Project P7: Design an A/B Test

Peter Eisenschmidt

April 14, 2017

1 Experiment Design

1.1 Metric Choice

The following metrics have been selected as **Invariant Metrics**:

Number of cookies

The number of unique cookies to visit the page should not be affected by the experiment, as someone visiting the course overview page has not seen the changes yet.

Number of clicks

The same applies to the number of clicks on the ”Start Free Trial” button; there should not be any impact of the experiment on this metric

Click-through-probability

As the CTR is defined as the number of unique cookies to click the ”Start free trial” button divided by the number of unique cookies to view the course overview page (both of which are invariant metrics), the click-through-probability is also an invariant metric

These three metrics are not impacted by the experiment and hence one can expect similar distributions between control and experiment groups.

The following metrics have been selected as **Evaluation Metrics**:

Gross conversion

Being defined as the number of user-ids to complete checkout and enroll in the free trial divided by the number of unique cookies to click the ”Start free trial” button, one would expect a lower gross conversion for the experiment as for the control group. The goal of the tested change is to reduce the number of frustrated students, so you could expect that students that are likely to drop out with the current design are filtered out early and do not complete the checkout.

Retention

Similarly, you would expect an increased retention as a result of the experiment, as the number of students that complete the checkout should reduce. At the same time, the number of students to make at least one payment should remain the same.

Net conversion — Net conversion is the combination of the two previously mentioned metrics. It is expected that net conversion remains the same for both control and experiment group, as the number of students to remain enrolled past the 14-day boundary as well as the number of unique cookies to click the "Start Free Trial" Button should remain the same.

For each of these metrics, a practical significance boundary d_{min} is defined. This indicates the minimum difference that needs to be observed between control and experiment group in order to determine whether the change is meaningful or not. This is important for the decision to whether or not launch the change.

For the above given evaluations metrics, the practical significance boundaries are $d_{min} = .01$ (for gross conversion and retention) and $d_{min} = .0075$ (for net conversion)

1.2 Measuring Standard Deviation

The analytical estimate of the standard deviation can be calculated as follows:

$$\sigma = \sqrt{\frac{p(1-p)}{N}} \tag{1}$$

where the probabilities are given in the baseline values:

- Probability of enrolling, given click (Gross Conversion): $p = .20625$
- Probability of payment, given enroll (Retention): $p = .5300$
- Probability of payment, given click (Net Conversion): $p = 0.1093125$

Given that the sample size to visit the course overview page is 5000 cookies, the number of units of analysis for each metric can be calculated as follows. For gross conversion, it is given by:

$$N = \frac{PageViews \times Cookies_{ClickFreeTrial}}{Cookies_{ViewPagePerDay}} = \frac{5000 \times 3200}{40000} = 400 \tag{2}$$

For retention it can be calculated as:

$$N = \frac{PageViews \times Enrollments}{Cookies_{ViewPagePerDay}} = \frac{5000 \times 660}{40000} = 82.5 \tag{3}$$

For net conversion, it is the same as gross conversion:

$$N = \frac{PageViews \times Cookies_{ClickFreeTrial}}{Cookies_{ViewPagePerDay}} = \frac{5000 \times 3200}{40000} = 400 \tag{4}$$

This results in the following standard deviations:

- Gross Conversion: $\sigma = .0202$

- Retention: $\sigma = .0549$
- Net Conversion: $\sigma = .0156$

For both gross and net conversion, the unit of analysis and the unit of diversion are the same (cookies), whereas the unit of analysis for retention is User ID. Therefore, the analytic estimate of the standard deviation is likely to be comparable to the empirical standard deviation for gross and net conversion but not for retention. For the latter it might be interesting to do an empirical estimate.

1.3 Sizing

1.3.1 Number of Samples vs. Power

In order to determine the number of samples, this calculator `http://www.evanmiller.org/ab-testing/sample-size.html` is used. For all three metrics, $1 - \beta$ is 80% and α is 5%, i.e. no Bonferroni correction is applied.

