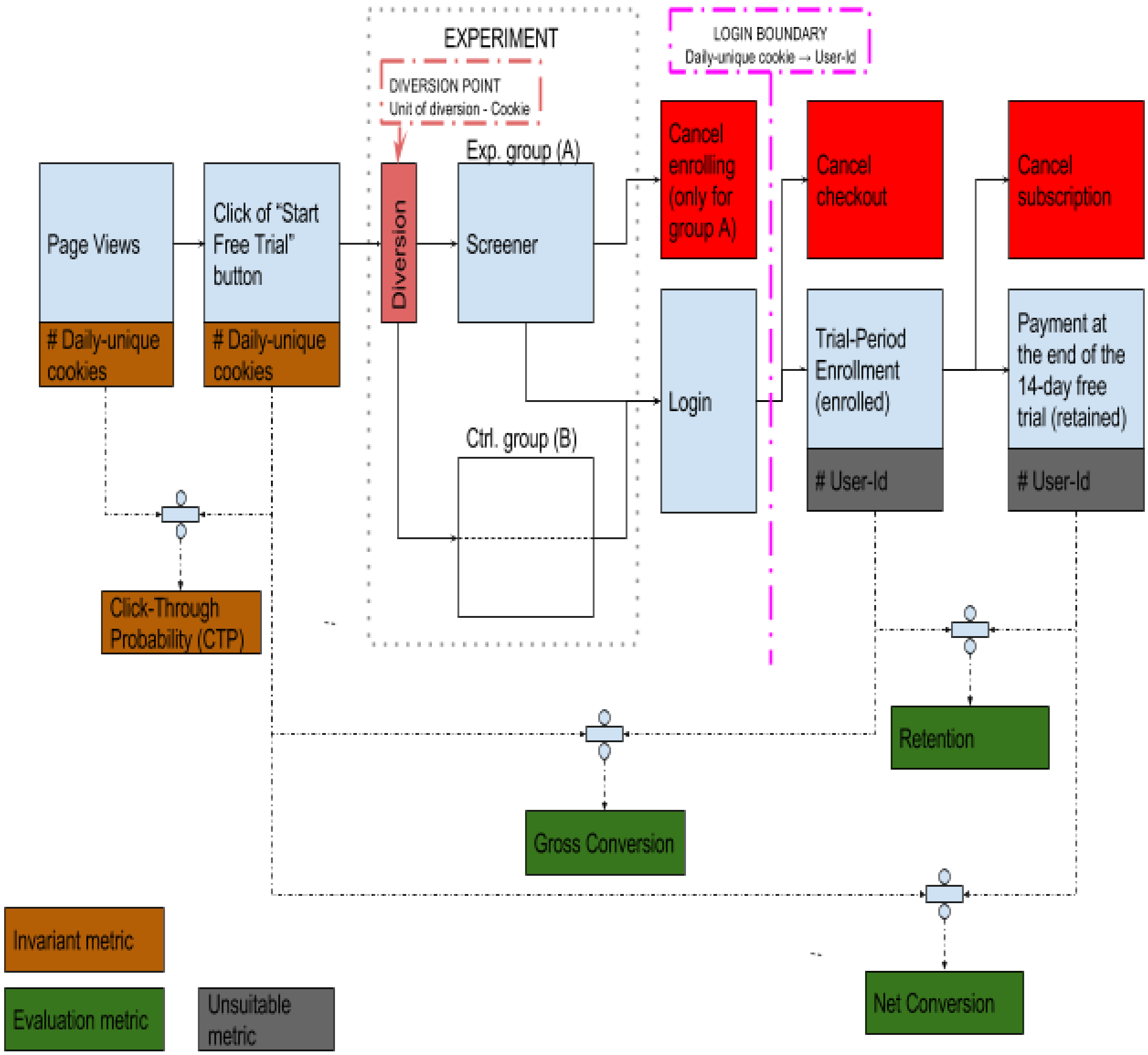
Data Analyst Nanodegree
Project #7: A/B Testing

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.
We expect these metrics to change since they are measured after the diversion point (contrary to the invariant metrics).
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 new design should avoid causing a decrease 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.
  Similarly to the Net conversion, in case of a successful experiment we would not want to see a decrease in this metrics either.

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. At the same time, we would NOT want to see any decrease in the other two evaluation metrics-”net conversion” and “retention”. This effectively means that the screener, while filtering out the unprepared enrollments, should not negatively affect the number of people that ultimately decide to pay.

Metric	Launch criteria
Gross Conversion	Decrease (less motivated/prepared people get filtered)
Retention	No decrease (screening does not decrease the number of people that ultimately decide to pay)
Net conversion	No decrease (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

If we decide to employ all three metrics, then we will be obliged to meet the most stringent pageview requirement - that of 4’739’879 for the “Retention” metric.

