Project 7: Design an A/B Test

- Free Trial Screener

Experiment Overview:

At the time of this experiment, 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.

Experiment Design

1. Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here.

- Invariant metrics: Number of cookies, Number of clicks, Click-through-probability
- Evaluation metrics: Gross conversion, Retention, 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:** Cookie is unit of diversion. Course overview page was visited before "Start free trial" clicked, therefore, Number of cookies would be the same in experiment and control groups.
- **Number of clicks**: As this happens before the free trial screener was triggered, both experiment and control groups still saw the same page.
- **Click-through-probability**: Since above two metrics are invariants, then the probability would be same on both groups.

- **Gross conversion:** After clicking the "Start free trial" button, users in experiment group should be asked by pop-up screener if they had enough time to study first. Users in control group should go to the checkout page directly, so Gross conversion is a good evaluation metric to test if screener would eliminate the users who left the free trial because they didn't have enough time. Gross conversion is our main evaluation metric. We expect that it would be lower in experiment group than in control group.
- **Retention:** We know that users in experiment group were aware of time commitment for successful completion before they enrolled the trial, enrolled trial users would be less and more likely they would remain enrolled after trial. In control group, it would be more enrolled trail users and more canceled users after free trail. Retention might be higher in the experiment group.
- **Net conversion:** Same as Gross conversion, each group treated different after clicking the "Start free trial" button. We expect to see that the screener would not significantly reduce the number of students to continue past the free trial and eventually complete the course.
- **Number of user-ids:** This is not a good invariant metric since we would see different value in the control and experiment groups. However, number of enrolled users, as a raw count, could be affected if experiment and control groups had different number of clicks. Therefore, it is not a good evaluation metric either. Metric, like ratio, probability, proportion or rate is more robust than a raw count metric.

In order to launch the experiment, we expect that the gross conversion would decrease both statistical and practical significance, since it would screen out some unserious users by introducing the new pop-up screener, while Net retention at lease would not decrease both statistical and practical significance.

2. Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics.

- Gross conversion: 0.0202
- Retention: 0.0549
- Net conversion: 0.0156

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.

When unit of diversion is equal to unit of analysis, which indicate the analytic estimate would be comparable to the empirical variability. Gross conversion and Net conversion meet above condition, they both use cookie as unit of analysis and unit of diversion. Retention is using **Number of user-ids** as unit of analysis, which is not equal to unit of diversion. Therefore, the analytic estimate and the empirical variability are different.

3. Sizing

3.1 Number of Samples vs. Power

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

- Will you use the Bonferroni correction: No
- Pageviews needed: 679,300

After I calculated the number of pageviews for Retention, I noticed that it would take almost four months to test our experiment even with full traffic (40,000 pageviews/ day), which is beyond our expectations, therefore I dropped the Retention from my evaluation metrics. My final evaluation metrics are Gross conversion and Net conversion.

3.2 Duration vs. Exposure

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

- Fraction of traffic: 0.75
- Length of experiment: 23

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

The experiment is totally harmless for customers and will not collect any sensitive information from customers. It would have a substantial impact on new enrollments, however, as long as we keep monitoring it during the experiment, we can find decline quickly. Therefore, the experiment is considered low risk, I would divert 75% traffic to this experiment in case of any bug.

Experiment Analysis

1. Sanity Checks

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.

- Number of cookies: CI: [0.4988, 0.5012]; observed: 0.5006; PASS
- Number of clicks on “Start free trial”: CI: [0.4959, 0.5041]; observed 0.5005; PASS
- Click-through-probability on “Start free trial”: CI: [-0.0013, 0.0013]; observed 0.0001; PASS

2. Result Analysis

2.1 Effect Size Tests

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.

- Gross conversion: [-0.0291, -0.0120], statistically significant, practically significant
- Net conversion: [-0.0116, 0.0019], not statistically significant, not practically significant

2.2 Sign Tests

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.

- Gross conversion: 0.0026, statistically significant
- Net conversion: 0.6776, not statistically significant

2.3 Summary

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.

- I did not use the Bonferroni correction, which is too conservative for this case, since gross conversion and net conversion metrics were high correlated. In this experiment, when performing multiple comparisons, the chance of a statistical error increases with our two metrics. As we know that false positives have the greatest impact when ANY metrics satisfied can trigger launch, and false negatives have the greatest impact when ALL metrics must be satisfied to trigger launch. The Bonferroni correction controls for false positives at increased false negatives. In our case, correction will not be needed, since the new screener would be launched only when both two metrics are satisfied.

3. Recommendation

Make a recommendation and briefly describe your reasoning.

The metrics I chose were Gross conversion and Net conversion. Gross conversion turned out to be negative and practically significant, which was what we expected. Net conversion turned out not to be statistically and practically significant, which was what we expected too. However, its confidence interval included negative numbers, which mean the new feature we introduced could lead paid users decreased.

According our result. I believe the screener feature definitely drew users attention positively. I would recommend testing on adding more features on the screener, before we decide whether to release or reject it.

Follow-Up Experiment

Give a high-level description of the follow up experiment you would run, what your hypothesis would be, what metrics you would want to measure, what your unit of diversion would be, and your reasoning for these choices.

Through the above experiment, I think the reason why users quit after trial, not just because they did not enough time, other issues, like under estimate the difficulty of the course, not enough learning motivation etc. could affect conversion.

It is always a big challenge for a new user to start learning new skills, especially the online course on an unfamiliar website. Screener is very good idea, which successfully reduces users who were lack of learning time, so that it can further optimize the allocation of resources, while reducing operating costs.

Next step, we need redesign the screener and add more features on the checkout page to increase conversion， In the follow-up experiment, I would like add a brief message on the bottom of pop-up screener, let them know the percentage of users had completed the first course to stimulate the sense of competition. And then, I would like to add the salary distribution or the future prospects of the relevant position on the checkout page to increase learning motivation.

Same as the original experiment, follow-up experiment will randomly be assigned to a Control and an Experiment group. Users in the Control group will remain unchanged. Users in the Experiment group will see the new screener and checkout page when signing up. Unit of diversion is cookie.

Null hypothesis: New designed screener and checkout page for free trial sign up will not increase Net conversion significantly and will not decrease Gross conversion significantly either.

Invariant metric: **Number of cookies** and **Number of clicks**, as those two metrics happen before the free trial screener is triggered, both experiment and control groups still see the same page.

Evaluation metric: Gross conversion and **Net conversion**, as each group treated different after clicking the “Start free trial” button. Therefore, we expect to see if new screener and checkout page would help decreasing the gross conversion ratio and increasing the ratio of user making payment over total users who clicked “Start free trial” button

If Gross conversion decrease statistical and practically significant, and Net conversion increase statistical and practically significant at the end of the experiment, then we can release the new screener and checkout page.

References：

- Controlled experiments on the web
- Bonferroni correction： From Wikipedia
- A/B testing: the most powerful way to turn clicks into customers / Dan Siroker, Pete Koomen.