The baseline conversion rate and the minimum detectable effect d_{min} is listed below for each metric as well as resulting number of samples.

- Gross conversion:
  - Baseline conversion: 20.625%
  - Minimum detectable effect: 1%
  - Samples = 25,835
- Retention:
  - Baseline conversion: 53%
  - Minimum detectable effect: 1%
  - Samples = 39,115
- Net conversion:
  - Baseline conversion: 10.93125%
  - Minimum detectable effect: .75%
  - Samples = 27,413

The number of pageviews can then be calculated as follows for gross and net conversion:

$$N_{PV} = 2 \times n_{Samples} \times \frac{Cookies_{ViewPagePerDay}}{Cookies_{ClickFreeTrial}} \tag{5}$$

For retention it is given by

$$N_{PV} = 2 \times n_{Samples} \times \frac{Cookies_{ViewPagePerDay}}{Enrollments} \tag{6}$$

The resulting page views are:

Gross Conversion	645,875
Retention	4,741,212
Net Conversion	685,325

Therefore, a total number of 4,741,212 page views is required if all metrics are to be used.

1.3.2 Duration vs. Exposure

The change tested in this experiment is a low risk for the participants (no collection of sensitive data, no exposure to physical harm as the change consists of asking the student how much time per week they were willing to invest in the course). Therefore, it can be assumed to be safe to divert 100% of the traffic to this experiment. With 40,000 page views per day, this results in the following durations:

- Duration (Gross conversion): 17 days
- Duration (Retention): 119 days
- Duration (Net conversion): 18 days

119 days is not feasible for this experiment, therefore retention is not retained as a metric. The resulting experiment duration is then 18 days.

2 Experiment Analysis

2.1 Sanity Checks

For **cookies**, there is a 50% probability of being either in the control or in the experiment group. In the experiment and control group there were 344,660 and 345,543 page views respectively. The standard deviation is therefore:

$$\sigma = \sqrt{\frac{2p}{N_{exp} + N_{cont}}} = .00060 \tag{7}$$

The margin of error (for a 95% confidence interval, i.e. z = 1.96) is then:

$$m = \sigma \times z = 0.00118 \tag{8}$$

Consquently, the upper and lower bound for cookies are 0.49882 and 0.50118 respectively.

The observed value is

$$p = \frac{N_{cont}}{N_{exp} + N_{cont}} = .50064 \tag{9}$$

which is in between the lower and upper bound.

For the **number of clicks**, the probability is also 50%. The number of clicks in the experiment and the control group are 28325 and 28378 respectively. The resulting standard deviation is therefore:

$$\sigma = \sqrt{\frac{2p}{N_{exp} + N_{cont}}} = .00210 \tag{10}$$

The margin of error is then:

$$m = \sigma \times z = 0.00412 \tag{11}$$

Consquently, the upper and lower bound for number of clicks are .49588 and .50412 respectively.

The observed value is

$$p = \frac{N_{cont}}{N_{exp} + N_{cont}} = .50047 \tag{12}$$

which also falls within the boundaries of the confidence interval.

Lastly, the observed **Click-through-probability** of experiment and control group are .08219 and .08213 respectively. The standard deviation for the control group is

$$\sigma = \sqrt{\frac{CTP_{cont} \times (1 - CTP_{cont})}{N_{cont}}} = .00047 \tag{13}$$

and the margin of error:

$$m = \sigma \times z = .00092 \tag{14}$$

which results in lower and upper bounds of .08121 and .08304. The CTP of the experiment group is within this confidence intervall.

It can therefore be concluded that all invariant metrics pass the sanity checks.

