P7 – Michail Kovanis

Experiment Design

Metric Choice

For this experiment, I can select out of 7 metrics to be either invariant or evaluation metrics. These metrics are:

1. The number of cookies (d_{min} = 3000)
2. The number of user-ids (d_{min} = 50)
3. The number of clicks (d_{min} = 240)
4. The click-through-probability (CTR) (d_{min} = 0.01)
5. The gross conversion (d_{min} = 0.01)
6. The retention (d_{min} = 0.01)
7. The net conversion (d_{min} = 0.0075)

Where d_{min} refers to the “difference that would have to be observed before that was a meaningful change for the business”.

Since our unit of diversion is a cookie, then the first invariant metric should be the number of cookies. In addition, good invariants are the number of clicks and the CTR because they measure events happening before the user enters the experiment, thus they are unlikely to be affected by it. The rest of the metrics consider events that happen after the user is exposed to the experiment and thus very likely not to remain invariant.

As evaluation metrics, I chose the gross and net conversion. These metrics are good for evaluation as they can show whether the new message helped to screen those that couldn’t dedicate at least 5 hours per week and how this affected those that would turn out to be paying customers. Retention could be an evaluation metric, but it would require the experiment to run for far longer than the previous two require and is unlikely that it will present us with a lot more information than net and gross conversion (for the purposes of this experiment). The number of user-ids wouldn't be a helpful metric because it is an absolute and not relative value.

Finally, to launch the experiment I will look whether the gross conversion decreases significantly without any significant decrease on the net conversion.

Measuring Standard Deviation

Since the evaluation metrics represent probabilities of an event that might happen or not, this should follow a binomial distribution and the theoretical standard deviation being equal to:

SQRT[p*q/N] = SQRT[p(1-p)/N]

Where p is the probability of the event and N the total number of trials (number of clicks coming from unique cookies). In our case the standard deviations for the two evaluation metrics are:

1. Gross conversion STD: 0.0202
2. Net conversion STD: 0.0156

For both evaluation metrics, the denominator of the metrics is the number of cookies that clicked the button. Since this is also the unit of diversion, the analytic estimate of the variance should be comparable to the empirical one.

Sizing

Number of Samples vs. Power

Using a = 0.05 and β = 0.2 (without the Bonferroni correction) as inputs to the online sample size calculator I see that I need 25,835 events for the gross and 27,413 events for the net conversion metric. Thus, at minimum I will need 27,413 events (or clicks). If 8% of the page-views result in clicks, then I will need 342,663 page-views.

Since the groups are two (control and experimental) this number needs to be multiplied by 2 and the total number of page-views should be equal to 685,325 to power the experiment appropriately.

Duration vs. Exposure

I would divert the whole traffic to the experiment (half to the control and half to the experimental group), which is about 20,000 cookies daily, and it will require 18 days to obtain 685,325 page-views. Since I need to obtain the same number of page-views for both groups, diverting half the traffic to the experiment will allow it to run for the shortest time-period.

This intervention appears before the user registers for an account and the unit of diversion is an anonymous cookie, therefore there are no concerns of sensitive user-data being collected. Moreover, this experiment is highly unlikely to cause any “physical, psychological and emotional, social, and economic concerns” and an informed consent would not be necessary. Thus, this experiment poses a minimal risk to Udacity and its users.

Experiment Analysis

Sanity Checks

For performing sanity checks on the invariant metrics, I first need to compute the standard error, then the margin of error and finally the upper and lower bounds. This procedure happens as follows:

1. SE:
    a. Number of cookies: SQRT [0.5 * (1 - 0.5) / (CO_{Con} + CO_{Exp})]
    b. Number of clicks: SQRT [0.5 * (1 - 0.5) / (CL_{Con} + CL_{Exp})]
    c. CTR: SQRT [(CL_{Con} / CO_{Con}) * (1 - CL_{Con} / CO_{Con}) / CO_{Con}]
2. Margin of error (m): 1.96*SE
3. Upper and lower bounds:

    a. Number of cookies: [0.5 – m, 0.5 + m]
    b. Number of clicks: [0.5 – m, 0.5 + m]
    c. CTR: [(CL_{Con} / CO_{Con}) + m, [(CL_{Con} / CO_{Con}) + m]

4. Observed:
    a. Number of cookies: CO_{Exp} / (CO_{Con} + CO_{Exp})
    b. Number of clicks: CL_{Exp} / (CL_{Con} + CL_{Exp})
    c. CTR: CL_{Exp} / CL_{Exp}

Where CO_{Con}, CO_{Exp} are the total number of cookies and CL_{Con}, CL_{Exp} the total number of clicks in the control and experimental groups respectively.

