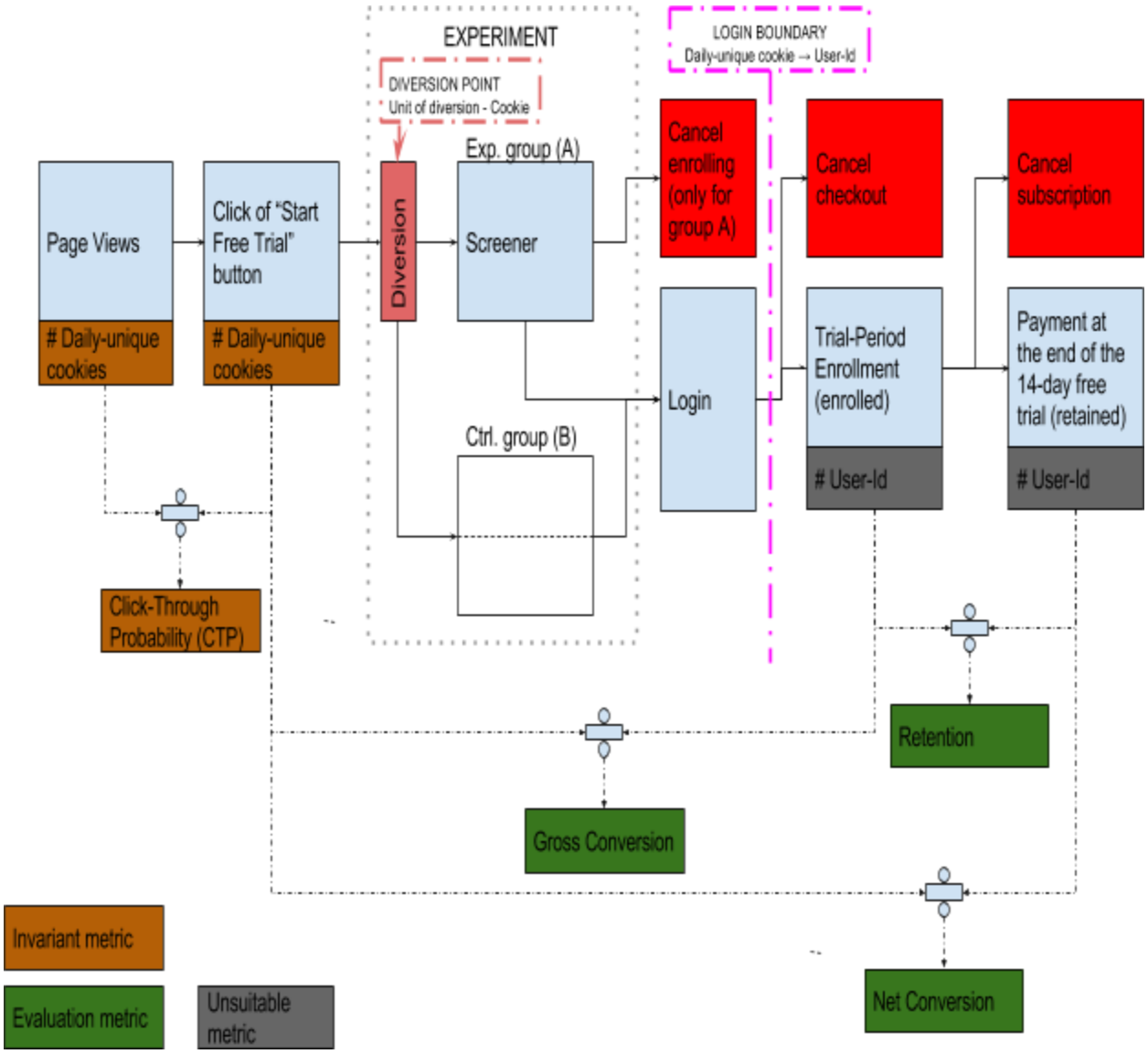
Data Analyst Nanodegree
Project #7: A/B Testing

Ivailo Kassamakov, Nov 10, 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.

Invariant Metrics

The invariant metrics are expected to be similar between the control and experiment group (the A and B group). As such they can be used to verify the correctness of the experiment setup and execution.

