PROJECT 7 DESIGN AN A/B TEST

REFERENCES

http://mathworld.wolfram.com/BonferroniCorrection.html

courses of project 7

http://www.evanmiller.org/ab-testing/sample-size.html

TEST DESIGN

MY CHOSEN METRICS

Which of the following metrics would you choose to measure for this experiment and why? For each metric you choose, indicate whether you would use it as an invariant metric or an evaluation metric.

NUMBER OF COOKIES

This is the number of unique cookies to view the course overview page.

I take this as one of the invariant metrics or evaluation metrics. There are several reasons for me to drop this metric up:

- user IDs is essentially number of enrollments. So, this is something we could use as an evaluation metric, to test how the popped window experiment may affect the enrollments. Thus, I do not take it as an invariant metric.
- However this is not an ideal evaluation metric as it is a raw count rather than a ratio. Everyday

NUMBER OF USER-IDS

This is the number of users who enroll in the free trial.

I do not take this as one of the invariant metrics, as the number of user-ids is totally affected by the experiment. Before a person is able to register for an ID, he/she will click on the 'start free trial' button, and the information window will pop up, asking if the person may expect at 5 hours of study each week. A large percent of people, who are tied up by work or family, will probably not have time to give 5 hours on their study each week, or are afraid of that if they could, and will be hesitated to finally register for and user-ID. Thus, such thing is not an ideal candidate for invariant.

NUMBER OF CLICKS

This is the number of unique cookies to click the “Start free trial” button (which happens before the free trial screener is trigger).

I take this as one of the invariant metrics. As the popped up window only appear after people click on the ‘start free trial’ button, this can be selected as it would not be affected by people’s self-evaluation.

CLICK-THROUGH-PROBABILITY

This I the number of unique cookies to click the “Start free trial” button divided by number of unique cookies to view the course overview page.

I do not choose this as one of the invariant metrics, as it looks repeated – if the number of cookies and number clicks are both defined, then the value of click-through-probability is also defined. Thus, we do not have to bother with this metric.

GROSS CONVERSION

This is the 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.

I take it as one of the evaluation metrics. Apparently, such value will be affected by the popped-up window, and will show differences in the experimental group, compared to the controlled group, as when people look at the popped-up window, most of them who do not think they will meet up the time criteria will back out before marching to the end of enrollment.

RETENTION (GROSS CONVERSION)

This is the 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.

This is also an ideal candidates of evaluation metrics, in my opinion. Completion of enrollment does not ensure the continuation of study over the 14-day trial window even with the pre-screening done by the popped window. This metric may help us examine the effect of the pre-screening, to see if more people will remain in the experimental group.

NET CONVERSION

This is the 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.

This is the superposition of gross conversion and retention, and will be certainly taken as one of the evaluation metrics to see the overall effect of the popped-up window.

LAUNCH CRITERIA

If we can see the gross conversion go down and the net conversion keep the same or even go up, too, then we can launch the experiment.

STANDARD DEVIATION CALCULATION

For each metric you selected as an evaluation metric, estimate its standard deviation analytically. Do you expect the analytic estimates to be accurate? That is, for which metrics, if any, would you want to collect an empirical estimate of the variability if you had time?

The sample number is 5000, and the given table sample number is 40000. According to the corresponding values in the given table, the estimated standard deviations are:

Gross conversion: sqrt(0.20625*(1-0.20625)/3200)*sqrt(40000/5000) = 0.0202

Retention: sqrt(0.53*(1-0.53)/660)*sqrt(40000/5000) = 0.0549

Net conversion: sqrt(0.1093125*(1-0.1093125)/3200)*sqrt(40000/5000) = 0.0156

Yes, I expect the analytic estimates to be accurate for gross conversion and net conversion, because the denominators for those two values are both number of cookies, which is the same as unit of diversion. Never the less, I do not expect the analytic estimate to be accurate in case of retention, of which the denominator is number of users enrolled in the courseware, which is not the same as the unit of diversion. For this value, I would like to collect an empirical estimate of the variability if time permitted.

CHOOSING NUMBER OF SAMPLES GIVEN POWER

Using the analytic estimates of variance, how many pageviews total (across both groups) would you need to collect to adequately power the experiment? Use an alpha of 0.05 and a beta of 0.2. Make sure you have enough power for each metric.

