A/B Testing Udacity’s Free Trial Screener

By:程泽华(Limber Cheng)

chengzehua@outlook.com

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.)

- **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:** Good invariant metric because the visits happen before the user sees the experiment, and are thus independent.Number of cookies is used as unit of diversion.
- **Number of user-ids:** Not a good invariant metric because the number of users who enroll in the free trial is dependent on this experiment. There may be some time that people chose not to enroll but without the Udacity count, they are unlikely to register. And people who chose yes must have a count. So it may be influenced by the experiment. The number of user IDs is usable as evaluation metric because it reflects whether we will reduce the number of students to continue past the free trial. Not an ideal evaluation metric because the number of visitors may be different between the experiment and control groups, which would skew the results, but it could potentially be an evaluation metric.
- **Number of clicks:** Good invariant metric because the clicks happen before the user sees the experiment, and are thus independent from it.
- **Click-through-probability:** Good invariant metric because the clicks happen before the user sees the experiment, and are thus independent from it.
- **Gross conversion:** To test the first part of our hypothesis (reducing students), we want to see whether the experiment group have practically significant lower Gross conversion than the control group, since the students enrolled in free trial aware the time commitment upfront.
- **Retention:** Not a good invariant metric because the number of users who enroll in the free trial is dependent on the experiment. Retention may be influenced by the experiment.the experiment may make some people do not complete check out. Good evaluation metric because it is directly dependent on the effect of the experiment, and also shows positive financial outcome of the change. Retention has use user ids count as denominator, so it is not a good evaluation metric for this test, though it may be an interest metric to look at in other settings.
- **Net conversion:** Not a good invariant metric because the number of users who enroll in the free trial is dependent on the experiment. Good evaluation metric because it is directly dependent on the effect of the experiment, and also shows positive financial outcome of the change.

To test the first part of our hypothesis (reducing students), we want to see whether the experiment group have practically significant lower Gross conversion than the control group, since the students enrolled in free trial aware the time commitment upfront. I will look at **Gross conversion** and **Net conversion**. The first metric will show us whether we lower our costs by introducing the screener. The second metric will show how the change affects our revenues.

In order to do the experiment, I would use **Gross conversion**, and **Net conversion**.

> For gross conversion metric, we are expecting to see that the gross conversion for the experiment group is lower than the gross conversion for the control group, since the students enrolled in free trial are aware of the time commitment upfront. So we are expecting to see this metric to pass both the statistical and practical significance test. *With the result of the Gross Conversion, we could got to know that how much does it count for student who use the free trial.*
>
> For net conversion metric, we are expecting to see that the net conversion in the experiment group is not significantly less than that of the control group. Even though we are expecting to see less students enroll in free trial in the experiment group, but if the hypothesis is true, we are expecting to see less free trial students to drop out after the trial as well. So we are hoping to see this metric not pass both the statistical and practical significance test.
>
> *P.S: This is mainly cite from ryancheunggit, udacity P7 , and I have add some of my ideas in it(**Marked in italics**).*

Measuring Standard Deviation

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

	Gross conversion	Net conversion
n	0.08* 5000 = 400	400
p	0.20625	0.1093125
SE	0.0202306	0.01560154

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 both have the number of cookies as their denominator, which is also our unit of diversion. We can therefore proceed using an analytical estimate of the variance.

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

I did not use the Bonferroni correction.

I will need 685,324 pageviews to power the experiment with these metrics.

That is, double (control + experiment groups) of the number of samples required for the more demanding of the two metrics, Net conversion.

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.

I would divert 70% of the traffic to the experiment. Given that, the experiment will take 25 days, which is a reasonable time for our needs.

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

The experiment is not extremely risky given that it does not affect existing paying customers, and is simple enough that there is a low chance of bugs occurring in the process. Nevertheless it may have a substantial impact on new enrollments, and diverting 100% of the traffic may thus not be advisable.

Risk assessment: Indicating needed time commitment is pretty harmless, it can stop student from, and the data is not sensitive neither.

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.

Item	Number of cookies	Number of clicks
experiment group	344660	28325
control group	345543	28378
SD	(0.25/(344660+345543))**0.5 = 0.0006018407	(0.25/(28325+28378))**0.5 =0.002099747
confidence interval	0.95	0.95
[0.5 - SD*1.96 , 0.5+SD*1.96]	[0.4988 ,0.5012]	[0.4959 ,0.5041]
observed	345543/(344660+345543) =0.5006; Pass	28378/(28325+28378)= 0.5005; Pass

For the click-through probability, the observed click-through probability for the control group is 0.082125814 and for the experiment group it is 0.082182441, and the pooled probability is:

$$\frac{28325+28378}{344660+345543} = 0.08215409 \tag{1}$$

The 95 percent confidence for the difference assume there is no difference between the probabilities should be:

$$[-0.001295679, 0.001295679] \tag{2}$$

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.

