Final Project for Unit 7 (A/B Testing)

Eric Perbos-Brinck
29 November 2016

- **Number of cookies:** That is, number of unique cookies to view the course overview page. (dmin=3000)
- **Number of user-ids:** That is, number of users who enroll in the free trial. (dmin=50)
- **Number of clicks:** That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger). (dmin=240)
- **Click-through-probability**: That is, 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)
- **Gross conversion:** That is, 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)
- **Retention:** That is, 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)
- **Net conversion:** That is, 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)

Experiment Design

Metric Choice

List which metrics you will use as invariant metrics and evaluation metrics here. (These should be the same metrics you chose in the "Choosing Invariant Metrics" and "Choosing Evaluation Metrics" quizzes.)

For each metric, explain both why you did or did not use it as an invariant metric and why you did or did not use it as an evaluation metric. Also, state what results you will look for in your evaluation metrics in order to launch the experiment.

ANSWER

A. Invariant Metrics: they should not change across control and experimental groups as the experiment will not have a direct effect on them.

1. Cookies (used): number of unique cookies to view the course overview page
This is a population sizing metric used to split the control and experiement groups evenly.

2. Clicks used): number of unique cookies to click the "Start free trial" button
This indicates the number of people choosing to start the trial, it should be identical as the experiment starts after.

3. Click-through-probability (used): number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page.
This should remain constant and is normalized so a good invariant metric to use.

4. Number of user-ids (not used): number of users who enroll in the free trial.
This metric isn't normalized -like a gross conversion rate is- so it is not used.

B. Evaluation metrics: they should change over the experiment with differences observed between control and experiment groups.
If the hypothesis is true, we should see a reduction in rate of enrollment, as un-prepared students do not enroll, without a decrease of payments : gross conversion should decrease while net conversion stays the same.

5. Gross conversion (used): 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.
The number of enrollments -and the resulting gross conversion- can be affected by the experiment and will be used as an evaluation metric.

6. Retention (used): 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.
It can be affected by the experiment as a lower number of user-ids (due to un-prepared students declining) may complete checkout thus will be used as evaluation metric.

7. Net Conversion (used): 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.
This is a critical metric affecting the bottom line (payments -> cash-flows) and must be used as an evaluation metric.

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

For each of your evaluation metrics, indicate whether you think the analytic estimate would be comparable to the the empirical variability, or whether you expect them to be different (in which case it might be worth doing an empirical estimate if there is time). Briefly give your reasoning in each case.

ANSWER

1. Gross Conversion
SD = sqrt [p * (1 - p) / N] where
p = 0.20625 (probability of enrolling, given click)
p1 = 0.08 (probability of click-through)
N=5000 * p1 = 5000 * .08 = 400 (number of click throughs for 5000 page views)
Gross conversion SD =0.0202

2. Retention
SD = sqrt [p * (1 - p) / N] where
p= 0.53 (probability of payment given enrollment)
p1= 0.08 (probability of click-through)
p2= 0.20625 (probability of enrolling, given click)
N= 5000 * p1 * p2 = 5000 * 0.08 * 0.20625 = 82.5 (number of payments for 5000 page views)
Retention SD =0.0549

3. Net Conversion
SD = sqrt [p * (1 - p) / N] where
p= 0.10931 (probability of payment, given click)
p1= 0.08 (probability of click-through)
N= 5000 * p1 = 5000 * .08 = 400 (number of click throughs for 5000 page views)
Net conversion SD =0.0156

Gross conversion and Net conversion metrics use the same unit of analysis and diversion (ie. cookies) so their analytical standard deviation should match thei empirical one.
While the Retention metric use different units of diversion (cookies) and analysis (user-ids) so its analytical estimate may differ from its empirical one.

Sizing

Number of Samples vs. Power

Indicate whether you will use the Bonferroni correction during your analysis phase, and give the number of pageviews you will need to power you experiment appropriately. (These should be the answers from the "Calculating Number of Pageviews" quiz.)

