Final Project for Unit 7 (A/B Testing)

Perry Radau
October 3, 2016

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:
The invariant metrics chosen were
1) cookies (“unique cookies to view page per day”) and
2) clicks (“unique cookies to click ‘start free trial’ per day”).
The evaluation metrics chosen were
1) gross conversion (“probability of enrolling, given click”) and
2) net conversion (“probability of payment, given click”).

For invariant metrics, the reason for choosing cookies and clicks (as defined above) is that the former gives an approximation of the number of people viewing the page for the day; the latter gives the number of people choosing to start the trial. In each case we expect this to be identical in the control and experimental conditions because the experiment modifies the user experience after the trial is chosen, and we want an equal sample of users that view and click the button. *The number of unique user ids is not ideal for the experiment because it is not normalized (i.e. gross conversion is a better metric), however it could potentially be an evaluation metric as it does include the students in the free trial and those who make payment.*
For evaluation metrics, the reason for choosing gross conversion is that we wish to determine if the experimental condition affects the rate of enrollment and gross conversion measures this. The reason for choosing net conversion is that we wish to determine if the experimental condition also affects the rate that individuals will not only enroll but also make payment.
The requirement for launch of the change specified by the experimental condition (i.e. the message to the students indicating that 5+ hours /week available time is usually required) is that this change will reduce the gross conversions, **and** the net conversions will be equal or greater, as compared to the control condition. This means that in the experiment it is hypothesized that fewer people will be enrolled in the free trial (and eliminating some of the students who would otherwise be frustrated and quit in the free trial period) without reducing the number of committed students who initiate paid enrollment.

The “retention” (probability of payment, given enrollment) was also considered as an evaluation metric but was rejected. The main reason was that the sample size calculation indicated that far more cookies would be required to achieve the minimum detectable effect, given the baseline conversion rates. This would cause the experiment to require much more time than is feasible. The second reason for rejecting this metric is that the net conversion metric should give similar information.

The “click-through probability” (also known as click-through rate or CTR) was also considered as an evaluation metric. In this experiment this would be defined as the number of unique cookies to click on the ‘start free trial’ button in a day, divided by the number of unique cookies to view the page per day, i.e. ‘clicks’ divided by ‘cookies’ as defined earlier. The reason for not selecting this evaluation metric is that this measures the fraction of students who view the page who also click on the ‘start’ button. However, it does not measure the number of students who complete the enrollment process (registering with a user name and password and other details) to begin the free trial, and these students are the experimental focus. Therefore CTR is a less direct measure of the number of students in the free trial than the ‘gross conversions’ metric described above where enrollment was required (not only clicking the ‘start trial’ button).

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)* (1/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)
SD =0.20230604
2) Retention
SD = sqrt(p*(1-p)* (1/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)
SD =0.054949012
3) Net conversion
SD = sqrt(p*(1-p)* (1/N)) where
p= 0.1093125 (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)

SD =0.015601545

For gross conversion and net conversion, the unit of analysis and unit of diversion are the same (i.e. cookies) and therefore the analytical and empirical estimates of variability should be similar. For retention the unit of analysis is enrollment, and the unit of diversion is cookie, therefore the analytical estimate is more likely to underestimate the variability and an empirical estimate may be needed.

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.)

Answer:

I did not use the Bonferroni correction because my expectation is that the evaluation metrics are highly correlated, and therefore the Bonferroni correction would be too conservative (i.e. more false negatives, less false positives) compared with an ideal correction. If I were to use this correction I would re-work the sample size calculation with alpha=0.025 reflecting the two evaluation metrics I chose in the final analysis.

I used the sample size calculator at http://www.evanmiller.org/ab-testing/sample-size.html

I did the calculation for

1) Retention

Baseline conversion rate = 0.53 (from baseline values sheet)

Minimum detectable effect = 0.01 (from the minimum business significance)

Sample size (enrollments) = 39115 (from online calculator)

Sample size (pageviews with unique cookie) = 39115 / (0.08*0.20625) = 2370606.1 (given probability of enrollment and click through)

Sample size (pageviews with half for experiment, half for control) = 2*2370606.1 = 4741213 (rounded up).

2) Gross conversion

Baseline conversion rate = 0.20625 (from baseline values sheet)

Minimum detectable effect = 0.01 (from the minimum business significance)

Sample size (clicks) = 25835 (from online calculator)

Sample size (pageviews with unique cookie) = 25835 / 0.08 = 322937.5 (given probability of click through)

Sample size (pageviews with half for experiment, half for control) = 2*322937.5 = 645875 (rounded up).

3) Net conversion

Baseline conversion rate = 0.1093125 (from baseline values sheet)

Minimum detectable effect = 0.0075 (from the minimum business significance)

Sample size (clicks) = 27413 (from online calculator)

Sample size (pageviews with unique cookie) = 27413/ 0.08 = 342662.5 (given probability of click through)
Sample size (pageviews with half for experiment, half for control) = 2*342662.5 = 685325 (rounded up).

