Project 7 by: Norbert Norit

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:

1. Number of cookies:
2. Number of clicks:
3. Click-through-probability:

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.

I use the number of cookies as invariant metric because it shall not change as a cause of the experiment. Actually, since the number of cookies is defined as number of unique cookies to view the course overview page, it is clear that at storing the cookie data, the experiment has not started yet, thus there shall not happen a change caused by the experiment.
The number of cookies cannot serve as an evaluation metric because it is not affected by the experiment.

The next invariant metric I chose is the number of clicks. Since it is defined as number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger), it shall not change due to the experiment. Basically because the number of clicks is stored before the experiment. On the other hand because of this click capturing before the experiment, the click metric cannot be used as an evaluation metric.

The third invariant metric I chose is the click-through-probability Since an invariant metric is shall not change as a cause of the experiment, I looked at the numerators and denominators and investigated whether they can be influenced by the experiment. The numerator is the number of clicks which I said before is an invariant metric, thus the numerator is invariant. The same goes for the denominator which represents the number of cookies which I spotted as invariant as my first metric.

Evaluation Metrics:

1. Gross conversion
2. Retention
3. Net conversion

Gross conversion is chosen as evaluation metric because that is one thing which is definitelly affected by the experiment. The unit of analysis is basically the denominator. That is the number of unique cookies to click the "Start free trial" button. Since this metric can definitelly be affected by the experiment, it cannot be an invariant metric.

The 2nd evaluation metric I picked is the Retention. It basically represents 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. The numerator and the denominator both can be definitelly be influenced by the experiment. That is why it is an evaluation metric. Consequently this metric cannot be used as an invariant metric because the whole metric can depend on the experiment.

The 3rd evaluation metric I chose is the net conversion. It represents 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 metric is a bit different from the the two metrics mentioned before. In the sense that the unit of analysis in the metric is used before as an invariant metric. But as a fraction where the numerator is definitely affected by the experiment, it can be handled as a great evaluation metric. Since the numerator is not independent on the experiment, it is not allowed to chose it as an invariant metric.

The Results in the evaluation metrics I will look for:

Gross conversion shall go down in order to achieve the goal that less users of the free trial quit.

Retention should go up in order to launch the experiment because one aims to increase the portion of people remaining at udacity after the 14 day period relativelly to the people completing the checkout. An increasing numerator and an decreasing denominator will drive this metric up which is a good thing in order to launch.

The net conversion shall go up or at least stay constant in order to achieve udacity's goal of not reducing the amount of students which continue.

Neither Evaluation Metric nor Invariant Metric:

1. Number of user-ids:

The number of user-ids represents the number of users who enroll in the free trial. Since the enrollment process starts after the experiment, there is definitely a high probability that this metric will change as a result of the experiment. This metric cannot be used as an invariant metric because of the fact that the experiment can influence the metric of the number of user-ids.

The reason this metric cannot serve as a great evaluation metric is that the unit of diversion is a cookie in the whole experiment and the problem would be that there is no consistency in the experiment and control group in the sense of assigning to the groups because one user id can have mulitple cookies because of the possiblity that one can log in on many devices, which result in different cookies. So one user id can probably be assigned to both groups and that would be violating the necessity of right distribution among the control and experiment group.

Measuring Standard Deviation

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

1. Gross conversion Std: 0.0202
2. Retention: Std: 0.0549
3. Net conversion: Std: 0.0156

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.

For the metrics Gross conversion and Net conversion it is not necessary to estimate the variability empirically because in both cases the unit of analysis (the Unique cookies to click "Start free trial" per day) is the unit of diverson as well.

But for the Retention metric the thing is that now the unit of analysis is the enrollments per day which is tracked by the user id. Unfortunatelly this is different to the unit of diversion which remains the cookie. That is why the analytical estimate will not match up in this case. So one needs to do an empirical estimation of it.

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.)

The bonferroni correction accounts for the issue that if you have multiple metrics you want to investigate, there will be a high probability that at least one metric has a significant impact. The thing is that this significant inmact can occur by chance due to the multiple metrics, especially if they are correlated among each other in some way. In this experiment this is the case. Thus it is clearly not a bad idea to calculate the significance of each metric using bonferroni method. Nevertheless this method is pretty conservative, meaning that it then often shows a non significant effect on a metric whereas if calculating normally there is some significance.