Demo in Lesson 1.22 "Quiz: Calculating Number of Page Views"
& Bonferonni correction in Lesson 5.14 "Multiple Metrics"

ANSWER

The Bonferroni Correction was not used during the analysis phase.
The initial number of page views required (largest sample size) to conduct the experiment was: 4,741,212

Using the online sample size calculator from "Evan's Awesome A/B Tools":
http://www.evanmiller.org/ab-testing/sample-size.html

1. Gross Conversion
Baseline conversion rate = 0.20625 (Final Project Baseline Values)
Minimum detectable effect = 0.01
Sample size (clicks) = 25835 (Sample Size Calculator)
Click-through-probability = 0.08
Unique Group Sample size (pageviews with unique cookie) = 25835 / 0.08 = 322937.5
Total Sample size (pv's half for experiment, half for control) = 2 * 322937.5 = 645,875

2. Retention
Baseline conversion rate = 0.53
Minimum detectable effect = 0.01
Sample size (enrollments) = 39115 (Sample Size Calculator)
Click-through-probability = 0.08
Probability of enrolling, given click = 0.2063
Unique Group Sample size = 39115 / (0.08 * 0.2063) = 2370606
Total Sample size = 2 * 2370606 = 4,741,212

3. Net Conversion
Baseline conversion rate = 0.1093125
Minimum detectable effect = 0.0075
Sample size (clicks) = 27413 (Sample Size Calculator)
Unique Group Sample size = 27413 / 0.08 = 342662.5
Total Sample size = 2 * 342662.5 = 685,325

Duration vs. Exposure

Indicate what fraction of traffic you would divert to this experiment and, given this, how many days you would need to run the experiment. (These should be the answers from the "Choosing Duration and Exposure" quiz.)

Give your reasoning for the fraction you chose to divert. How risky do you think this experiment would be for Udacity?

ANSWER

Because of the lack of sensitive data being collected and the low risk on Audacity operations, I chose to divert all the traffic for 100% exposure.
- Low risk: the experiment is a simple message of caution to the potential recruits and should not have a dramatic impact on the key metric Net Conversion and the bottom line (payments).
- Lack of sensitive data: no confidential information is recorded in the experiment.

But the initial Sample Size of 4,741,212 involved a very long-running experiment (up to 119 days or 4 months) and Audacity doesn't want to spend that long.

So I chose the second largest sample size from Net Conversion (685,325) which required 18 days to run. (685,325 / (40,000 cookies per day * 100%)) = 17.13 days rounded-up to 18 full days.

Experiment Analysis

Sanity Checks

For each of your invariant metrics, give the 95% confidence interval for the value you expect to observe, the actual observed value, and whether the metric passes your sanity check. (These should be the answers from the "Sanity Checks" quiz.)

For any sanity check that did not pass, explain your best guess as to what went wrong based on the day-by-day data. **Do not proceed to the rest of the analysis unless all sanity checks pass.**

Demo in Lesson 1.25: "Quiz: Calculating Results"

ANSWER

Note: calculations for sanity checks should come from the control group only
https://discussions.udacity.com/t/sanity-check-on-click-through-probability/39676

1. Number of Cookies
Number of cookies (control) Nc = 345543
Number of cookies (experiment) Ne = 344660
p (probability event is assigned to control group) = 0.5
Standard error (binomial) SE = sqrt [p * (1-p) / (Ne + Nc)] =
sqrt [0.5*0.5/(344660+345543)] = 0.000602

Margin of error, m= SE * Z(alpha = 0.05, two-tail) = 0.0006 * 1.96 = 0.0012
Confidence interval lower limit, CI(low) = 0.5 - 0.00118 = 0.4988
Confidence interval upper limit, CI(up) = 0.5 + 0.00118 = 0.5012
Observed fraction, f = Nc / (Ne + Nc) = 0.5006
CI(low) < f < CI(up) : OK
Conclusion: The observed fraction falls within the confidence interval, and the sanity check passes.

2. Number of Clicks
Number of clicks Nc = 28378
Number of clicks Ne = 28325
p = 0.5
SE = 0.002100
m = 0.0041
CI(low) = 0.5 - 0.0041 = 0.4959
CI(up) = 0.5 + 0.0041 = 0.5041
Observed fraction, f = Nc / (Ne + Nc) = 0.5005
CI(low) < f < CI(up) : OK
Conclusion: The observed fraction falls within the confidence interval, and the sanity check passes.