Given that the Retention requirements are too high, it was dropped from further analysis.
The final sample size chosen was the maximum of those calculated for the evaluation metrics (2) and (3), which was the net conversion result:
685325 page views (as measured by unique cookies per day).

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:
I chose 100% traffic exposure for an experiment duration of
Duration = (page views required) / (baseline value for page views per day * exposure)
Duration = 685325 / (40000 * 100%) = 17.1331 days
The final duration was rounded up to ensure there were sufficient days, i.e. 18 days.
The reason for choosing 100% exposure (i.e. all traffic diverted) was that the experiment was simply a cautionary message to the users, therefore would not be expected to have a dramatic impact on a key metric (e.g. paid enrollments) and so the risk of exposing this much traffic to the experimental condition was small. There is no chance that anyone would be seriously impacted by the nature or duration of the experiment, and we are not dealing with sensitive data (e.g. political attitudes, confidential information, etc.) Therefore the cost of running a longer experiment (lower exposure) that would also prevent other experiments is not justified.

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.**

Answer:
1) Number of cookies (pageviews with unique cookies)
Number of cookies (experiment) Ne = 344660

Number of cookies (control) Nc = 345543
p (proability 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)) = 0.000601841
Margin of error, m= SE * Z(alpha=0.05, 2-tail) = 0.0006 * 1.96 = 0.00118
Confidence interval lower limit, CI(low) = 0.5 - 0.00118 = 0.49882
Confidence interval upper limit, CI(up) = 0.5 + 0.00118 = 0.50118
Observed fraction, f = Nc / (Ne + Nc) = 0.50064
Therefore f > CI(low) and f < CI(up)
Conclusion: The observed fraction falls within the confidence interval, and the sanity check passes.

2) Number of clicks
Number of clicks (experiment) Ne = 28325
Number of clicks (control) Nc = 28378
p (proability event is assigned to control group) = 0.5
Standard error (binomial) SE = sqrt(p*(1-p)/(Ne+Nc)) = sqrt(0.5*0.5/(28378+28325)) = 0.00210
Margin of error, m= SE * Z(alpha=0.05, 2-tail) = 0.0021 * 1.96 = 0.00412
Confidence interval lower limit, CI(low) = 0.5 - 0.00412 = 0.49588
Confidence interval upper limit, CI(up) = 0.5 + 0.00412 = 0.50412
Observed fraction, f = Nc / (Ne + Nc) = 0.50047
Therefore f > CI(low) and f < CI(up)
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.)

Answer:
1) Gross conversion (probability of enrolling given click)
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)*(1/Ne+1/Nc)) =
sqrt(0.20861*(1-0.20861)*(1/17293+1/17260)) = 0.004372
Difference estimate, d = Xe/Ne – Xc/Nc = 3423/17260 – 3785/17293 = -0.02055
Margin of error, m= SE * Z(alpha=0.05, 2-tail) = 0.004372 * 1.96 = 0.008568
Confidence interval lower limit, CI(low) = -0.02055 - 0.008568 = -0.02912
Confidence interval upper limit, CI(up) = -0.02055 + 0.008568 = -0.01199

Statistical significance:

Zero is NOT within the CI = [-0.02912,-0.01199], thus the result has statistical significance.

Business (practical) significance:

For business significance the Confidence Interval (CI) must exceed the minimum level dmin in either the positive or negative directions.

i.e. CI must not overlap with NS = [-dmin, dmin] for significance.

In this case CI does NOT overlap with [-0.01,+0.01], thus the result has business significance.

Conclusion: Gross conversions are smaller in the experimental condition than the control condition.

2) Net conversion (probability of payment given click)

Number of clicks (control) Nc = 17293

Number of payments (control) Xc = 2033

Number of clicks (experiment) Ne = 17260

Number of payments (experiment) Xe = 1945

p (pooled probability of payment) = (Xc + Xe) / (Nc + Ne) = 0.11513

Pooled Standard error (binomial) SE = sqrt(p*(1-p)*(1/Ne+1/Nc)) =
 sqrt(0.11513*(1-0.11513)*(1/17293+1/17260)) = 0.00343

Difference estimate, d = Xe/Ne – Xc/Nc = 1945/17260 – 2033/17293 = -0.00487

Margin of error, m= SE * Z(alpha=0.05, 2-tail) = 0.00343 * 1.96 = 0.00673

Confidence interval lower limit, CI(low) = -0.00487 - 0.00673 = -0.01160

Confidence interval upper limit, CI(up) = -0.00487 + 0.00673 = 0.00186

Statistical significance:

Zero is within the CI = [-0.01160, 0.00186], thus the result does NOT have statistical significance.

Business (practical) significance:

For business significance the Confidence Interval (CI) must exceed the minimum level dmin in either the positive or negative directions.

i.e. CI must not overlap with NS = [-dmin, dmin] for significance.