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)	16.06 → 17
Retention	118.5 → 119
Nett Conversion (NC)	16.9 → 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 experiment risky since it doesn't involve manipulation of sensitive information (such as medical or financial), nor does it have any potential of harming the participants. It doesn't pose any significant business risk either - which could be for example a significant drop in the conversion rate - and even if it causes such a negative effect, this can be quickly detected and remedied. This low-risk nature theoretically allows 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. To comply with the client expectations, we have to limit the duration to "several weeks". A maximum value of 30 days looks acceptable, which translates to a traffic exposure of minimum 56.6% (22'643 pageviews daily).

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 shows 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 Conversion	-0.0206	-0.0291	-0.0120	0.01	Statistically 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.

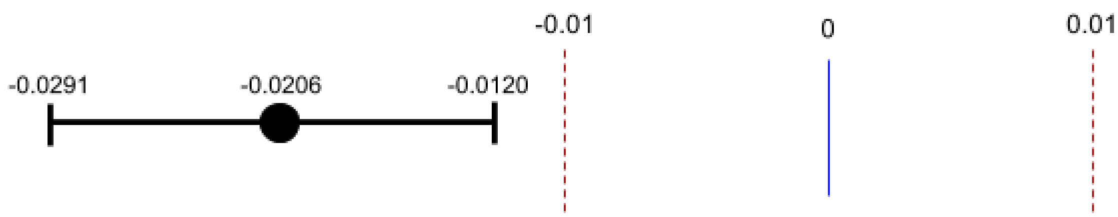
The experiment was designed with the following features:

- 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.
- The NC metrics while showing both practical and statistical significance, has a negative point estimate, which is NOT consistent with our goal of avoiding a decrease in the conversion rates.

Recommendation

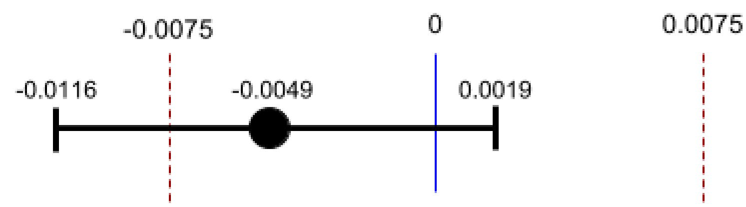
Make a recommendation and briefly describe your reasoning.

The GC showed a practically significant DECREASE of the students opting for the free trial, which is in line with our expectations and consistent with our goal of reducing the number of unprepared enrollments.



Gross Conversion's 95% CI

However, the NC was found to be statistically insignificant (CI contains 0) and possibly practically insignificant (the point estimate is within the practical difference boundaries). Worse, its CI contains the negative practical difference boundary, which means that there is a potential for practically significant DECREASE in the conversion rate. Unfortunately, given our launch criteria, this poses a business risk that cannot be accepted.



Net Conversion's 95% CI

Therefore, I would recommend NOT to launch this design change, but instead to focus on coming up with an improved idea for screening out the unmotivated enrollments.

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 of the above experiment in the following way: The screening of students who cannot dedicate enough time to the course certainly reduces the number of unmotivated enrollments. However, the potential slight decrease 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 design of such an experiment is very similar to the original one and will look like this:

- The free trial screener is extended with YES/NO questions asking the candidate if he possesses certain skills/knowledge/experience.
- After the screener is filled in, the candidate receives a recommendation whether enrolling for the free trial is reasonable or not.
- Our hypothesis is that candidates who are screened both in terms of available time AND prior skills/knowledge have LESS chance of dropping from the free trial due to frustration. At the same time the fact that they are required to spend some time filling the questionnaire can also filter people with insufficient general motivation.
- In case our hypothesis is true we would like to see a DECREASE in the free trial enrollments, while the conversion rates remain the same or INCREASE.
- The unit of diversion is daily-unique cookies (since login is obligatory only immediately before the enrollment, so no user-id is normally available at the time of diversion).
- To verify the correct execution of the experiment we will monitor several invariant variables. Suitable metrics are ones collected before the diversion point - number of pageviews (counted via the daily-unique cookies), and number of free-trial-button clicks (again counted via the cookies).
- The launch decision will be taken based on the analysis of the evaluation metrics. Suitable metrics are ones involving quantities measured after the diversion point. Appropriate ones to monitor are GC and NC rates. The Retention rate would not be appropriate since, as we saw in the first experiment, it requires a lot of traffic to gather enough statistics.

- To take a launch decision, we would want to see the following simultaneous dynamic of the evaluation metrics: GC decreasing due to a reduction of the unwanted enrollments, NC remaining the same or probably increasing.
- The risk associated with this experiment is very low since it doesn't collect sensitive information or harms the users in any way. This allows us to expose any fraction of the traffic to it (from 56% up to 100%, corresponding to a duration from 30 days down to 17 days).
- Like in the original metrics, there is no issue collecting the required statistics (cookies and user-ids) using the available client infrastructure.

>> END <<