Udacity Report: A/B Testing.

Anirudh Ramesh, Dec 20, 2016

The Objective of this A/B testing exercise is to reduce the number of students who cancel their subscription to a course in the 'free trial' phase, without affecting the number of students moving past free trial towards a nanodegree.

Experiment Design

Metric Choice:

- **Invariants:**

1. **Number of Cookies:** I choose to keep the number of *unique* cookies visiting the page in a day constant in control and experiment groups. Any significant difference would mean uneven distribution of traffic – which would make the whole process invalid.
2. **Number of Clicks:** Number of clicks on the start free-trial button also needs to be an invariant. Any significant difference between Control and experiment in this metrics would mean something is clearly different on how the button looks/functions - as one could tell from click-through-probability. Number of clicks is also used in measuring Gross-conversion and Net-Conversion, and this parameter should be set before measuring evaluation metrics, since it is not affected by the experiment.
3. **Click-through-Probability:** Any change is click through probability would mean different version/variants of the button might have present in the two groups. This also might be due to intrinsic bias in the population we might have ignored between control and experiment.

- **Evaluation Metrics:**

Users are prompted to only move forward if they can dedicate 5 hours or more weekly for the courses. To find out if this indeed has improved student experience, one way is to measure number of Users who click 'Start free trail' even after the prompt (meaning they do have 5 hours or more / they understand the risk) and subsequently measure the number of users who enroll/pay. There should be a significant difference if the experiment has worked.

1. **Gross-Conversion:** I choose gross-conversion as an evaluation metric – which is (Number of users who checkout and enroll / Number of Cookies which click start free trial). This is pretty straight forward measure, as people who think they can keep up will probably enroll and people who cannot cope with and dedicate 5 hours weekly, wouldn't have clicked 'Start free trial' in the first place! (If this experiment worked). It is easier to get if the prompt served well from this metric.
2. **Net Conversion:** The reason I am using Net conversion as one of the evaluation metric is because:
   (Number of users who checkout and pay / Number of Cookies clicking start free trial) – although this gives mostly the same insight as gross-conversion, for a given course and given number of users, net-conversion should be unaffected.

I wanted to use 'retention' as well, but conducting an A/B test using 3 evaluation metrics:

1. Using retention required largest sample size (of the size 4.7 million approximately) which would need 100% of the traffic to be diverted to this, even If I hope to complete it in period of 2 months. There is also the issue that there might be some unforeseen effects that I might have missed which may cause significant damage testing on 100% of the population.
2. The insight given by retention can be derived using net-conversion as well – i.e. for a given course and given number of students, the retention should be the same ('Without affecting') , therefore I decided to use only gross-conversion and net-conversion only, since net-conversion also follows the same direction as retention, but needs fewer page views.

I will now have to see if the experiment side yields significantly difference compared to the control side.

I expect a decrease in Gross-Conversion in experiment group, if the experiment indeed has worked – because although the number of users who enroll would remain same as before, if the prompt worked, denominator would get larger than if there were no prompt!

I expect **net-conversion** to remain unaffected according to our hypothesis.
I wouldn't mind an increase as well, which I think might happen since if the experiment had indeed worked, people who had checked out to free trial are the ones who understood the prompt, and this should remove any students who actually had < 5 hours to spend weekly. This would remove any cancellations from students who have <5h per week (Without affecting those who might have went on to enroll in a nano-degree in the first place). i.e.

People who have at least one Payment in past 14 days / People who checked-out to free trial

➔ Numerator should remain the same and denominator should have decreased if the prompt worked.

However, since the hypothesis expects no change in number of students who go on to enroll in nano-degree I would settle for no-change in net-conversion.
These two evaluation metrics alone should provide insights as to if I should launch the experiment or not.

Since, the objective is to improve student experience – **I would launch, if there is a significant difference in Gross-Conversion and net-conversion is unaffected.**

Measuring Standard Deviation:

Standard Deviation: ((p * (1-p)) / n) ^ (0.5)
P: Probability of event
n = Sample size.

Gross Conversion :

40,000 Cookies visiting each day have click-through probability of 0.08.
Therefore, for a sample size of 5,000 cookies:
n =

Number of clicks : 400 (Page views x Click-Through-Probability)
Gross Conversion : 0.020230604

Net Conversion :

40,000 Cookies visiting each day have click-through probability of 0.08.
Therefore, for a sample size of 5,000 cookies:
n =

Number of clicks : 400 (Page views x Click-Through-Probability)
Net Conversion : 0.015601545

For both the metrics, since the Unit of diversion and unit of analysis is the same – I expect it to match to be close to the analytical estimate – without any significant difference.

Sizing:

Number of Sample Size vs. Power: I decided against using Bonferroni Correction in my estimate as I feel the evaluation metrics are somewhat correlated and I do not want too conservative of an estimate.

Sample size for :
Gross Conversion = Number of Unique Cookies who Enroll in Free trial / **Number of unique cookies to click button**
Net Conversion = Number of Unique Cookies who pay after free-trial (in past 14 days) / **Number of unique cookies to click button**
Online Calculator values :
Gross Conversion :
Alpha = 0.05 , Sensitivity = 80% , Baseline Value : 20.625% , **to Observe a change of 1%**
Number of Clicks required : 25,835 Per group.
Net Conversion :
Alpha = 0.05 , Sensitivity = 80% , Baseline Value : 10.9312% , **to Observe a change of 0.75%**
Number of clicks required : 27,413 Per group.

I decided to use the largest value of 27,413 click per group – for a click-through probability of 0.08 , I would require (27,413 * 0.08) per group and a total of (27413 * 0.08 * 2) for the entire set (Control + Experiment) – which is roughly – **6,85,325 page views**.

