, driving students past the 14 days trial period and make at least one payment and even complete the course. In other words, the aim is to increase retention, by encouraging students to commit more time and complete the course once they are enrolled in the free trial program.

My hypothesis is that, the popup message and the links following it will encourage students to commit more time to successfully complete the trial period and enroll in the course and possibly complete it. Using this hypothesis, I expect the number of frustrated students to decrease and thus resulting in an increase in retention.

Unit of diversion is going to be **user-id**. Once a user is registered during the free trial period, they are tracked by their user-id from free trial period all the way to completion of the course.

Invariant metric: is going to be **user-id**, which is already created during login to the trial page and will not be affected by the introduction of the popup message.

Evaluation metric is going to be **retention.** In this analysis, retention shows the changes achieved towards reducing attrition rate of enrolled students.

If retention shows statistically and practically significant increase, then the experiment will be launched on the Udacity webpage.

References:

1. http://mathworld.wolfram.com/BonferroniCorrection.html
2. https://www.optimizesmart.com/understanding-ab-testing-statistics-to-get-real-lift-in-conversions/
3. http://onlinelibrary.wiley.com/doi/10.1111/opo.12131/full