2.2 Result Analysis

2.2.1 Effect Size Test

The **Gross Conversion** of experiment and control group are:

$$\begin{aligned} GC_{exp} &= \frac{X_{exp}}{N_{exp}} = .19832 & (15) \\ GC_{cont} &= \frac{X_{cont}}{N_{cont}} = .21887 & (16) \end{aligned}$$

where X_{exp} and X_{cont} are the number of enrollments in both groups. Please note that only rows are used where Enrollments are not null. This results in $N_{exp} = 17,260$ and $N_{cont} = 17,293$.

The observed difference is therefore:

$$\hat{d} = GC_{exp} - GC_{cont} = -.02055 \tag{17}$$

The pooled probability is hence:

$$\hat{p}_{pool} = \frac{X_{exp} + X_{cont}}{N_{exp} + N_{cont}} = .20861 \tag{18}$$

The pooled standard error is then given by:

$$SE_{pool} = \sqrt{\hat{p}_{pool} \times (1 - \hat{p}_{pool}) \times \left(\frac{1}{N_{exp}} + \frac{1}{N_{cont}}\right)} = .00437 \tag{19}$$

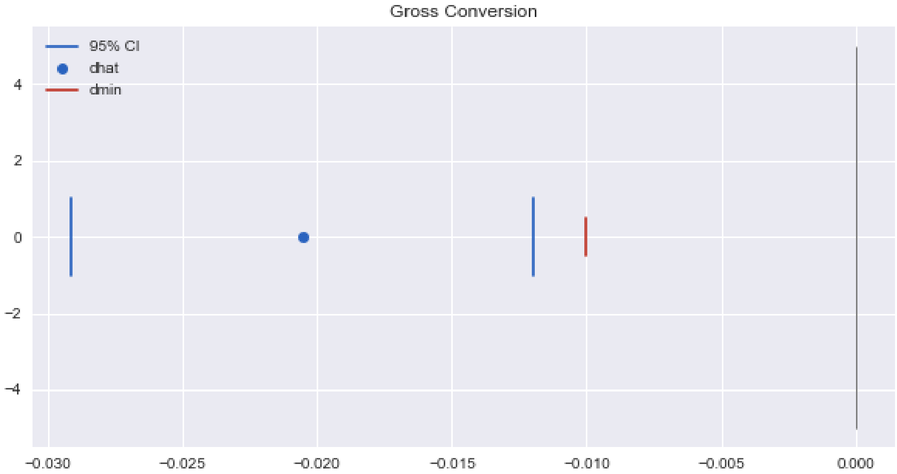
which in turn gives a margin of error (with $z = 1.96$ for a 95% confidence interval):

$$m = z \times SE_{pool} = .00857 \tag{20}$$

The confidence interval for **Gross Conversion** is therefore:

$$CI = (-0.02912, -0.01199) \tag{21}$$

This means that gross conversion is both statistically and practically significant (recall that the practical significance boundary is .01).



Similarly, the **Net Conversion** of experiment and control group are:

$$NC_{exp} = \frac{X_{exp}}{N_{exp}} = .11269 \tag{22}$$

$$NC_{cont} = \frac{X_{cont}}{N_{cont}} = .11756 \tag{23}$$

where X_{exp} and X_{cont} are the number of payments in both groups. As before, only rows where payments are not null are used.

The observed difference is then:

$$\hat{d} = NC_{exp} - NC_{cont} = -.00487 \tag{24}$$

The pooled probability is hence:

$$\hat{p}_{pool} = \frac{X_{exp} + X_{cont}}{N_{exp} + N_{cont}} = .11513 \tag{25}$$

The pooled standard error is then given by:

$$SE_{pool} = \sqrt{\hat{p}_{pool} \times (1 - \hat{p}_{pool}) \times \left(\frac{1}{N_{exp}} + \frac{1}{N_{cont}}\right)} = .00343 \tag{26}$$