As there are only 660 enrollments, I decided not to take the retention as the evaluation metric this time, as it will take too long to finish this experiment. With the sample size calculator, here we have:

Net conversion sample needed	**27413**
Page view sample needed	27413 / 0.08 = 342663
Total page view sample needed	342663 * 2 = 685325

CHOOSING DURATION VS. EXPOSURE

What percentage of Udacity's traffic would you divert to this experiment (assuming there were no other experiments you wanted to run simultaneously)? Is the change risky enough that you wouldn't want to run on all traffic?

Given the percentage you chose, how long would the experiment take to run, using the analytic estimates of variance? If the answer is longer than a few weeks, then this is unreasonably long, and you should reconsider an earlier decision.

I would like to divert 100% of the Udacity's traffic (40000/day) to this experiment, assuming there were no other experiments run simultaneously. Considering the critical elements of human subjects risk:

- is sensitive information (e.g. medical information) collected?

No, there is no sensitive information, such a medical information, personal financial information is collected in this experiment.

- could anyone be harmed by this experiment?
  No, no one will be harmed by his experiment as there is no potentially harmful practices done onto the attendants (the people who browsed the Udacity website in this case), such as receiving vaccines, taking financial risks, playing as a student of a new education method, and in their gold age.

The change is not risky to prevent me from running on all traffic, as such experiments are mild, and would not stop any students who are interested in data analysis from enrolling and paying. Thus, it will take about 17 days to collect enough data.

DATA ANALYSIS

SANITY CHECKS

Start by checking whether your invariant metrics are equivalent between the two groups. If the invariant metric is a simple count that should be randomly split between the 2 groups, you can use a binomial test as demonstrated in Lesson 5. Otherwise, you will need to construct a confidence interval for a difference in proportions using a similar strategy as in Lesson 1, then check whether the difference between group values falls within that confidence level.

If your sanity checks fail, look at the day by day data and see if you can offer any insight into what is causing the problem.

Here I choose the 95% confidence interval. The values that I calculated are shown below:

	Lower bound	Upper bound	Observed	Passes
• Number of cookies	0.4988	0.5012	0.5006	☑
• Number of user-ids				☐
• Number of clicks on "Start free trial"	0.4959	0.5041	0.5004	☑

CHECK FOR PRACTICAL AND STATISTICAL SIGNIFICANCE

Next, for your evaluation metrics, calculate a confidence interval for the difference between the experiment and control groups, and check whether each metric is statistically and/or practically significance. A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident there was a change), and it is practically significant if the confidence interval does not include the practical

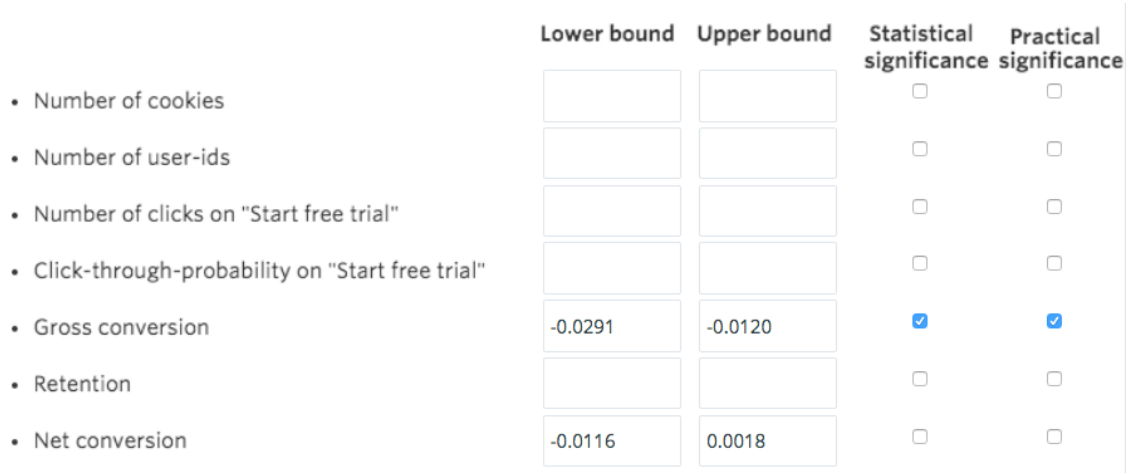
significance boundary (that is, you can be confident there is a change that matters to the business.)

If you have chosen multiple evaluation metrics, you will need to decide whether to use the Bonferroni correction. When deciding, keep in mind the results you are looking for in order to launch the experiment. Will the fact that you have multiple metrics make those results more likely to occur by chance than the alpha level of 0.05?