In this case CI does overlap with [-0.0075,+0.0075], thus the result does NOT have business significance.

Conclusion: Net conversions are not significantly different in the experimental condition than the control condition.

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.)

Answer:

To calculate the sign tests' probability value listed below, I used the calculator at:
http://graphpad.com/quickcalcs/binomial1.cfm

A “success day” for the calculation of Ns below means that the probability of the relevant rate (e.g. click rate for gross conversions, payment rate for net conversions) is greater in the experimental condition than in the control condition for a given day.

1) Gross conversion (probability of enrolling given click)

Number of successes (days where enroll rate is higher in experiment than control), Ns = 4

Number of trials (total number of days with enrollment data), Nt = 23

Probability for a success day under null hypothesis, p0 = 0.5

Probability value (2-tail) from sign test: p = 0.0026

Given that p = 0.0026 < alpha = 0.05

2) Net conversion (probability of payment given click)

Number of successes (days where payment rate is higher in experiment than control), Ns = 10

Number of trials (total number of days with enrollment and payment data), Nt = 23

Probability for a success day under null hypothesis, p0 = 0.5

Probability value (2-tail) from sign test: p = 0.6776

Given that p = 0.6776 > alpha = 0.05

Conclusion: The gross conversion metric is significant and the net conversion is not significant from the sign tests.

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.

Answer:

I did not use the Bonferroni correction because in this experiment we need *all* metrics to match the expectations in order to launch the change. In this case the risk of a false negative (Type II error) increases as the number of metrics increase. The Bonferroni correction is not suitable in this case, as it is intended to reduce the risk of a false positive (Type I) error at the expense of Type II errors. Bonferroni correction would be more appropriate if we were intending to launch the change if only a *single* metric were to test positive.

There was agreement between the results of the effect size hypothesis tests and the sign tests. In both cases the gross conversion test was significant whereas the net conversion test was not significant. Therefore there was no discrepancy to explain.

Recommendation

Make a recommendation and briefly describe your reasoning.

Answer:

The experiment provided information to students that they should devote 5+ hours per week if they wish to enroll in the program, and the “hypothesis that this might ...reduce the number of

frustrated students who left the free trial without significantly reducing the number of students who continue past the free trial and eventually complete the course."
The experimental results gave evidence that this experiment condition did reduce the number of gross conversions (people who enrolled after clicking on the free trial) which was the desired result. The statistical significance test indicated that the experiment did not change the number of net conversions (people who enrolled and paid after clicking on the free trial), which was also the desired effect.
However, the confidence interval for the net conversions [-0.0116, 0.0019] was not wholly above the lower boundary of the interval for practical significance (-0.0075). Therefore we do not have 95% confidence that the net conversions were above our criterion for practical significance, and thus the change may have a negative impact on the business as this would imply the possibiity of fewer paying students.

Based on this evidence, I would not launch the change i.e. not make the experimental condition a permanent change for enrolling students.

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:
The follow-up experiment that I propose is that the students are given a fee reduction incentive that is maximum if they pay on Day 1 after enrollment (e.g. 14% off) and linearly decreases to zero on Day 14. This would create urgency, a well known technique to persuade consumers.

Invariant metric:
The user-ids (i.e. enrollees in the free trial) would be the invariant metric, as there should be the same number in experimental and control groups.
Evaluation metric:
1) "Retention" (payment probability) defined by the number of enrollees that complete payment prior to their (individual) 14 day expiry period, divided by the total number of enrollees. The "total number of enrollees" would be determined by the sample size needed to reach adequate power for statistical testing. The expectation is that the experimental condition would increase the Retention for a given group of enrolled students.
2) "Payment Velocity" would be a second (**optional**) metric. This is defined as the Retention multiplied by Expiration Remaining, where the latter is the number of days remaining until the free trial expires. This would be interesting to test if the students are more likely to pay earlier in the free trial period with the incentive, than in the control condition. (It should be noted that this is interesting but does not directly address the question of reducing free trial students while not reducing paid students. Therefore it is not discussed further.)

Null hypothesis, H_0: The change in Retention (d) in the experimental condition and control conditions is zero (alpha=0.05).

Alternate hypothesis, H_A: The Retention in the experimental condition after subtracting the Retention in the control condition is not equal to zero (alpha=0.05). This difference (d) in Retention is the variable to be tested for significance.
Unit of diversion: The user-ids of users enrolled in the free trial is the unit of diversion. This was chosen because I'm only interested in the students that have already decided to enroll (whereas the using "clicks" on the free trial button would include more noise from the students who don't complete the subsequent steps to enroll).
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. (d_{min} = 0.01 according to the provided information.) This means that if the confidence interval is represented as
$CI = [C_{low}, C_{up}]$
then the launch requirement is that
$C_{low} > d_{min} > 0$

Conclusions: If the launch criteria are met, then a larger proportion of users who begin the free trial were found to continue and complete payment. Therefore it would be my recommendation to launch the change, introducing the time-limited incentive for future enrollees.