Item	Gross conversion	Net conversion
experiment group	3423/17260	1945/17260
control group	3785/17293	2033/17293
SE	0.004371675385	0.003434133513
m	SE * 1.96 = 0.00856848375	SE * 1.96 = 0.0067
Pooled Probability	0.2086	0.2086
D hat	-0.02055	-0.0049
Confidence Interval(CI)	[-0.0291, -0.0120]	[-0.0116, 0.0019]
Conclusions	**statistically significant, practically significant**	**not statistically significant, not practically significant**

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 metric
  - There are 4 out of 23 days on which the gross conversion is higher in the experiment group, the two-tailed p value for this sign test is 0.0026
  - **statistically and practical significant**
- Net conversion metric
  - There are 10 out of 23 days on which the net conversion is higher in the experiment group, using the online calculator, the two-tailed p value for this sign test is 0.6776
  - **statistically and practically insignificant**

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 a Bonferroni correction because we are only testing one variation. It might be useful to apply the Bonferroni correction if we decide to do post-test segmentation on the results, for example based on browser type or countries of origin.

Recommendation

Udacity should not use the change.

Make a recommendation and briefly describe your reasoning.

The metrics I was interested in were **Gross conversion** and **Net conversion**.

The result shown that Gross Conversion was practically significantly decreased.This is a good outcome because we lower our costs by discouraging trial signups that are unlikely to convert.This is a good news for Udacity team since coach can now focusing on more quality students.

Net conversion unfortunately ended up being statistically and practically insignificant and the confidence interval includes negative numbers. Therefore, there is a risk that the introduction of the trial screener may lead to a decrease in revenue. So it would be risky if we launch the change.

We should therefore consider test other designs of the screener before we decide whether to release the feature, or abandon the idea entirely.

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.

From my personal experience, the more we have paid for the bill, the deeper our preference will become. It's originate from our human nature that we hate the feeling of losing. And once we have paid our price, the remark of the stuff will change. So I could say that the student that regarded as valuable to the company, would probably love their choice. So people who have already pay for the lesson will persuade people who do not.

Udacity does a great job at offering a variety of ways for students to get help, through many different ways, for example the discussion forum, office hours, or project reviews, and the most valuable feedback is project reviews(as far as I am concern). So now,we can easily see that what features Udacity seized:

1. They require the student to make the first move, which allow each of the student to take an active learning attitude, that is the guarantee of the progress.
2. They tend to expose the student to numerous people, with no one clear point of contact throughout the Nanodegree. (In the real world, this person would be either your Study advisor, or Faculty advisor, or assigned Senior in your Freshman year at college.) So it could infer that Udacity has made the change that everyone could be one's advisor and everyone could be one's student, which also bring the community vitality so the community(let us take the forum in discussion online into account) will keep growing.

Udacity could consider implementing a similar system for the Nanodegree program. When a user joins the trial program, he or she will receive an on-site message and a subsequent email from a randomly assigned member of the Udacity team. This message will introduce them to their concierge//guru/mentor or whatever the most appropriate wording would be for Udacity’s customer base (this could potentially be A/B tested as well, if the original experiment turns out to be a success), and encourage them to reach out to this team member whenever they need help. Also an very famous research have said that, during the education process, the most important thing is to get enough feedback to student. So that is **the feedback frequency would affect the students' study process.** So with such amount of feedback resources, the student will receive more information than what have been show in the video. (Some tips for example)

> P.S : I have forgot the origin paper and it's name, what I have wrote were merely depends on my personal memory.

My **null hypothesis** is that *assigning fewer point of contact to new trial signups will not increase Retention by a practically significant amount.*

And my **alternative hypothesis** is that *assigning fewer point of contact to new trial signups will increase Retention by a practically significant amount.*

New free trial signups will randomly be assigned to a Control and an Experiment group. The experience for users in the Control group will remain unchanged. Users in the Experiment group will be assigned a random member of the Udacity team and receive an on-site onboarding message and one email follow up from that person.

The **unit of diversion** will be the **user-id**, as this change only impacts what happens after a free trial account is created.

The **invariant metric** will be the **Number of user-ids**, because the users sign up for the free trial before they are assigned a point of contact and are exposed to the new onboarding messages.

The **evaluation metric** will be **Retention**, which, if positive and practically significant, will show an increase in revenue resulting from this change.

If Retention is positive and practically significant at the end of the experiment, we can launch the new feature, and expand it with more regular follow up emails and personalized on-site messages throughout the Nanodegree program.

Resources

- Nine Common A/B Testing Pitfalls and How to Avoid Them
- A/B Test, Wikipedia
- AB Test Blog
- AB Test
- ryancheunggit, udacity P7
- Final Project Instruction