- **Number of cookies:** the number of unique cookies to view the course overview page.
  Since the unit of diversion is "cookie", it is logical to assume that the number of visiting cookies will be equally distributed between the A & B groups.
- **Number of clicks:** the number of unique cookies to click on the "Start Free Trial" button (before the free trial screener has appeared).
  Since the experiment takes place *after* the "Start Free Trial" has been clicked, we expect to see similar number of unique-cookie clicks between the A and B groups (given that the number of unique cookies viewing the page is also similar).
- **Click-through probability (CTP):** the number of clicks (as defined above) divided by the number of cookies (as defined above).
  Since both constituent quantities in this ratio are expected to be invariant, ditto for the ratio itself.

Evaluation Metrics

The evaluations metrics are supposed to have different values for the control and experiment groups, which can be used to decide if the experiment is successful.
Each evaluation metric needs to show a difference between the experiment branches bigger than a minimum practical significance level d_{min} to be considered an indicator of success.

- **Gross conversion:** the number of user-ids to complete the checkout and enroll in the free trial divided by the number of clicks (as defined above).
  Given the filtering nature of the experiment, we would want to see LESS people completing the checkout in the experiment group (the one that sees the free trial screener).
- **Net conversion:** the number of users to remain enrolled past the 14-day trial period divided by the number of clicks (as defined above).
  The ultimate goal of the experiment is to achieve an INCREASE in the net conversion.
- **Retention:** the number of users to remain enrolled past the 14-day trial period divided by the number of users to enroll in the free trial.
  In case of a successful experiment we would expect to see an INCREASE in this metrics.

If the proposed UI change works as expected, and the free-trial screener filters well, we would see a decrease in the "gross conversion" metric for the experiment group.
On the other hand, in the same conditions, the other two evaluation metrics-"net conversion" and "retention" should see an increase for the experiment group, since we suppose that the

people who make it to the free-trial enrollment after the screening will be more motivated to continue the course.

Metric	Behavior in a successful experiment
Gross Conversion	Decrease (less motivated people get filtered)
Retention	Increase (people enrolling for the trial are more motivated)
Net conversion	Increase (same reason)

Unsuitable Metrics

- **Number of user-ids:** the number of user-ids to complete the checkout and enroll in the free trial.

  This metric is not suitable for our analysis. First it cannot serve as invariant since user-ids are recorded only after the user has decided to enroll for the free trial, and we can't expect it to have similar values between the control and experiment group.

  Theoretically, it could be used to track retention beyond the 14-days trial period, however it is not very convenient due to its absolute (non-normalized, non-ratio) nature.

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.

We possess the following baseline values of the metrics

Metric	Baseline value
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

Based on this we calculate the following analytic SD for the evaluation metrics:

Evaluation Metric	SD	Unit of analysis	Analytical var = Empirical var?
Gross conversion	.0202	cookie	yes
Retention	.0549	user-id	no
Net conversion	.0156	cookie	yes

We can expect to see similar empirical and analytical variability for those metrics whose unit of analysis coincides with the unit of diversion (cookies). This is the case for Gross conversion and Net conversion, but not for Retention. The Retention metric will require an empirical calculation of its 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.)

In the beginning, during the metric selection process, **no Bonferroni** correction will be done.

The following sample sizes were calculated using the empirical_sizing.R script provided with the lessons. The calculations were done for significance level alpha=0.05 and power level beta=0.2.

Metric	Required sample size for 1 group	Required sample size for both groups	Factor for converting to page views	Required page views
Gross Conversion	25699	51398	0.08 clicks/view	642’475
Retention	39104	78208	0.0165 enrolls/view	4’739’879
Nett Conversion	27172	54344	0.08 clicks/view	679’300

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?

To get a feeling of the best-case duration required to run the test, we'll start by assuming that 100% of the daily traffic (40'000 page views) is exposed to the test. This gives:

Metric	**Duration with 100% traffic exposure [days]**
Gross Conversion (GC)	16
Retention	119
Nett Conversion (NC)	17

We easily see that the Retention metric is very expensive in terms of duration. Waiting for 119 days with 100% allocated traffic poses unnecessary business risks (takes up resources and excludes the running of other experiments in parallel).

For this reason we will have to **drop** the Retention metric and leave only GC and NC as our evaluation metrics. In this case with 100% traffic exposure we would need 17 days to complete the experiment.

