A-B Testing

By Olivia Jonah May 17 2017

==========

Experiment Design

Metric Choice

> List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

>

> 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.

Used metrics

- **Invariant:** Number of cookies, Number of clicks
- **Evaluation:** Gross conversion, Retention, Net conversion

Selection reasoning

Invariant Metrics

- **Number of cookies:** This is a good invariant, since it is used as unit of diverse and the number of cookies assigned to the control and experiment groups should be random and approximately equal.
- **Number of clicks:** Since clicks happen right before the experiment , the number of clicks shouldn’t be affected by the experiment.

- **Click-through-probability:** This invariant is the number of unique cookies to click the "Start free trial" button divided by the number of cookies to view the course homepage , it is normalized for each group of users, and is highly correlated with two other metrics that I already using.

####Evaluation Metrics

- **Gross conversion:** This is important evaluation metric for our experiment, because it depends directly on number of enrollments.

- **Net conversion:** Since the experiment aims at reducing the number of frustrated students and increase percent of paying users, this metric reflects the potential preservation of paying students who complete the course.

####Unused Metrics

- **Number of user IDs: this variant is not normalized so it won't be a good pick and robust compared to cookies as unit of diversion in this experiment.
- **Retention: Primary reason not to keep this metric is the high duration due to the required sample. We also are expecting retention to increase based on the numerator –(number of user IDs to make a payment) not decreasing by much and the denominator (number of user IDs to enroll in the free trial) to decrease. It has already been captured both in the via gross conversion and net
  conversion. With expected changes in the numerator and denominator, it would be somewhat complex to identifying the individual effect of each. Additionally, the unit of diversion (a cookie) is different from the unit of analysis (the denominator of retention, user IDs in a free trial). This can affect our experiment in size or length.

Final decision making

As will be shown later **Retention** is hard to use, because to measure it we need too much page views (more than 4m), which will take too much time. So, in this experiment I will only use **Gross conversion** and **Net conversion** as evaluation metrics.

After the experiment, to be able to make the decision to accept the change the following must be true:

- **Gross conversion:** Should significantly decrease (Number of trials should be lower)

- **Net conversion:** Should not significantly decrease (Number of paying customers should not be lower)

Measuring Standard Deviation

> List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

>

> 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.

###Calculating Variability

Type of metric	Distribution	Estimated variance
Probability	Binomial(normal)	P^(1-P^)/N

P is found in the table of base line values. N is the answer when the number of clicks on Start Free Trial is divided by 8 which is used to scale the number of cookies in the baseline table of 40,000 to the sample size of 5,000 we now have

- **Gross conversion:** (p=0.20625, N= 3,200/8= 400, therefore standard deviation= 0.0202)

- **Net conversion (probability of payment given click):** (p= 0.1093125, N=3,200/8=400, standard deviation =0.0156)

The analytical estimates of variability might be similar or close to the empirical estimates and as such we might ignore the empirical estimates. This is because the unit of analysis (i.e. the denominator of both evaluation metrics) is the number of unique cookies to click the "Start free trial" button. Since the unit of diversion (a cookie) is the same as the unit of analysis for both metrics.

Sizing

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. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

Metric	Value	Std	Std /5000	Sample size	Probability	Pageviews
Gross conversion	0.2063	0.0072	0.0202	25835	0.0800	645875
Retention	0.5300	0.0194	0.0549	39115	0.0165	4741212
Net conversion	0.1093	0.0055	0.0156	27413	0.0800	685325

Bonferroni correction will not be used, since we only have few metrics and it adjust for experiments with multiple metrics. The metrics used in this experiment meet the criteria. Quoting from Udacity Coach Sheng Kung “the fewer metrics that you require to be significant to make a decision, and the more independent these metrics are, the stricter you need to be with your significance level to constrain your overall error rate.” We are already conservative in this experiment by choosing Gross and net conversions and we expect to launch if all our evaluation metrics meet the criteria.

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. (These should be the answers from the "Choosing Duration and Exposure" quiz.)

>

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

Given that we require 685,325 pageviews and Udacity gets 40,000 page views per day we would require 21 days to run the experiment. As 21 days is the maximum allowable duration for iteration, and 81.6% percent of traffic will be given to this experiment.

If it is set for 18 days which is also possible, it will require 100% traffic and all the trouble that will bring, namely, collection of data over a short period of time is at risked of being influenced by weekend and holidays. Running for a long period increases sample size, traffic over the weekend is normally different

from weekdays so making it 21 days will give us more weekends, fixing of bugs in the new instance which could be caught when there is 100% traffic.

This experiment is not risky as one including life and death like changing a person's medication. Also, the addition of the feature also won't affect the user experience once a user decides to enroll, even if the feature breaks, like the change of a database and it subsequently breaking would. The only risk is losing revenue when paying customers drop or cancel, but if we are committed to the current design of our experiment, so we can keep it running for as long as we could to be able to achieve the pageviews of 685,325. The only sensitive information I think will be present in this experiment is the credit/debit card details of paying students

Experiment Analysis

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. (These should be the answers from the "Sanity Checks" quiz.)

