AB Test Design and Analysis

Experiment Overview: Ads conversion

We designed an ads asking people who viewed the ads to become a member. To simplify, we only show the click option of single ride (casual) and membership. After clicking the ads, the landing page shows the following text for experiment and control group:

Welcome to Cyclistic
$5 single ride

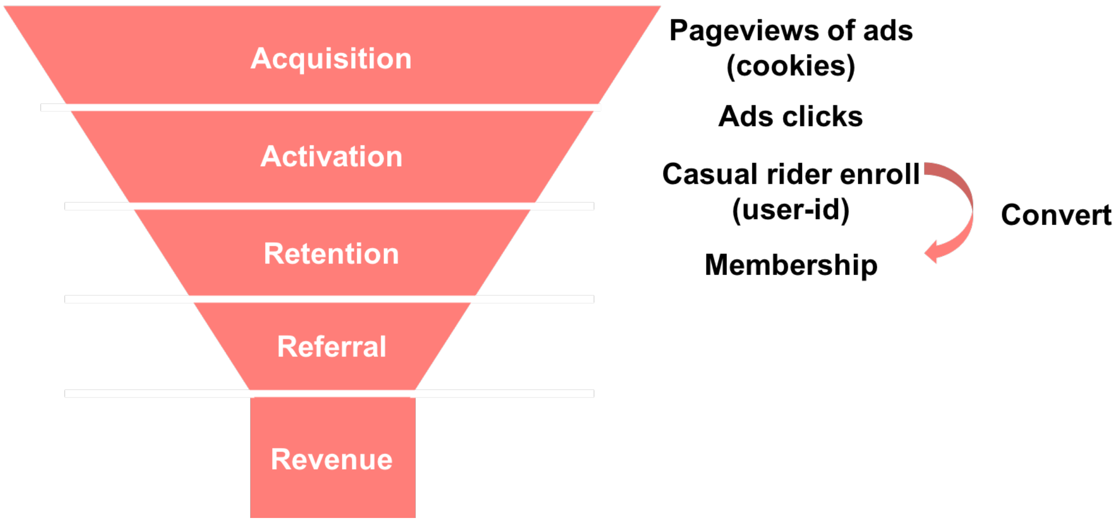
$99 annual membership

Welcome to Cyclistic
$5 single ride

$99 annual membership:
Go dinner after work
Go shopping
Go with kids

We can use cookies to track which users are directed from the target website. Once user made member for single/ full-day passes, they create account and get user-id. Base on previous analysis, most conversion happen within 14 days after purchase of single/ full-day passes; after 14 days the conversion rate drops. So, we define some metrics as:

- **Number of cookies:** That is, number of unique cookies to view the website page showing ads. (d_{min}=3000)
- **Number of clicks:** That is, number of unique cookies to click the ads button (d_{min}=240)
- **Click-through-probability:** That is, number of unique cookies to click the ads button divided by number of unique cookies to view the page with ads. (d_{min}=0.01)
- **Number of casual user-ids:** That is, number of users who created account and paid for single ride. (d_{min}=50)
- **Gross conversion:** That is, number of casual user-ids divided by number of unique cookies to click the ads page. (d_{min}= 0.01)
- **Net conversion:** That is, number of member user-ids divided by number of unique cookies to click the ads page. (d_{min}= 0.01)
- **Retention:** That is, number of casual user-ids paid for members within 14 days divided by number of casual user-ids. (d_{min}=0.01)



Metric Choice

Evaluation metrics

Retention

The hypothesis is by adding ways of riding, users are more likely to consider pay for membership. So, the number of casual user-ids to become members later divided by total number of casual user-id is expected to increase if hypothesis tested true.

Net conversion

As the ultimate goal is to have more members, it is also possible that people will start as member if they know how to enjoy rides.

Invariant metrics

Number of cookies

The experiment only affect behavior after ads button clicks, so number of cookies before that will not changes.

Number of clicks

Same as above, the clicks happen upstream of expected changes.

click-through-probability (CTP)

CTP is calculated from invariant metrics above.

Number of casual user-id

Same as above, this happen upstream of expected changes.

Gross conversion

As the test is designed to increase membership purchase, gross conversion of casual user from ads click is not expected to change.

Measuring Variability

Unique cookies to view page with ads per day:	40000
Unique cookies to click ads per day:	3200
Casual user per day:	660
Click-through-probability on ads:	0.08
Probability of casual, given click:	0.20625
Probability of member, given casual:	0.53
Probability of member, given click	0.1093125

For each metric you selected as an evaluation metric, estimate its standard deviation analytically. Do you expect the analytic estimates to be accurate? That is, for which metrics, if any, would you want to collect an empirical estimate of the variability if you had time?

Retention
N = 660
p = Probability of member, given casual = 0.1093125

$\mathrm{sd} = \sqrt{\mathrm{p} * \frac{(1-\mathrm{p})}{\mathrm{N}}} = \sqrt{0.53 * \frac{(1-0.53)}{3200}} = 0.0194$

Net conversion
p = Probability of payment, given click = 0.1093125
sd = 0.0055

Both retention and net conversion are both unit of analysis and unit of diversion, thus the estimate is comparable to empirical estimation of variance.