How many pageviews per each evaluation metric are necessary?

Retention metric: I used the online calculator from the lecture to determine the number of page views necessary.
For an alpha of 0.05, a beta of 0.2, the baseline conversion rate of 0.53 and a minimum dectectable effect of 0.01, the online calculator gave me 39115 pageviews which are necessary. Nevertheless that number is not the one we are looking for. The thing is that I have to account for the probabilites that this metric will have based on 39115 pageviews. In other words, I have to account for the probability that how much of the 39115 pageviews can be used. That is why one have to divide the pageviews by Click-through-probability on "Start free trial", which is 0.08 and Probability of enrolling, given click, which is 0.20625 in order to get the necessary pageviews.
Further more one have to double that number in order to finally get the pageviews needed, because the online calculator just takes 1 group into account, but we have a control and an experiment group.
So in total we need 39115 / 0.08 / 0.20625 * 2 = 4741212 page views.

This logic holds true for calculating the page views needed for net and groos conversion as well.
Only one thing is different and that is here we divide the number of page views needed from the online calculator only by 0.08 and not with 0.20625 too, that is because here it is not the case that we have to account for the probability of enroll given click. Basically because here the unit of analysis, both for gros- and net conversion is Unique cookies to click "Start free trial" per day and not like in the retention metric the number of enrollments per day.

Pageviews needed for gross conversion: 645875
Pageviews needed for net conversion: 685325

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.)

I take a fraction of 50% for the reasons following below. Since the daily traffic is 40000 page views, 20000 will account to the experiment. Therefore 10000 to the control group and 10000 to the experiment group. So in total we need if we look at the retention metric 4741212 / 20000 = 237.0606
days. Well, if looking at the other two evaluation metrics, we would need remarkably less days and that is what I chose to implement. I discard the retention metric and stick to the others. Since the second largest pageviews needed comes from Net conversion, I decided to use this number of pageviews in order to calculate the days which we need to run the experiment.
Pageviews needed for net conversion is 685325. Deviding by 20000 gives me the days needed. So roughly 34.266 days.

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

To put it in a nutshell, one can access the fraction of traffic exposed to the experiment by analyzing the risks. For example if there is delt with sensitive data like personal information. That is not a problem in this experiment set up because only a screener is added to the user experience. Another issue could exist if the experiment can have a large duration which could make udacity suffering user leavments. But this is not the case in this experiment because the screener can be viewed and clicked through very shortly. Summing up one could even run the experiment on all traffic because these kind of risks are not existent in the experiment set up.

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.)

Invariant metric: Number of cookies

Page views in control group: 345543
Page views in experiment group: 344660
Standard deviation: sqrt(0.5 * 0.5 / (345543 + 344660)) = 0.0006018
Margin of error = 1.96 * 0.0006018 = 0.0011796
Lower bound: 0.5 - 0.0011797 = 0.4988
Upper bound: 0.5 + 0.0011797 = 0.5012
Fraction which was observed: 345543 / (345543 + 344660) = 0.5006

This metric passes my sanity check because the observed fraction of 0.5006 falls in the confidence interval. Meaning there is not a significant difference at the 95% niveau.

Invariant metric: Number of clicks

Clicks in control group: 28378
Clicks in experiment group: 28325
Standard deviation: sqrt(0.5 * 0.5 / (28378 + 28325)) = 0.0021
Margin of error = 1.96 * 0.0021 = 0.0041
Lower bound: 0.5 - 0.0041 = 0.4959
Upper bound: 0.5 + 0.0041 = 0.5041
Fraction which was observed: 28378 / (28378 + 28325) = 0.5005

This metric passes my sanity check because the observed fraction of 0.5005 falls in the confidence interval. Meaning there is not a significant difference at the 95% niveau.