>

> 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.

Invariant Metric	Lower Bound	Upper Bound	Observed Value	Passes?
Number of cookies	0.4988	0.5012	0.5006	Yes
Number of clicks	0.4959	0.5041	0.5005	Yes
Click-through-probability	-0.0012	0.0013	0.0001	Yes

For each invariant metric, above are the 95% confidence intervals for the values we expect to observe, the actual observed value, and sanity test pass/fail determination. The values for number of cookies are number of cookies in the control group divided by the number of cookies in both groups, where the expected fraction is 0.5. The same is true for number of clicks. The values for click-through-probability are the difference between the control and the experiment

click-through-probabilities, where the expected difference is 0.

Result Analysis

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. (These should be the answers from the "Effect Size Tests" quiz.)

Evaluation Metric	Lower Bound	Upper Bound	Statistically Significant? (α = 0.05)	Practical Significance Boundary	Practically Significant?
Gross conversion	-0.0291	-0.0120	Yes	0.01	Yes
Net conversion	-0.0116	0.0019	No	0.0075	No

For each evaluation metric, above are the 95% confidence intervals around the difference between the experiment and control groups. Statistical and practical significance determinations are made, as well.

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. (These should be the answers from the "Sign Tests" quiz.)

Evaluation Metric	Sign Test p-value	Statistically Significant? ($\alpha = 0.05$)
Gross conversion	0.0026	Yes
Net conversion	0.6776	No

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.

Bonferroni correction was not used because there are only metrics and they meet the criteria, and to make final decision to launch or not launch experiment, we required the effects of both metrics, so the risk to reject null hypothesis is still 5%.This decrease would have come at the expense of our ability to detect a true effect.

There are no discrepancies between effect size and sign tests. Both declare the difference between the control group and the experiment group for gross conversion as statistically significant, but not for net conversion.

Recommendation

> Make a recommendation and briefly describe your reasoning.

I recommend that the free trial screen is not launched. However, there is some plus in this experiment namely; the confidence interval for gross conversion was entirely below the -1% boundary, which would benefit Udacity by freeing up coach capacity and their ability to improve student experience. The lower

bound of the confidence interval for net conversion, however, was below the -0.75% boundary. This potential lack of preservation of paying students would not be good for Udacity's revenues, since a decrease in net conversion below -0.75% is not acceptable for the business, more work is to be done in this area. The lower bound of the confidence interval for net conversion is -0.0116, which is not that far off from -0.0075. Making it probably possible to get different result if ran on different set of days with these holidays; Halloween, Thanksgiving, and the lead up to Christmas could cause an abnormal drop in payments as for some people this is an expensive time of year. Also we could run the experiment past our sized number of pageviews to tighten our confidence interval, though we should be aware of increasing our false positive rate .

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

The following is a potential follow-up experiment to help reduce the number of students who cancel early in the course.

Change

Let the students who get enrolled in the free trial be properly informed by sending them an email frequently and regularly encouraging them and helping them with any frustrations and or disappointments, maybe within a week of them enrolling. They must be encouraged as to the time need to spend per week on the course. This timing spent on course should be emphasized.

Hypothesis

Receiving this email within a week say 5 days of enrolling in the free trial, students will be less likely to quit the course upon hearing the required time commitment. I can safely assume that if one cancels or quits it is not because they are frustrated. This change any who is intimidated by the time commitment. These students might be able to dedicate less time to be successful, or they might be able to find time in their schedule if they enjoy the course. After trying the course, they might be more likely to stick past the free trial period.

Unit of Diversion

User ID helps with the tracking of a student who is enrolled in the free trial, therefore the unit of diversion should be user ID. Since the unit of diversion should match up with how we identify users at the point where the dividing mechanism (email or no email) will be implemented.

Metrics

Invariant Metric: Number of User IDs that enrolls in a free trial, since it would be randomized between our experiment's two branches.

Evaluation metric: Retention, if receiving emails with all the information discussed above will help increase the ratio of users who pay over those who try another program. So using the Retention Metric which is a User ID based, it will ascertain that the initiative will increase the revenue by having more people sign up for the program.

References

- https://discussions.udacity.com/t/metrics-selections/42440/3

- https://discussions.udacity.com/t/retention-or-net-conversion-metric/187232/9

- http://math.stackexchange.com/questions/838107/relationship-between-binomial-and-bernoulli

- https://discussions.udacity.com/t/unit-of-diversion/179147

- https://discussions.udacity.com/t/when-to-use-bonferroni-correction/37713/6

- https://discussions.udacity.com/t/what-does-success-in-sign-test-means/169553

- https://discussions.udacity.com/t/what-exactly-are-units-of-diversion-and-invariant-metrics/161779

- http://graphpad.com/quickcalcs/binomial1/

- https://discussions.udacity.com/t/single-metrics-example-quiz/180337

- http://www.had2know.com/academics/normal-distribution-table-z-scores.html

- http://ncalculators.com/math-worksheets/calculate-standard-error.htm