Design an A/B Test

Chun Zhu

Experiment Design

Metric Choice

- `Number of cookies`: invariant metric

  `Number of cookies` is the number of unique cookies to view the course overview page (dmin=3000). Since the experiment doesn't affect the overview page, this metric should be invariant.

- `Number of user-ids`: none

  `Number of user-ids` is the number of users who enroll in the free trial (dmin=50). Since the enrollment may be affected by the experiment, I expect to see some differences of this metric between control group and experimental group. This metric is not a good choice for evaluation metric either because the number of enrolled users is strongly correlated with the number of "*start free trial"* clicks on a given day.

- `Number of clicks`: invariant metric

  `Number of clicks` is the number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). (dmin=240) Since this event happens before the free trial screener is triggered, it should not affected by the experiment. This metric should be invariant.

- `Click through probability`: invariance metric

  `Click through probability` is the number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page (dmin=0.01). Since `Number of clicks` and `Number of cookies` are both considered invariant metric, the ratio of the 2 variables should also be invariant.

- `Gross conversion`: evaluation metric

  `Gross conversion` 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. (dmin= 0.01) The

number of user-ids to enroll in the free trial may be influenced by the experiment. At the same time, `Gross conversion` marginalizes variances in the empirical count of user-ids. So `Gross conversion` is a good evaluation metric.

- `Retention`: evaluation metric

  `Retention` 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. (dmin=0.01) Like `Gross conversion`, `Retention` is affected by the experiment, which makes it OK for evaluation metric. However, I won't use it in this test since it takes a really long time to collect the data. I will explicitly illustrate this problem in the analysis later.

- `Net conversion`: evaluation metric

  `Net conversion` 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. (dmin= 0.0075) Actually `Net conversion` is the most interesting parameter. The ultimate goal of this A/B test is to check whether the change will increase `Net conversion`. It is a good evaluation metric.

To launch the experiment, I will require `Gross conversion` to have a practically significant decrease, and `Net conversion` to have a statistically significant increase.

Measuring Standard Deviation

For the 2 evaluation metrics I chose, the unit of diversion and the unit of analysis are both the number of unique cookies to click the "Start free trial" button. So the analytic variance is likely to match the empirical variance for both `Gross conversion` and `Retention`.

Given a sample size of 5000 cookies visiting the courseview page and assuming these 2 metrics follow Binomial distribution, the analytic estimate of the standard deviations for the 2 evaluation metrics are calculated below:

- `Gross conversion`

$$p = 0.20625\ (given\ in\ the\ table\ of\ baseline\ values)$$
$$N = 5000 \times 0.08 = 400$$
$$sd = \sqrt{\frac{p(1-p)}{N}} = 0.0202$$

- `Net conversion`

$$p = 0.1093125\ (given\ in\ the\ table\ of\ baseline\ values)$$

$$N = 5000 \times 0.08 = 400$$

$$sd = \sqrt{\frac{p(1-p)}{N}} = 0.0156$$

Sizing

Number of Samples vs. Power

I will not use the Bonferroni correction in the analysis since I think these 2 evaluation metrics are highly correlated which will give too conservative results in Bonferroni correction. The online calculator is used to calculate the sample sizes to power the experiment appropriately. Here $\alpha = 0.05, \beta = 0.2$.

Evaluation Metric	Baseline conversion rate	Minimum Detectable Effect (dmin)	Sample size needed	Number of pageviews needed (both control and experimental groups)
`Gross conversion`	20.625%	1%	25,835	645,875
`Retention`	53%	1%	39,115	4,741,212
`Net conversion`	10.93125%	0.75%	27,413	685,325

From the above table, we can calculate the time needed for data collection. The number of unique cookies to view page per day is 40000, so it will take 117 days of complete site traffic to achieve the necessary number of pageviews for retention metric. This is too long for A/B testing. So retention is discarded for evaluation metric. While with `Gross conversion` and `Retention` as evaluation metrics, it only takes 18 days to gather the data at complete site traffic. The number of pageviews needed is 685325.

Duration vs. Exposure

I will divert 100% traffic to the test subjects, with 50-50 split between the control and experimental groups. So each group will have 50% traffic. If the experiment tends to have a negative impact on the business, only 50% of the visitors will be at risk. This risk is tolerant. If we reduce the risk the experiment duration will be lengthened from 18 days, which is not desirable for A/B test.

Experiment Analysis

Sanity Checks

- `Number of cookies`: invariant metric

$$N_{control} = 345543, \qquad N_{experimental} = 344660$$
$$sd = \sqrt{\frac{0.5 \times 0.5}{N_{control} + N_{experimental}}} = 0.0006018$$
$$ME = 1.96 \times sd = 0.0011796$$
$$lower\ bound = 0.5 - ME = 0.4988\ (95\%\ confidence\ interval)$$
$$upper\ bound = 0.5 + ME = 0.5012\ (95\%\ confidence\ interval)$$
$$observed = \frac{N_{control}}{N_{control} + N_{experimental}} = 0.5006$$

The observed value is within the bounds, so `Number of cookies` passes the sanity check.