Invariant metric: Click-through-probability

Average Click through probability in control group: 0,0820901744
Average Click through probability in experiment group: 0,082190521
Pooled click through probability: 0,0821540909
Pooled SE: 0,00066106
Observed difference between the click through probabilites of the groups: 0,0001003466
Margin of Error: Pooled SE * 1.96 = 0,001295679
Lower bound: -0,0011953324
Upper bound: 0,0013960256

This metric passes my sanity check because the observed value of 0,0001003466 falls in the confidence interval. Meaning there is not a significant difference at the 95% niveau. Furthermore the zero is a part of the confidence intervall, so there is no differene at the underlying sifnificance level.

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.)

I chose not to use bonferoni method because it is very conservative and will likely not show significant results because the metrics are likely to be correlated amongst each other.

Evaluation metric: Gross conversion

Average Gross conversion in control group: 0.2188746892
Average Gross conversion in experiment group: 0.1983198146
Variance Gross Conversion in control group: 0.00000988657605
Variance Gross Conversion in experiment group: 0.000009211417482
Variance of both is the sum of each individual variance: 0.00001909799252
Observed difference between the average gross conversion of the groups: -0.020554874
Margin of Error: 0.0085652
Lower bound: -0.0291
Upper bound: -0.0120

The observed value of falls in the confidence interval. Further it is practically significant because dmin= -0.01 is not icluded in the confidence interval. Since zero is not included in the confidence intervall, the difference is statistically significant at the 95% level.

Evaluation metric: Net conversion

Average net conversion in control group: 0.1175620193
Average net conversion in experiment group: 0.1126882966
Variance net Conversion in control group: 0.000005999027983
Variance net Conversion in experiment group: 0.000005793142782
Variance of both is the sum of each individual variance: 0.00001179217076
Observed difference between the average net conversion of the groups: -0.004873723
Margin of Error: 0.0067228
Lower bound: -0.0116
Upper bound: 0.0018

The confidence interval does include zero which means there is no statistically significant effect. Furthermore dmin= -0.0075 is included. Thus there is also no practical significance.

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.)

I used an online calculator to perform the sign tests.

Gross conversion:

In 4 out of 23 days one can see an improvement in the experiment group.

With probability 0.5 I get a p-value of 0.0026. That is smaller than alpha = 0.05, meaning that he change is statistical significant and occured not randomly.

Net conversion:

In 10 out of 23 days one can see an improvement in the experiment group.

With probability 0.5 I get a p-value of 0.6776. Thus it is very likely with chance of 67.76% that the result occured due to chance. So it is not statistically significant.

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 chose not to use bonferoni method because our goal of the experiment was to achieve a significant result in both metrics in order to launch the experiment. The thing is that if we look at somehow similar metrics, there is pretty high chance that a significant result occured randomly. But that would be a problem which Bonferoni method would account for. Nevertheless Bonferroni method is likely to be very conservative thus calculating significance more rarly. Bonferroni would be great if we just based our decision of at least one metric to be significant. That is not the case because we want both metrics to be significant in order to launch the experiment for the whole.

For gross conversion we see that the sign test and the effect size test generate the same conclusion. Further for net conversion the results of the different tests match up too. So no additional investigation is needed.

- Gross conversion:

  Statistically significant in sign test

  Statistically significant in Effect size test and practically significant.

- Net conversion:

  Not significant in sign test.

  Not statistically significant in Effect size test and not practically significant.

  I think the discrepancy occured heavily because if one assume equal probability and there are 10 succeses out of 23, it is clear that this cannot occur not random, thus there will always be a not significant sign test showing up.

  Further I conclude that the effect size test does not account for random distribution like the sign test does.

Recommendation

Finally the screener achieved to reduce the number of students who quit after the free trail. Nevertheless it affected the number of students to stay past the trial negativelly which is not what we wanted. Based on this, we should not launch the experiment for the whole.

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.

In order to reduce the number of people who quit after the 14 day period, one can try to increase the motivation of the students in some sort of way. One could use a banner which pops up after the log in if the student who did not dedicate a certain amount of time to learning at udacity. So buzy students will not see the banner. The banner will show a picture of a student with the text „I was hired by Tesla" because of this Nano degree I accomplished".
The metric I will messure is retention. Retention should not show a difference in the controll group but definitelly it will have an influence on the experiment group. As unit of diversion I will use the user id because it is easy to track and is most suitbable here because we are investigating students only who allready created an account by enrolling and thus having a user id.
I am looking for an increasing retention metric in the experiment group.