We don't consider this site change very risky since we don't envisage an enormous reduction in the enrollment in case the free trial screener is used. This theoretically could allow us to expose all of the traffic to the test. However, other business needs could dictate a lower exposure, e.g. to allow other experiments to run at the same time. With an exposure of 50% we would need around 35 days to complete the experiment, which is still an acceptable value.

Multiple-metrics setup and Bonferroni correction

An important moment in our experiment design is that we would want to see practically significant changes in **both** metrics (GC and NC) to be able to decide in favour of launching the experiment.

The issue we have with this multiple-metrics setup where we require the acceptance of ALL alternative hypothesis to deem the experiment as success, is that the type II error probability is increased. This renders the use of the Bonferroni correction **unsuitable**, since it controls the familywise error rate (the type I error rate for the ensemble of metrics) at the expense of the type II error rate (which is exactly what we would like to avoid).

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.

We would expect the “number” invariants to be equal between the two branches of the experiment (i.e. the ratio of the “control-group” metric to the “both-groups” metric to be 0.5). For the “probability” metrics we would expect a 0 difference between the control and experiment groups’ metric.

Metric	Expected value	Lower Bound	Upper Bound	Observed value	OK?
Number of cookies	0.5	0.4988	0.512	0.5006	YES
Number of clicks	0.5	0.4959	0.5041	0.5005	YES
CTP	0.0821	0.0812	0.0831	0.0822	YES

The sanity check show a correctly setup and performed experiment.

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.)

The metrics are statistically significant if they don’t contain 0 in their confidence interval.
The metrics are practically significant if their confidence interval is outside of the d_{min} minimal practical difference.

Metrics	Difference (control-exp)	Lower Bound	Upper Bound	Min. practical difference	Decision
Gross	-0.0206	-0.0291	-0.0120	0.01	Statistically

Conversion					and practically significant!
Net Conversion	-0.0049	-0.0116	0.0019	0.0075	Statistically and practically NOT 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.)

The calculation of the p-values for the sign tests were performed with the calculator at http://graphpad.com/quickcalcs/binomial2/
The decision of statistical significance is done for alpha=0.05.

Metrics	**# "successes"**	**# "trials"**	**p-value**	**Stat. significant?**
Gross Conversion	19	23	0.0026	Yes
Net Conversion	10	23	0.6776	No

The results of the sign test confirm the verdict of the hypothesis tests that the NC metrics is not significant.

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.

The experiment was designed with the following features:

- Three invariant metrics were selected and later were shown to pass the sanity check: number of pageviews, number of clicks and CTP
- Initially three evaluation metrics were assessed: GC, Retention and NC. However, the Retention metrics was found to require an impractically long duration of the experiment, and as a result it was dropped. The experiment was staged further as a multiple-metrics test, with the requirement that all metrics be found practically significant in order to approve the launch of the tested design change.

- As explained above, the Bonferroni correction was found unsuitable, since it would increase the Type II error rate.
- The hypothesis and sign tests performed on final experiment results showed that the GC metrics showed a practically and statistically significant DECREASE in the number of enrolling students. Unfortunately, the NC metrics failed to show both practical and statistical significance.

Recommendation

Make a recommendation and briefly describe your reasoning.

The GC showed a decrease of the students opting for the free trial, which is in line with our expectations. However, the NC did not show any increase in the converted students, which is our ultimate business goal.

Therefore, I would recommend NOT to launch this design change.

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.

We can interpret the results in the following way: The screening of students who cannot dedicate enough time to the course certainly reduce the number of enrollments from people who will ultimately get frustrated and quit due to insufficient time resources. However, the lack of increase in the net conversion rate indicates that there might be other reasons for getting frustrated which have not been addressed by the free-trial screener. We could think e.g. of course descriptions that have not been precise or detailed enough, or of lack of necessary knowledge and skills in the student to complete the course.

A very straightforward follow-up experiment could be set up to address the latter frustration reason, i.e. the lack of prerequisite knowledge. In this new experiment, the original free-trial screener can be extended with a questionnaire that tries to evaluate if the candidate possesses the needed knowledge. The experiment design, including diversion unit, metrics choices and sizing can remain the same.

>> END <<