- `Number of clicks`: invariant metric

$$N_{control} = 28378, \qquad N_{experimental} = 28325$$
$$sd = \sqrt{\frac{0.5 \times 0.5}{N_{control} + N_{experimental}}} = 0.0021$$
$$ME = 1.96 \times sd = 0.0041$$
$$lower\ bound = 0.5 - ME = 0.4959\ (95\%\ confidence\ interval)$$
$$upper\ bound = 0.5 + ME = 0.5041\ (95\%\ confidence\ interval)$$
$$observed = \frac{N_{control}}{N_{control} + N_{experimental}} = 0.5005$$

The observed value is within the bounds, so `Number of clicks` passes the sanity check.

- `Click through probability`: Invariance metric

$$control\ value = \frac{28378}{345543} = 0.0821258, \qquad N_{experimental} = 344660$$
$$sd = \sqrt{\frac{0.0821258 \times (1 - 0.0821258)}{N_{experimental}}} = 0.000468$$
$$ME = 1.96 \times sd = 0.00092$$
$$lower\ bound = 0.0821258 - ME = 0.0812\ (95\%\ confidence\ interval)$$
$$upper\ bound = 0.0821258 + ME = 0.0830\ (95\%\ confidence\ interval)$$
$$observed = \frac{28325}{344660} = 0.821824$$

The observed value is within the bounds, so `Click through probability` passes the sanity check.

Result Analysis

Effect Size Tests

- `Gross conversion`

$$p = \frac{3785 + 3423}{17293 + 17260} = 0.2086$$
$$sd = \sqrt{p(1-p)\left(\frac{1}{17293} + \frac{1}{17260}\right)} = 0.00437$$
$$ME = 1.96 \times sd = 0.0085652$$
$$d = \frac{3423}{17260} - \frac{3785}{17293} = -0.02055$$
$$lower\ bound = d - ME = -0.0291\ (95\%\ confidence\ interval)$$
$$upper\ bound = d + ME = -0.0120\ (95\%\ confidence\ interval)$$

This metric is statistically significant since the 95% confidence interval does not include 0. And it is practically significant because it does not include the practical significant boundary $[-0.01, 0.01]$.

- `Net conversion`

$$p = \frac{2033 + 1945}{17293 + 17260} = 0.1151$$
$$sd = \sqrt{p(1-p)\left(\frac{1}{17293} + \frac{1}{17260}\right)} = 0.00343$$
$$ME = 1.96 \times sd = 0.0067228$$
$$d = \frac{1945}{17260} - \frac{2033}{17293} = -0.0048$$
$$lower\ bound = d - ME = -0.0116\ (95\%\ confidence\ interval)$$
$$upper\ bound = d + ME = 0.0019\ (95\%\ confidence\ interval)$$

This metric is not statistically significant since the 95% confidence interval include 0. And it is not practically significant either because it overlaps with practical significant boundary $[-0.0075, 0.0075]$.

Sign Tests

The online calculator is used to perform the sign tests.

Evaluation Metric	Number of Success (Number of days with improvement in experiment compared with control group for the evaluation metric)	p-value	Statistically significant ($< \alpha$)
`Gross conversion`	4	0.0026	True
`Net conversion`	10	0.6776	False

Summary

I do not use the Bonferroni correction in the analysis since I think these 2 evaluation metrics are highly correlated which will give too conservative results in Bonferroni correction. Based on the results of the effective size test and the sign test, `Gross conversion` will decrease while `Net conversion` will not be significantly impacted.

Recommendation

I do not recommend launching the proposed change of introduction of the trial screener as the A/B test shows that there will not have a significant increase on `Net conversion`. `Gross conversion` decreased as expected which means we would lower our costs by discouraging trial signups that are unlikely to convert. However, the most important `Net conversion` does not significantly increases with the change. Furthermore, the 95% confidence interval of `Net conversion` contains negative values, which means that there is a risk that the change may lead to a decrease in revenue. So this change does not meet the business goal and should not be launched.

Follow-Up Experiment

To reduce the number of frustrated students who cancel early in the course, I think the follow-up experiment can be adding “enroll now with a discount” option right after the user click the “start the free trial” button. This feature allow students skip free trial portion and in exchange they get a tuition discount. If the student finishes the course within a set time frame, the student can get the tuition discount. So the number of enrollments may increase as students who are already determined to take the course will enroll directly and will not have to make a rush decision in just 14 days.

My null hypothesis is that by providing a tuition discount with direct enrollment, `retention` will not increase. The components of the analysis are the following:

- Unit of diversion: `user-ids`
  `User-id` identifies the user with a free-trial account. Since my proposed change only impacts what happens after a free-trial account is created, so `user-ids` can be the unit of diversion.

- Invariant metric: `Number of user-ids`
  The users sign up for free trial before they see the “enroll now with discount” option, so `Number of user-ids` should not be affected by the change.

- Evaluation metric: `Retention`
  If `retention` is positive and practically significant, it means that the change will increase the revenue.

If `retention` is positive and practically significant at the end of the experiment, we can launch the new feature. Further experiments can be done on the discount value offered.