The 95% confidence intervals and the observed values for all invariant metrics are:

1. Number of cookies:
    a. CI: [0.4988, 0.5012]
    b. Observed: 0.5006
2. Number of clicks:
    a. CI: [0.4959, 0.5041]
    b. Observed: 0.5005
3. CTR:
    a. CI: [0.0812, 0.0830]
    b. Observed: 0.0822

All observed values of the invariants are inside the 95% confidence intervals and thus the metrics pass the sanity checks.

Result Analysis

Effect Size Tests

Using the formulas for the pooled x hat and the pooled SE (from lesson 1; pooled standard error) I obtained the 95% confidence intervals for the evaluation metrics, which can be seen below:

1. Gross conversion: [-0.0291, -0.0120] (Statistically and practically significant)
2. Net conversion: [-0.0116, 0.0019] (95% CI include 0, thus change is neither statistically nor practically significant)

Sign Tests

To perform a sign test, I counted the number of days that the number of enrolments or payments were higher in the experimental than the control group. For the enrolments, it was 4 days and for the payments it was 10 days. Thus, the respective two-tail p values were 0.0026 (statistically significant) and 0.6776 (not statistically significant).

Summary

For the above analysis, I didn’t use the Bonferroni correction. The Bonferroni correction is generally necessary when having multiple metrics and one would need any of them to fulfil a condition. Then the probability of having rejecting the null hypothesis just by chance is higher (Type I error). However, here I used two metrics and I required that both fulfil a condition, which means that the risk of Type I error is not very high, therefore the Bonferroni correction wasn’t necessary. Finally, both the effect size hypothesis tests and the sign tests showed significance in the gross conversion and not in the net conversion, therefore no discrepancies between these two exist.

Recommendation

The results of the experiment show that the intervention produced a statistically significant drop in the gross conversion rate without significantly decreasing the net conversion rate. Even though the lower bound of the 95% confidence interval of the net conversion is below the negative practical significance boundary, this is not a significant result because the range includes the 0. This means that the intervention is successful into screening out from the free trial people that would enrol to it but not continue to be paying customers. Thus, I would recommend the intervention to be adopted.

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.

I would perform a follow-up experiment to evaluate an intervention to reduce early cancelation rate. An early cancelation can be defined as students that make one payment but cancel before the second. Those early cancelations may happen for a variety of reasons, however in this experiment I will focus only on those that cancel due to frustration from the course being too difficult for them.

The intervention would be that users who click the “Start free trial” button are presented with a pop up which includes the minimum requirements for a student to be able to successfully attend and complete the course. The users would need to confirm that they have the necessary skillset, otherwise they would be prompted to access the course material instead. The main hypothesis is that users who would like to enrol, but don’t have the required skillset yet will spend some time first to raise it to the necessary level for the course. Thus, they would be more unlikely to cancel due to a mismatch between their skillset and the course requirements.

The cookie is an appropriate unit of diversion (and invariant) since the intervention would start before a new user registers for an account. The number of clicks and the click-through-rate refer to events happening before the intervention, thus they are also appropriate invariant metrics. The evaluation metrics will be the net conversion and, a new metric, the two-month net conversion (probability to make two payments per cookies that click the button). We expect the net conversion rates to drop and we would like the two-month net conversion to raise. However, the two-month net conversion is a variable that would require in the most optimal case an extra month to obtain than the other two (since number of samples will be decrease the expected time should be around ~3 months extra than the first experiment). Since that time is long and in the case this experiment proves to be risky, we could use the other evaluation metric to assess whether we stop the experiment at the 35 days or not. For example, net conversion is expected to drop, but the users that do not go beyond the trial should not be more than those that cancel early. Thus, we could set a threshold below which if the net conversion falls then the experiment should be stopped.

References

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

Sign test calculator: http://graphpad.com/quickcalcs/binomial1.cfm

Bonferroni correction: http://onlinelibrary.wiley.com/doi/10.1111/opo.12131/full