Sizing

Choosing Number of Samples given Power

Using the analytic estimates of variance, how many pageviews **total** (across both groups) would you need to collect to adequately power the experiment? Use an alpha of 0.05 and a beta of 0.2. Make sure you have enough power for **each** metric.
Using tools by https://www.evanmiller.org/ab-testing/sample-size.html

Net conversion
Baseline conversion rate: 10.93125%
Minimum Detectable Effect: 0.75 %
Sample size: 27413
Number of cookies (pageview) = number of clicks / CTP * 2 groups = 27413/0.08 *2 = 685325

Retention
Baseline conversion rate: 53%
Minimum Detectable Effect: 1 %
Sample size: 39115
Number of cookies = 39115* 1/(Probability of casual, given click) * 1/CTP * 2 groups = 758594
Thus, we need at least 758594 cookies to achieve enough power.

Choosing Duration vs. Exposure

What percentage of the website's traffic would you divert to this experiment (assuming there were no other experiments you wanted to run simultaneously)? Is the change risky enough that you wouldn't want to run on all traffic?
Given the percentage you chose, how long would the experiment take to run, using the analytic estimates of variance? If the answer is longer than a few weeks, then this is unreasonably long, and you should reconsider an earlier decision.

Although it's common to start with a small percent (1%, 10%) of traffic and expand to a bigger percentage in AB test [1], we might start with 50% or greater because:
1/ the experiment is not risky, it is unlikely to annoy users or cause any negative effect
2/ there are not technical difficulties that might hurt user experience
3/ with more traffic daily can reduce time of running, and speed up the test cycle iteration
Days required for test = 758594 * 50% / 40000 = 10

Analysis

Sanity Checks

Start by checking whether your invariant metrics are equivalent between the two groups. If the invariant metric is a simple count that should be randomly split between the 2 groups, you can use a binomial test as demonstrated in Lesson 5. Otherwise, you will need to construct a confidence interval for a difference in proportions using a similar strategy as in Lesson 1, then check whether the difference between group values falls within that confidence level.
If your sanity checks fail, look at the day by day data and see if you can offer any insight into what is causing the problem.

95% confidence interval, z score = 1.96
Probability of a cookie in the control group p = 0.5

Number of cookies
Total number of cookies in control = N(cookies_ctl) = 345543
Total number of cookies in experiment = N(cookies_exp) = 344660
Total number of cookies = N(cookies_total) = 690203
$\text{sd} = \sqrt{\text{p} * \frac{(1-\text{p})}{\text{N}}} = \sqrt{0.5 * \frac{(1-0.5)}{690203}} = 0.0006$
margin of error m = z score * sd = 0.0006 * 1.96 = 0.0012
lower bound = p – m = 0.4988
upper bound = p + m = 0.5012
Observed fraction of cookies in control groups = N(cookies_ctl)/N(cookies_total) = 345543/690203= 0.5006
The value 0.5006 is in 95% confidence interval [lower= 0.4988, upper= 0.5012], thus number of cookies passes sanity check.

(Apply the same formula above in R script)

Number of clicks
Total number of clicks in control = 28378
Total number of clicks in experiment = 28325
Total number of cookies = 56703

sd = 0.0021
margin of error m = 0.0041
lower bound = 0.4959
upper bound = 0.5041
Observed fraction of clicks in control = 0.5005
The value 0.5005 is in 95% confidence interval, thus number of clicks passes sanity check.

CTP
CTP_ctl = N(clicks_ctl)/N(cookies_ctl) = 28378/345543 = 0.0821
CTP_exp = N(clicks_exp)/N(cookies_exp) = 28325/346660 = 0.0822
p = (N(clicks_ctl) + N(clicks_exp)) / (N(cookies_ctl) + N(cookies_exp)) = (28378+28325)/(345543+344660) = 0.0822

$\mathrm{sd} = \sqrt{\mathrm{p} * (1-\mathrm{p}) * (\frac{1}{\mathrm{N(cookies\_ctl)}} + \frac{1}{\mathrm{N}(cookies\_exp)})}$ = 0.00066

margin of error m = 0.0013
lower bound = -0.0013
upper bound = 0.0013
Observed difference of CTP = CTP_exp – CTP_ctl = 0.001
The value 0.001 is in 95% confidence interval, thus CTP passes sanity check.

Check for Practical and Statistical Significance

Next, for your evaluation metrics, calculate a confidence interval for the difference between the experiment and control groups, and check whether each metric is statistically and/or practically significance. A metric is statistically significant if the confidence interval does not include 0 (that is, you can be confident there was a change), and it is practically significant if the confidence interval does not include the practical significance boundary (that is, you can be confident there is a change that matters to the business.)
If you have chosen multiple evaluation metrics, you will need to decide whether to use the Bonferroni correction. When deciding, keep in mind the results you are looking for in order to launch the experiment. Will the fact that you have multiple metrics make those results more likely to occur by chance than the alpha level of 0.05?