3. Click-through Probability
Nc = 345543
Xc = 28325
CTRc = 0.0821
CTRe = 0.0822
CTRc SE = 0.0005
m = 0.0009
CI(low) = 0.0812
CI(up) = 0.0830
f = 0.0822
CI(low) < f < CI(up) : OK
Conclusion: The observed fraction falls within the confidence interval, and the sanity check passes.

Result Analysis

Effect Size Tests

For each of your evaluation metrics, give a 95% confidence interval around the difference between the experiment and control groups. Indicate whether each metric is statistically and practically significant. (These should be the answers from the "Effect Size Tests" quiz.)

Demo in Lesson 1.26: "Quiz: Confidence Interval Cases" Launch/ No_Launch/ More_tests
& Lesson 5.14: "Multiple Metrics"

ANSWER

Remember: we dropped the Retention metric earlier due to extensive duration.
Also the actual duration of the experiment was 23 days as per "Final Project Results" spreadsheet.

1. Gross Conversion
Number of clicks (control) Nc = 17293
Number of enrollments (control) Xc = 3785
Number of clicks (experiment) Ne = 17260
Number of enrollments (experiment) Xe = 3423
p (pooled probability of enrollment) = (Xc + Xe) / (Nc + Ne) = 0.20861
Pooled Standard error (binomial) SE = sqrt [p * (1 - p) / (Ne + Nc)] =
sqrt [0.20861 * (1 - 0.20861) / (17293 + 17260)] = 0.004372
Difference estimate, d = (Xe / Ne) – (Xc / Nc) = (3423 / 17260) – (3785 / 17293) = -0.02055
Margin of error, m= SE * Z(alpha = 0.05, two-tail) = 0.004372 * 1.96 = 0.00857
Confidence interval lower limit, CI(low) = -0.02055 - 0.008568 = -0.0291
Confidence interval upper limit, CI(up) = -0.02055 + 0.008568 = -0.0119

- Statistical significance: Null/Zero IS NOT within the Confidence Interval limits [-0.0291, -0.0119];
thus the result has statistical significance.

- Practical significance: the CI limits [-0.0291, -0.0119] both exceeds the minimum level dmin=0.01 in practical significant bounderies, negative on in this case, [-0.01, +0.01];
so there is zero overlap.
thus the result has practical significance.

Conclusion: Gross Conversion is smaller in the experimental group than in the control one.

2. Net Conversion
Nc = 17293
Xc = 2033
Ne = 17260
Xe = 1945
p = (Xc + Xe) / (Nc + Ne) = 0.11513
Pooled SE = sqrt [0.11513 * (1-0.11513) * (17293 + 17260)) = 0.00343
d = (1945 / 17260) – (2033 / 17293) = -0.00487
m = 0.00343 * 1.96 = 0.00673
CI(low) = -0.00487 - 0.00673 = -0.0116
CI(up) = -0.00487 + 0.00673 = 0.0019

- Statistical significance: Null/Zero IS within the Confidence Interval limits [-0.0116, +0.0019];
thus the result does NOT have statistical significance.

- Practical significance: the CI limits [-0.0116, +0.0019] does OVERLAP both zero AND the minimum level dmin=0.0075 in negative value [-0.0075 , +0.0075];
thus the result does NOT have practical significance.

Conclusion: Net Conversion is NOT significantly different in the experimental group than in the control one but may require additional tests to act upon launch vs no_launch.

Sign Tests

For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant. (These should be the answers from the "Sign Tests" quiz.)

Demo in Lesson 5.9: "Single metric example"

ANSWER

Online calculator used for Sign and Binomial Test: http://graphpad.com/quickcalcs/binomial1.cfm
Two-tail P value: probability of getting the observed number of success/failure by chance alone

Success day: when the probability of click-through-rate for Gross Conversion or payment-rate for Net Conversion is higher in the Experiment group than in the Control group.