My calculations are:

(As there are only enrollments and payments for the first 23 days, I took them all here)

	Control group	Experimental group
Clicks	17293	17260
Enrollment	3785	3423
payments	2033	1945
Gross conversation	0.2189	0.1983
Net conversation	0.1176	0.1127



RUN SIGN TESTS

For each evaluation metric, do a sign test using the day-by-day breakdown. If the sign test does not agree with the confidence interval for the difference, see if you can figure out why.

	Gross conversion	Net conversion
number of success	19	19
Number of trials	23	23
Two-tailed p-value	0.002599	0.6776
Alpha value	0.05	0.05
Statistically significant	Yes	No
Practically significant	No	No

RESULTS SUMMARY

Here I did not use the Bonferroni correction as I used two evaluation metrics, gross conversion and net conversion here, which mean multiple hypotheses are tested.

In order to launch, we would need them both to match our expectations, in another word, we look for a decrease in gross conversion and for a no decrease in the net conversion. This is different than the situation where at least one of the them should match our expectation. For the fomer situation (our case), the risk of a Type II error increases as the number of metrics increases, for the former, the risk of a Type I error increases as the number of metrics increases.

We risk not to launch because if at least one metric (out of 2) fail to reject the null (Type II error).

Conversely if we were to launch the experiment when any metric would match our expectations (so we would launch if just one metric out of 2 does what we expect) then we would have to use Bonferroni. In an extreme condition, if there are 20 metrics, out of 20 metrics, the risk that just one rejects the null by pure chance (Type I error), would be very high. Bonferroni is designed to reduce this type of risk.

Thus, no Bonferroni correction is performed to avoid the Type II error.

MAKE A RECOMMENDATION

Finally, make a recommendation. Would you launch this experiment, not launch it, dig deeper, run a follow-up experiment, or is it a judgment call? If you would dig deeper, explain what area you would investigate. If you would run follow-up experiments, briefly describe that experiment. If it is a judgment call, explain what factors would be relevant to the decision.

Apparently, I would not like to launch this experiment, because:

1) The Gross Conversion went down by at least the practical significance boundary, which is actually good. We want to decrease the cost (whether that is monetary, or the cost of unsatisfied students) of enrollments that aren't likely to stick out, so decreased enrollment is what we were looking for.
2) For the Net Conversion, we analyzed that there has been no statistically significant change, but the confidence interval does include the negative of the practical significance boundary. That is,

it's possible that this number went down by an amount that would matter to the business. This is not acceptable risk for Udacity to launch, as it is against the second part of our hypothesis: "...without significantly reducing the number of students to continue past the free trial and eventually complete the course."

FOLLOW-UP EXPERIMENT: HOW TO REDUCE EARLY CANCELLATIONS

If you wanted to reduce the number of frustrated students who cancel early in the course, what experiment would you try? Give a brief description of the change you would make, what your hypothesis would be about the effect of the change, what metrics you would want to measure, and what unit of diversion you would use. Include an explanation of each of your choices.

- Design and describe an experiment that helps reduce early cancellations:
  - **Current problem:** The studied popped-up-window method will not help lift up the value of net conversion – no matter if the students are persuaded not to take the course by the popped-up window, or are scared away by the course itself when they actually take the course, the net conversion rate will remain the same, in other worlds, the revenue of Udacity.
  - **My experiment**: If we want to reduce the number of frustrated students who cancel early in the course, we may try to add a 'self-adaptive' function to the data analyst nano degree course study, instead of using popped-up window discussed here. The 'self-adaptive' function can be realized by a series of web pages, or a more interactive animation, where the fundamental knowledge, skills and working routine of data scientist is introduced with quiz, from simple to more complex level. Based on the answers provided by the students, the 'self-adaptive' function will list a customized courses this student needs to take, begging with the easiest one, which is closest to the current level of this student, so that this student will always feel happy and encouraged during this nano degree study, as the new course is always prepared by the course taken by this student, recommended by this 'self-adaptive' function.
- Propose and clearly state an hypothesis:
  - Hypothesis 1: The gross conversion rate is decreased.
  - Hypothesis 2: The net conversion rate remains the same or even increased. (encouraged by such self-adaptive function)
- Construct and explain proper metrics to evaluate it and explain how they inform the hypothesis:
  - Evaluation metrics: gross conversion (retention) and net conversion.
  - If gross conversion goes down, it means the hypothesis 1 is successful.
  - If net conversion goes up, it means the hypothesis 2 is successful.
  - In this experiment, we will expect both of these hypotheses to be successful.
- Propose a coherent unit of diversion and invariant metric(s).
  - Unit of diversion: cookie (before enrollment), user-id(after enrollment)
  - Invariant metrics: number of cookies, user- and number clicks.