Duration and Exposure:

I decided to use 80% of the traffic for the experiment. The data being collected is not very sensitive, this is **not risky at all for Udacity**. However, this also gives me a room of 20% of the population, to check if there are any irregularities, and see how the population under experiment and the 20% of the population who do not get the prompt behave.

Diverting 80% of the traffic requires 21 days to completely get the required number of page views for the analysis. I feel this gives a reasonable amount of time to arrive at a conclusion.

Experiment Analysis:

Sanity Checks:

I performed sanity checks to see if the data is properly distributed between control and experiment groups.

Number of Cookies :	**Within Permissible range**	
Control :	345543	
Experiment :	344660	
Chance of being assigned randomly :		0.5
Standard error :	0.000601841	3.62212E-07
Margin of error :	0.001179608	
Upper bound value :	0.498820392	
Lower bound value :	0.501179608	
Observed value :	0.500639667	

Number of clicks on start free-trial :	**within Permissible range**	
Control :	28378	
Experiment :	28325	
Chance of clicking button :		0.5
Standard Error :	0.002099747	4.40894E-06
Margin of error :	0.004115504	
Upper bound value :	0.504115504	
Lower bound value :	0.495884496	
Observed value :	0.500467347	

Click-through probability on start-free-trial: **Within permissible range**	
Control	
Click-through probability :	0.082125814
Standard error :	0.000467068
Margin of error :	0.000915454
Upper bound value :	0.083041267
Lower bound value :	0.08121036
Observed value (experiment) :	0.082182441

From these results, I found no irregularities. Next, I proceeded to see if I could launch the experiment.

Result Analysis:

Effect Size tests:

For 95% confidence interval:

Gross-Conversion: Both Statistically and Practically significant.

Net-Conversion: Neither statistically nor practically significant.

Gross Conversion Calculation :	
Control :	
Total Enrollments :	3785
Total clicks :	17293
Ratio :	0.218874689
Experiment :	
Total Enrollments ;	3423
Total clicks :	17260
Ratio :	0.198319815
Pooled variance :	0.208607067
	1.91115E-05
	0.004371675
Standard Error :	**0.0044**
Margin of error :	0.008624
d-cap :	-0.020554875
CI Upper :	-0.011930875
CI lower :	-0.029178875

Net Conversion Calculation :	
Control :	
Total Payments:	2033
Total clicks :	17293
Ratio :	0.117562019
Experiment :	
Total Payments ;	1945
Total clicks :	17260
Ratio :	0.112688297
Pooled variance :	0.115127485
	1.17933E-05
	0.003434134
Standard Error :	**0.00343413**
Margin of error :	0.006730895
d-cap :	-0.004873723
CI Upper : (Rounded)	0.0019
CI lower : (Rounded)	-0.0116

Sign Tests:

Metric	P-Value :
1. Gross Conversion	0.0026 (statistically significant)
2. Net Conversion	0.6776 (Not statistically significant).

Summary:

I am not using Bonferroni correction for this experiment. Bonferroni correction is designed for cases where a single metrics is sufficient to accept/reject a experiment. In our experiment, we need two-metrics: Gross and net-conversion to stick to a certain pattern.

With such small values of alpha, I might ignore movements, which I will be able to see with higher values of alpha i.e I am more likely to **not reject null** for an alpha for 0.025 (which would be the case if I used Bonferroni correction) than when I use alpha value of 0.05. In addition, I reject the experiment even if one of the metrics fails to behave as expected i.e. I am more likely to have a type II error (Reduces statistical power).
Therefore, I have decided against using using Bonferroni correction.

When observing the data, I noticed the net conversion for the most part, (although it is not statically significant metrics) is high for 3 out of 4 Sunday in the dataset. There can be plenty of reasons for this – but net-conversion looks to show significant improvement during Sundays. This is one of the irregularities I noted.

Recommendation:

From the experiment, it is clear that the prompt has worked to improve student experiment. This is great news – First part of the hypothesis has been successfully completed.
However, the movement of net-conversion has a range of possibilities – it can decrease or increase, but it has more range of values in the negative direction. Also, the negative range of values have the practical significant boundary(in the negative direction) – therefore, increasing power could very well decrease net-conversion, which would be detrimental to udacity.This would mean, most likely Udacity would be compromising on revenues at the cost of improving student experience.

Based on the current insights alone, I recommend against launching the change.

Follow-Up Experiment: How to Reduce Early Cancellations:

If a user is already past the free-trial period and enrolled to a course after checkout (after a period of 14 days) – any early cancellation, I feel, can be attributed to a set of reason : Difficulty might have been higher than expected, or the time dedicated for the course , weekly , has been over-estimated by the

user , amongst few others. For the follow-up experiment, I would like to explore the irregularities in time dedicated by user - aspect of this.

Since the user has already subscribed to the course, **the unit of diversion is the User ID itself.**

I would implement few changes from the free-trial phase itself:

1. I would introduce a weekly deadline based on the time the User had said he/she could spend in a week. This would give the user a realistic idea about the actual time he needs to be able to afford. On the other hand, if the user is ahead of the schedule, there should not be any notification until he is lagging again.
2. I would also display(not essentially notify , but the user can see if he/she wants to) the time taken to complete the nano-degree based on the time entered by User, taking average from previous users and how long they took to complete the nanodegree – by dedicating said time weekly. Therefore, each module in the Udacity should have deadlines, based on this weekly estimate.

I would use number of courses enrolled and number of User-IDs as invariants.

I would use Retention as an evaluation metric. (Alpha = 0.05, Sensitivity = 80%).

Significant increase in Retention would indicate that the population has given some heed to the notifications from Udacity. Using retention, I should be able to arrive at a conclusion!