A/B Testing

Ada Lee

December 17, 2015

Experiment Design

Metric Choice

Candidate metrics:

- `Number of cookies`: Since unit of diversion is cookies, so this should be invariant between experiment group and control group.
- `Number of user-ids`: This may decrease in experiment group, and number is not a good evaluation metrics because it is not normalized. So number of user-ids is neither invariant metrics nor good evaluation metrics.
- `Number of clicks`: Since cookies are randomly assigned to two groups, and clicks happen before our experiment, so this should be invariant.
- `Click-through-probability`: Since number of cookies and number of clicks are invariant, click-through-probability equals to number of clicks divided by number of cookies, so this is invariant as well.
- `Gross conversion`: In experiment group, when people click "Start free trial" button, if they have less than 5 hours' time in a week, they are suggested not to enroll, so number of user-ids to enroll will probable decrease, and we already know number of cookies is invariant, so gross conversion may decrease. Gross conversion is about probability to succeed, so this is good evaluation metrics.
- `Retention`: People who click the "Start free trial" button and have less than 5 hours a week is suggested not to enroll, this would decrease number of people who enroll and are less likely to pay, so retention would increase in experiment group. Retention is about probability to succeed, so this is good evaluation metrics.
- `Net conversion`: This may decrease in experiment group since some people would click "Start free trial" button and has less than 5 hours each week are suggested not to enroll, thus would decrease people who enroll, thus may decrease number of people pay. Net conversion is about probability to succeed, so this is good evaluation metrics.

In summary, invariant metrics that I choose are `number of cookies`, `number of clicks`, `click-through probability`, good evaluation metrics are `gross conversion`, `retention` and `net conversion`. In our experiment, we want to reduce number of people who don't pay because they could not investigate enough time, but we don't want number of payments decreases, so we don't want accidentally decrease number of users that would pay but they claimed they have less than 5 hours' time each week and were recommend not to enroll and thus not pay due to our recommendation. We would launch the experiment only when we reduce number of enrollments but do not reduce number of payments, this is equivalent to gross conversion significantly decrease but net conversion does not decrease. Since retention equals to net conversion divided by gross conversion, retention would significantly increase.

Measuring Standard Deviation

base line table

Unique cookies to view page per day:	40000

Unique cookies to click “Start free trial” per day:	3200
Enrollments per day:	660
Click-through-probability on “Start free trial”:	0.08
Probability of enrolling, given click:	0.20625
Probability of payment, given enroll:	0.53
Probability of payment, given click:	0.1093125

For 5000 page view,

$$\text{number of clicks} = 5000 \times 0.08 = 400$$

$$\text{number of enrollment} = 5000 \times 0.08 \times 0.20625 = 82.5$$

We know that

$$\text{stand deviation} = \sqrt{\frac{p \times (1-p)}{n}}$$

So

$$\text{analytical stand deviation for gross conversion} = \sqrt{\frac{0.20625 \times (1-0.20625)}{400}} = 0.0202$$

$$\text{analytical stand deviation for retention} = \sqrt{\frac{0.53 \times (1-0.53)}{82.5}} = 0.0549$$

$$\text{analytical stand deviation for net conversion} = \sqrt{\frac{0.1093125 \times (1-0.1093125)}{400}} = 0.0156$$

For gross conversion and net conversion, their denominators are number of clicks, which is also unit of diversion, so their analytic variance are likely to match their empirical variance. For retention, its denominator is number of enrollments, which is not unit of diversion in our experiment, so its empirical variance would be much higher than analytic variance.

Sizing

Number of Samples vs. Power

We want gross conversion significantly decrease AND net conversion does not significantly decrease. Bonferroni correction is suitable when use `OR` and does not suitable when use `AND`, so we would not use Bonferroni correction.

$\alpha = 5\%, \beta = 20\%$

- `Gross conversion` (base conversion rate = 20.625%, $d_{min} = 1\%$)
- `Retention` (base conversion rate = 53%, $d_{min} = 1\%$)
- `Net conversion` (base conversion rate = 10.93125%, $d_{min} = 0.75\%$)

From this calculator, we get samples

- `Gross conversion`: 25835 clicks for each group
- `Retention`: 39115 enrollments for each group

- `Net conversion`: 27413 clicks for each group

So we need pageviews

- `Gross conversion`: $25835 \times 40000 \div 3200 = 322937.5$ pageviews for each group
- `Retention`: $39115 \times 40000 \div 660 = 2370606$ pageviews for each group
- `Net conversion`: $27413 \times 40000 \div 3200 = 342662.5$ pageviews for each group

Duration vs. Exposure

First trial (Use `Gross conversion`, `Retention` and `Net conversion` as evaluation metrics):

- Number of pageviews: $2370606 \times 2 = 4741212$
- Fraction: 1.0
- $days = 4741212 \div 1.0 \div 40000 = 119$

Experiment of 119 days is too long.

Second trial (Only use `Gross conversion` and `Net conversion` as evaluation metrics)

- Number of pageviews: $342662.5 \times 2 = 685325$
- Fraction: 1.0
- $days = 685325 \div 1.0 \div 4000 = 18$

Experiment of 18 days is short enough, so we would choose `Gross conversion` and `Net conversion` as evaluation metrics, and we abandon `Net conversion` as evaluation metrics since this would require too long duration (119 days).

