Design and Analyze A/B Testing

by Prasad Pagade, 3/10/2017.

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.)

- Invariant metrics: Number of cookies, Number of clicks and Click-through-probability.
- Evaluation metrics: Gross conversion, Retention and Net Conversion.

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.

- **Number of cookies** : This is essentially number of unique cookies to visit the course overview page. This is a good population metric and can be easily split evenly between control and experiment group. This is also independent from the experiment. Hence, this is my choice of an invariant metric.

- **Number of clicks**: The clicks happen before the experiment is triggered and hence is independent. The number of clicks will be the same for control and experiment group and hence this is a good invariant metric.

- **Click-through-probability**: This is again independent from the experiment page since the clicks occur before the users reach the experiment page. Therefore, this metric does not depend on our test and is a good invariant metric.

- **Number of User-ID**: This is not suitable for evaluation & invariant metric. The number of users enroll in the free trial are dependent on the experiment, we expect to see different value in the control and experiment group, therefore it cannot be a good invariant metric. Meanwhile, user-id is a count of users accessing the page and is not normalized like gross conversion which includes the user-ids and is a better way to track the effect of the experiment. Therefore, we do not use user-ids as an evaluation metric.

- **Gross conversion**: This is a great evaluation metric since it's directly dependent on the effect of the experiment and allows us to show whether we managed to decrease the cost of enrollment by decreasing the number of students who are likely to cancel before the free trial of 14 days. The experiment group users will get to decide based on the time commitment if they want to enroll or continue with the free courses. The underlying hypothesis is that these students who enroll after viewing the "time commitment page" are more focused than the control group and are likely to be retained after the 14 days trial.

- **Retention** (*Number of user-ids to remain enrolled for 14 days trial period and make their first payment/Number of users who enrolled in the free trial*): The retention of students is directly affected from this experiment. Based on our hypothesis, the control group retention is expected to be lower compared to experiment group as they may lack the time commitment required to complete the courses resulting into cancellation. This expectation is already set forward with the experiment group before they enroll in the course.

- **Net conversion**: For the experiment group, users are made aware of the time commitment requirement through the screener page and from there they can make a decision to remain enrolled for the course past 14 day trial-period. However, on the control side, they would be able to continue the payment without being aware of the minimum time requirement. The net conversion is the final conversion after gross conversion and retention and we expect to see whether the users in the experiment group could make better informed decision and stay committed leading to first payment for Udacity.

  The evaluation metrics were chosen as they are expected to have different distributions in the control and experiment group. The aim of the screener page is to set the time expectation with the students upfront so that it leads to less frustration later and causes the student to stay enrolled past the 14 day boundary and make the 1st payment. With this in mind, in order to launch the experiment either of the following must be observed:

  - Increased retention (more students staying beyond the free trial in the experiment group)
  - Decreased Gross Conversion leading to **unchanged or increased** Net Conversion (less students enrolling in the free trial but more students staying beyond the free trial)

Measuring Standard Deviation

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

Link to the Project Baseline Values: data

```
> Gross conversion: se = ROUND(SQRT((0.20625*(1-0.20625))/(5000*(3200/40000))),4) =
0.0202

> Retention: se = ROUND(SQRT((0.53*(1-0.53))/(5000*(660/40000))),4) =0.0549

> Net conversion: se =ROUND(SQRT((0.1093125*(1-0.1093125))/(5000*(3200/40000))),4) =
0.0156

Final Results:
Gross conversion: 0.0202
Retention: 0.0549
Net conversion: 0.0156
```

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.

The unit of diversion for Gross and Net conversion is **cookies**. Since cookies is the unit of diversion and unit of analysis, the analytical estimate would be comparable to the empirical variability.

For retention, the unit of diversion(cookies) is not the same as unit of analysis(user-ids). Hence, if we choose retention as our metric then we need to calculate the analytical estimate and empirical variability.

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.)

My approach will not deploy the Bonferroni correction, as the metrics in the test have a high correlation. Pageviews required for each metric were calculated using an alpha value of 0.05 and beta value of 0.2.