1. Gross Conversion
Number of successes observed, Ns = 4
Number of trials or experiments, Nt = 23
Probability for a success day under null hypothesis, p0 = 0.5
Probability value (two-tail) from sign test: P = 0.0026
P = 0.0026 < alpha = 0.05
Gross Conversion is statistically significant by sign test.

2) Net conversion
Ns = 10
Nt = 23
p0 = 0.5
Two-tail P = 0.6776
P = 0.6776 > alpha = 0.05
Net Conversion is NOT statistically significant by sign test.

Summary

State whether you used the Bonferroni correction, and explain why or why not. If there are any discrepancies between the effect size hypothesis tests and the sign tests, describe the discrepancy and why you think it arose.

"When to use Bonferroni correction" thread in DAND P7_Design_A/B_Test forum
https://discussions.udacity.com/t/when-to-use-bonferroni-correction/37713
Controlling procedures, Wikipedia
https://en.wikipedia.org/wiki/Family-wise_error_rate#Controlling_procedures

ANSWER

The Bonferonni correction was not used because it's a conservative method to reduce the risk of Type I errors (false positive) at the expense of Type II errors (false negative).
It is useful when evaluating multiple metrics in a "OR" condition, that is only one significant metric is needed.

Since we are using two metrics Gross Conversion & Net Conversion in a "AND" condition, this is being conservative in the same way.

The Effect-size and Sign tests both showed Gross Conversion to be statistically significant, while Net Conversion is not.
Therefore there is no discrepancy to explain.

Recommendation

Make a recommendation and briefly describe your reasoning.

ANSWER

The experiment was providing additional information to potential students regarding the expected "5 hours per week" workload before they chose to enroll in a program at Audacity.

The hypothesis was that this may lower the total number of new students registering by elimination of un-prepared students who drop-out before the first payement is due 14 days later, thus a lower Gross Conversion.
Without impacting negatively, as in reducing, the core number of students who remained in the program after payement due, thus maintain Net Conversion.

The statistical significance of experiment showed that it was successful in reducing Gross Conversion while maintaining Net Conversion.
However the confidence interval for Net Conversion [-0.0116, +0.0019] overlaped the lower boundary of the interval for practical significance -0.0075. Thus we do not have 95% confidence that the change will NOT have a negative impact on Net Conversion which would translate in lower revenues (ie. fewer paying students).

My recommendation: do NOT launch the change until an additional test with greater power on Net Conversion is conducted and Audacity is confident that its bottom line will not suffer.

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.

ANSWER

As a direct follow-up experiment, I would consider re-running the same experiment but with a reduced "start-free-trial" period from 14 days to 10 days or even 7 days (To be discussed with management and product owners).

My reasoning is "Make hay while the sun shines":

1. We saw that the pop-up window "How much time can you devote to the course ?" was efficient in reducing the Gross Conversion, thus generating a more qualified audience/base to convert to full-time paying students.

2. What is unclear at this stage is whether this pop-up window has any impact, positive or negative, on the Net Conversion. Maybe it doesn't because it's just not correlated or more actually not a causal trigger.

3. When a student starts the free-trial, he/she has entered his credit card details -big show of confidence- so **the motivation to study is real and immediate**: it shouldn't take more than a week for the student to invest 5+ hours into the program and make up his/her mind.
On the contrary, allowing for a long period of grace may be counter-productive by letting the student drift away.
So let's see if shortening the trial-period will help to "close the deal".

Experiment details:

A. Null hypothesis, H_0: the change in Gross Conversion and Net Conversion (d) in the experimental condition and control conditions is zero (alpha=0.05).

B. Alternate hypothesis, H_A: the change in Gross Conversion and Net Conversion (d) in the experimental condition and control conditions is not equal to zero (alpha=0.05).
The differences are to be tested for business/practical significance.

C. Invariant Metrics: same as before (Cookies, Clicks, CTR)

D. Evaluation Metrics: same as before (Gross Conversion, Net Conversion)

E. Unit of diversion: same as before (Cookies)

F. Launch Criteria: to meet the launch criteria, the statistical testing must indicate that the null hypothesis should be rejected (95% confidence).
Furthermore, the 95% confidence interval for d must be greater than the minimum (d_{min}) requirement for business/practical significance.