In this experiment, when students want to enroll we ask how much time each week they have for the course they want to enroll and recommend students not to enroll if they could not investigate more than 5 hours' time. This does not harm to students, students who don't have enough time could choose to access course materials, watch videos and do quizzes, when they think they need to enroll, they could enroll at any time. So this experiment does not harm to students. Also we don't ask personal information in this experiment, so there are not privacy problems. This experiment does not significantly increase burden of Udacity website, does not change database much, so the website would not be harmed by this experiment. Since this experiment does not harm to students and website, we could use 100% traffic, 18 days is long, if we decrease fraction of traffic, we would need more times to run this experiment, so I prefer divert 100% fraction of Udacity's traffic to this experiment.

Experiment Analysis

Sanity Checks

- Control group: #pageviews = 345543, #clicks = 28378,
- Experiment group: #pageviews = 344660, #clicks = 28325,

1. Number of cookies

$$SD = \sqrt{\frac{0.5 \times 0.5}{345543 + 344660}} = 0.000602$$
$$\text{margin of error} = SD \times 1.96 = 0.00118$$
$$\text{confidential interval} = (0.5 - 0.00118, 0.5 + 0.00118) = (0.49882, 0.50118)$$
$$\hat{p} = \frac{345543}{345543 + 344660} = 0.50064 \in (0.49882, 0.50118)$$

So number of cookies pass sanity check.

2. Number of clicks on "Start free trial"

$$SD = \sqrt{\frac{0.5 \times 0.5}{28378 + 28325}} = 0.0021$$
$$\text{margin of error} = SD \times 1.96 = 0.004116$$
$$\text{confidential interval} = (0.5 - 0.004116, 0.5 + 0.004116) = (0.495884, 0.504116)$$
$$\hat{p} = \frac{28378}{28378 + 28325} = 0.500467 \in (0.495884, 0.504116)$$

So number of clicks on "Start free trial" pass sanity check.

3. Click-through-probability

Since click-through-probability = $\frac{\text{Number of clicks on "Start free trial"}}{\text{Number of cookies}}$, and both Number of clicks on "Start free trial" and Number of cookies pass sanity check, click-through-probability would also pass sanity check.

Result Analysis

- Control group: #clicks=17293, #enrollment=3785, #payments=2033
- Experiment group: #clicks=17260, #enrollment=3423, #payments=1945

Effect Size Tests

1. Gross conversion

$$\hat{p} = \frac{3423 + 3785}{17293 + 17260} = 0.2086$$
$$SE = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{N_1} + \frac{1}{N_2})} = \sqrt{0.2086(1-0.2086)(\frac{1}{17293} + \frac{1}{17260})} = 0.004372$$
$$\hat{d} = \frac{3423}{17260} - \frac{3785}{17293} = -0.02055$$
$$m = SE * 1.96 = 0.00857$$
$$\text{confidential interval} = (\hat{d} - m, \hat{d} + m) = (-0.02055 - 0.00857, -0.02055 + 0.00857) = (-0.0291, -0.0120)$$

Since $(-0.0291, -0.0120) \subset (-\infty, 0), (-0.0291, -0.0120) \subset (-\infty, -0.01)$, gross conversion in experiment group is both statistically significant and practically significant less than gross conversion in control group.

2. Net conversion

$$\hat{p} = \frac{1945 + 2033}{17293 + 17260} = 0.1151$$

$$SE = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{N_1} + \frac{1}{N_2})} = \sqrt{0.1151(1-0.1151)(\frac{1}{17293} + \frac{1}{17260})} = 0.003434$$

$$\hat{d} = \frac{1945}{17260} - \frac{2033}{17293} = -0.00487$$

$$m = SE * 1.96 = 0.00673$$

$$\text{confidential interval} = (\hat{d} - m, \hat{d} + m) = (-0.00487 - 0.00673, -0.00487 + 0.00673) = (-0.0116, 0.0019)$$

Since $0 \in (-0.0116, 0.0019)$, net conversion in experiment group is neither statistically significant nor practically significant different from net conversion in control group.

Sign Tests

Use this calculator

1. Gross conversion

- Number of “successes” you observed = 4
- Number of trials or experiments = 23
- Probability = 0.5

We get two-tailed p value 0.0026, $0.0026 < 0.05$, also $4 < 11$, so gross conversion in experiment group is significantly less than gross conversion in control group.

2. Net conversion

- Number of “successes” you observed = 10
- Number of trials or experiments = 23
- Probability = 0.5

We get two-tailed p value 0.6776 , since $0.6776 > 0.05$, so net conversion in both group are not significantly different.

Summary

We want gross conversion significantly decrease AND net conversion does not significantly decrease. Bonferroni correction is suitable when use `OR` and does not suitable when use `AND`, so we do not use Bonferroni correction in this experiment. For both effect size hypothesis tests and the sign tests, gross conversion in experiment group is significantly less than gross conversion in control group, net conversion in both groups has not significantly differ.

Recommendation

This experiment significantly decrease gross conversion but do not significantly decrease net conversion. This means that this new change would significantly reduce enrollment that does not pay, but does not significantly reduce enrollment that pay. In our experiment net conversion in both groups has not significantly differ, so we are not sure that this experiment would not decrease net conversion, we could not launch the experiment if we are not sure about that, thus without further experiments we could not decide to launch the experiment.

Follow-Up Experiment

I think that if we provide 50 dollars of discount in first month if the student arrange a one-to-one discussion with coach during free trial period, we will reduce early cancel of enrollments. When student experiment the support of coach they would feel supported and have more confidence that they will accomplish the course, thus they would be more probable to pay. We want to use user-ids as unit of diversion because this more stable than cookies. We define cancel rate as number of users who click “Start free trial”and cancel enrollment in 14-days divided by number of users who click “Start free trial” button. We use cancel rate as evaluation because this is about probability of early cancel, and we want to know that weather probability of early cancel decrease in our experiment, so this would be good metrics.