Gross Conversion

- Baseline Conversion: 20.625%
- d_min: 1%
- α: 5%
- β: 20%
- 1 - β: 80%
- sample size = 25,835 enrollments/group
- Number of groups = 2 (experiment and control)
- total sample size = 51,670 enrollments
- clicks/pageview: 3200/40000 = 0.08 clicks/pageview
- pageviews = 645,875

Retention

- Baseline Conversion: 53%
- d_min: 1%
- α: 5%
- β: 20%
- 1 - β: 80%
- sample size = 39,155 enrollments/group
- Number of groups = 2 (experiment and control)
- total sample size = 78,230 enrollments
- enrollments/pageview: 660/40000 = .0165 enrollments/pageview
- pageviews = 78,230/.0165 = 4,741,212

Net Conversion

- Baseline Conversion: 10.9313%
- d_min: 0.75%
- α: 5%
- β: 20%
- 1 - β: 80%
- sample size = 27,413 enrollments/group
- Number of groups = 2 (experiment and control)
- total sample size = 54,826

- clicks/pageview: 3200/40000 = .08 clicks/pageview
- pageviews = 685,325

Pageview selected: 4,741,212

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.)

Since our experiment is not risky as it is only a screener page and does not contain any sensitive data on the users we can choose to expose 100% of the users to the experiment. With a daily traffic of 40,000 pageviews our experiment will take 4,741,212/40,000 = **119 days** to run. This is a long duration and may not be feasible.

Thus, we switch to **Net Conversion** for the pageviews. To accommodate any other experiments that may be running in parallel, we choose to direct 70% of the traffic to the experiment. So, 28,000 out of the 40,000 pageviews will be directed to the experiment. This will take 685,325/28,000 = **25 days** which is far more acceptable.

Thus, we discard **Retention** as the evaluation metric.

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

I choose to divert 70% of the traffic to the experiment as our experiment does not pose a high risk to the users. This experiment is not high risk as the screener page suggests the users who want to enroll about the minimum time required to put into the course. Users must choose to enroll or move to the free courses thereafter. The operation of the website is not affected in any way. I kept 30% in the control group so that it can provide traffic for any other experiments that may be run in parallel to this one.

Experiment Analysis

Sanity Checks: Data

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.)

After the experiment is run, we expect that the cookies and clicks are evenly divided distributed with a probability P = 0.5. We will test this at the 95% confidence interval.

Metric	Lower Bound	Upper Bound	Expected Value	Observed Value	Result
Number of Cookies	0.4988	0.5012	0.5000	0.5006	Pass
Number of clicks on "start free trial"	0.4959	0.5041	0.5000	0.5005	Pass

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.)

95% Confidence interval for the difference between the experiment and control group for evaluation metrics.

Metric	dmin	Difference	Lower Bound	Upper Bound	Result
Gross Conversion	0.01	-0.0205	-0.0291	-0.0120	Statistically and Practically Significant
Net Conversion	0.0075	-0.0048	-0.0116	0.0019	Neither Statistically Nor Practically Significant

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.)

Metric	p-value	Statistically Significant α = 0.05
Gross Conversion	0.0026	Yes
Net Conversion	0.6776	No

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.

Our launch decision is based on the significant difference of the two metrics (gross and net conversion) rather than a single metric out of many options. Also, the success of the experiment depends upon the magnitude and direction of **both** the metrics, hence we choose not to use the Bonferroni correction.

The sign tests allow for an additional form of analysis on top of effect size test that helps us verify the significance level of our metrics. Had we any discrepancies about the significance of the evaluation metrics between the sign and effect size tests, further study would be warranted. In this case, both tests agree and our conclusions about both metrics are strongly supported.

Recommendation

Make a recommendation and briefly describe your reasoning.

My recommendation is that we should *NOT* launch the experiment because of the following reasons:

For the Gross conversion, the experiment was effective at reducing the number of student enrollments who didn't have the time commitment required to take the course. This is within our expected behavior to launch the experiment. But for the net conversion, we noted that the difference is not significant. In fact, the confidence interval includes the negative practical boundary, so it's possible that the number of enrollments went down by an amount that would matter to the business. This is not an acceptable risk for Audacity and hence the final recommendation is not to launch the experiment.

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.

From our experiment, we found that the number of students retained past the 14-day trail did not improve. My suggestion to improve the net conversion would be to motivate students at day 7 by showing them series of success stories of past students - who initially struggled, then succeeded through course to eventually make a transition to a new career.

Hypothesis: The success stories pop-up will motivate students and increase the retention past the 14-day trial period.

Null Hypothesis: The success stories pop-up will worsen the retention past 14-day trial period.

Invariant Metric: The invariant metric would be user-ids as we would equally distribute students to either control group or experiment group.

Unit of Diversion: The unit of diversion would be user-id as the experiment will show up when the students enroll the course and progress through it during the 14-day trial period.

Unit of Evaluation: The unit of evaluation would be retention as we would like to track if there was an increase in number of students who saw the experiment enrolled past the 14-day trial period. A statistically and practically significant increase in retention would enable us to launch this experiment.