Net conversion (Sat, Oct 11 to Sun, Nov 2)
p = (N(pay_ctl) + N(pay_exp)) / (N(clicks_ctl) + N(clicks_exp)) = (2033+1945)/(17293+17260) = 0.1151

$\mathrm{sd} = \sqrt{\mathrm{p} * (1-\mathrm{p}) * (\frac{1}{\mathrm{N(clicks\_ctl)}} + \frac{1}{\mathrm{N(clicks\_exp)}})}$ = 0.0034

margin of error m = 0.0067
p(pay_ctl) = N(pay_ctl)/N(clicks_ctl) = 0.1176
p(pay_exp) = N(pay_exp)/N(clicks_exp) = 0.1127
p(diff) = p(pay_exp) - p(pay_ctl) = -0.0049
lower bound = -0.0116
upper bound = 0.0019
dmin = 0.0075

As the confidence interval include 0 but include dmin boundary, the decrease in gross conversion is not statistically nor practically significant. We can accept the null hypothesis that there is no difference between the 2 groups.

Retention (Sat, Oct 11 to Sun, Nov 2)
p = (N(member_ctl) + N(member_exp)) (N(casual_ctl) + N(casual_exp)) =
(2033+1945)/(3785+3423) = 0.5519

$$\mathrm{sd} = \sqrt{\mathrm{p} * (1 - \mathrm{p}) * (\frac{1}{\mathrm{N(casual\_ctl)}} + \frac{1}{\mathrm{N(casual\_exp)}})} = 0.0117$$

margin of error m = 0.0229
p(member_ctl) = N(member_ctl)/N(casual_ctl) = 0.5371
p(member_exp) = N(member_exp)/N(casual _exp) = 0.5682
p(diff) = p(member_exp) - p(member_ctl) = 0.0311
lower bound = 0.0082
upper bound = 0.054
dmin = 0.01
As the confidence interval does not include 0 but include dmin boundary, the increase in retention is statistically but not practically significant. We can accept the null hypothesis that there is no difference between the 2 groups.

Run Sign Tests

For each evaluation metric, do a sign test using the day-by-day breakdown. If the sign test does not agree with the confidence interval for the difference, see if you can figure out why.
Using tools by https://www.graphpad.com/quickcalcs/binomial1.cfm

Net conversion
Number of "successes": 13
Number of trials (or subjects) per experiment: 23
Sign test. If the probability of "success" in each trial or subject is 0.500, then:
The two-tail P value is 0.6776
Retention
Number of "successes": 13
Number of trials (or subjects) per experiment: 23
Sign test. If the probability of "success" in each trial or subject is 0.500, then:
The two-tail P value is 0.6776
The decrease in gross conversion is statistically significant while the change in retention is not.

Make a Recommendation

Finally, make a recommendation. Would you launch this experiment, not launch it, dig deeper, run a follow-up experiment, or is it a judgment call? If you would dig deeper, explain what area you would investigate. If you would run follow-up experiments, briefly describe that experiment. If it is a judgment call, explain what factors would be relevant to the decision.

		state of nature	
		H0 true	H1 true
action	accept H0	True negative	False negative (fail to reject, type II error)
	reject H0	False positive (fail to reject, type I error)	True positive

Bonferroni correction was not applied as there were not many evaluation metrics in this case. Multiple testing correction reduce type I errors at the cost of type II errors. In the case here, retention is a more important metrics as it is more related to the downstream of the funnel. If we falsely accepted H0 that there is no change in net conversion, we will risk the reality that it might decrease, the ads will hurt the business. So, the correction is not practical in net conversion.

The ads do not help to increase membership.

For net conversion, although we accepted H0 that there is no change in experiment and control groups, the 95% confidence interval include practical significant boundary on the negative side, suggesting that net conversion might decrease and hurt business. We should further check 99% intervals.
For retention, although we see possible increase, it is not significant either.

Follow-Up Experiment: How to further increase retention

We want to study the behavior and features of casual users who ride >=3 time without converting to members. We start with the assumption that users do not convert because of financial reason that they feel member cost is too high. We want to offer users selected financial reason a 10% discount for the first member year.

Survey design
We want to first get users' age, education level, family education level, job income, etc. in addition to the survey specific for this retention goal:

financial	Membership fee is too high
	I ride with family/kids, thus need to purchase multiple
usage	Not enough stations
	I don't ride enough

Hypothesis
H0: no change in retention
H1: increased retention rate
Users received discount are more likely to complete first member.

Evaluation metrics
Retention: number of casual user-ids paid for members within 14 days divided by number of casual user-ids.
(other upstream/ downstream metrics should be monitored as well, as retention in the second year might decrease when fee returned to normal)

References

1/ Hacking Growth: How Today's Fastest-Growing Companies Drive Breakout Success; Morgan Brown, Sean Ellis;
https://books.google.com/books/about/Hacking_Growth.html?id=LcS_DAAAQBAJ