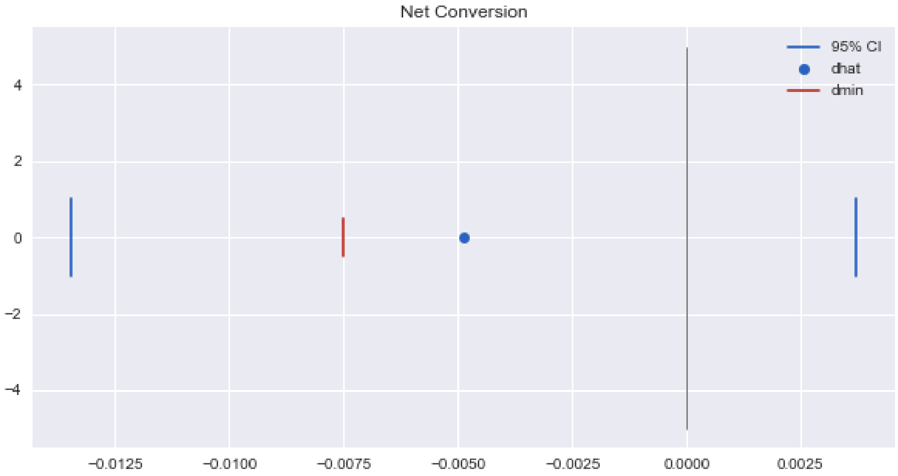
which in turn gives a margin of error (with $z = 1.96$ for a 95% confidence interval):

$$m = z \times SE_{pool} = .00673 \tag{27}$$

The confidence interval for **Net Conversion** is therefore:

$$CI = (-0.01160, 0.00186) \tag{28}$$

As zero is contained within the net conversion confidence intervall, the results are neither statistically or practically significant.



2.2.2 Sign Test

In order to perform the sign test, the number of days with a positive change, i.e. the gross and net conversion of the experiment group is greater than those of the control group, are counted. The results are

Gross Conversion Number of successes: 4, total number days: 23

Net Conversion Number of successes: 10, total number days: 23

Using `http://graphpad.com/quickcalcs/binomial1.cfm`, this gives the following two-tail P values:

$$\begin{aligned} p_{GC} &= .0026 < \alpha = .05 \qquad (29) \\ p_{NC} &= .6776 > \alpha = .05 \qquad (30) \end{aligned}$$

This indicates that gross conversion is statistically significant whereas net conversion is not.

2.2.3 Summary

Bonferroni correction was not used, as only two metrics are used. So, if gross and net conversion were completely independent, the probability of a false positive would be 9.75%. However, both metrics are correlated (students that make a payment need to enrol first), so 9.75% is too high. Therefore, applying a Bonferroni correction would be too conservative.

The effect size tests yields that gross conversion is statistically and practically significant. Net conversion is neither. The statistical significance of gross conversion is confirmed by the sign test. It also shows that net conversion is not statistically significant.

2.3 Recommendation

The experiment shows that the number of students to enroll in the 14-day trial are effectively reduced, which was one of the goals. However, at the same time the number of students to continue past the free trial and make at least one payment should not be reduced. This could not be shown in the experiment as the practical significance boundary for net conversion ($d_{min} = .0075$) falls within the confidence interval. So, there is a risk that the proposed change also negatively affects the number of students to make a payment. Therefore, the recommendation is not to launch the change.

3 Follow-Up Experiment

The initial experiment proved successful in terms of reducing the number of students to enrol in the free trial. However, it seems that it also reduces the number of students to continue past the 14-day trial. It is possible that there a students that initially indicate that they would spend more than 5 hours per week but for whatever reason do not invest the required amount of time.

The follow-up experiment could consist in a check after the first week; if a student spends less than 5 hours, another screener is displayed to ask whether the student would like to schedule a one-to-one meeting with a coach. The idea is that this motivates the student to invest more time in the second week and subsequently continues past the free trial.

As this concerns only students that have already enrolled in the free trial, the **Unit of Conversion** would be User ID. The metric for the follow-up experiment would be **Retention**, as the goal would be to see an increased number of students to make a payment.