A/B Testing

By Luca Taroni

Experiment Design

Metric Choice

Invariant metrics:

1. ‘Number of cookies’
2. ‘Number of clicks’
3. ‘Click-through-probability’

Those figures are not affected by the change, they happen in the process before the modification we want to test and are therefore suitable to be used for sanity checks.

‘Number of user ids’ varies but it’s not a metric in that, by itself, it cannot help assessing if there are changes, it needs to be related to other traffic measures to become meaningful.

Evaluation metrics:

1. Gross conversion
2. Retention
3. Net conversion

The Gross Conversion Is not the best metric to address our hypothesis, it would help us understand how many people think twice and give up the free trial but that is not answering to our specific hypothesis: we would not be able to assess how many student abandon the free trial because they don’t have enough time so we wouldn’t know if we have managed setting cleared expectations.

The Retention would answer to our question partly. We would understand if ‘warned’ students are more successful in remaining enrolled, in other words if the message proves useful selecting more motivated student and discouraging less motivated students but it would not help us understand the second part of the hypothesis: are we managing not to significantly reduce the amount of student to continue past the free trial?

The Net Conversion would possibly answer another part of the question: The result would derive from the combination of both the ‘discouragement’ effect (how many people don’t meet the 14 days goal because they drop off and don’t complete checkout) and the number of students that, after completing checkout would remain enrolled. So the probability would be the result of two distinct elements.

Concluding: My understanding is that we would need to use all the 3 metrics to have a clear picture, namely to understand to which extent the change is discouraging students, if the change is successful in selecting more motivated students and which the combined effect is. This last part (the combined effect) is best described by the Net Conversion that could be the metric of choice if we were to select a single metric for some reasons. A full understanding of the figures though would be achieved by considering the 3 metrics and their individual meaning and interacting contribution to achieve an appropriate reading of the figures emerging from the experiment.

Regarding the Bonferroni correction: My choice is to use the three metrics and there is a potential need to account for the increased likelihood of witnessing a rare event, more specifically a Type I error (False Positive). That likelihood is increasing with the number of tests. My opinion is that 3 tests is not such a 'large number'. Considering our alfa (0.05) the chance of a False Positive would be 14,3% **if** the metrics were independent, but our metrics might reasonably have a certain degree of correlation and are possibly not independent, therefore the previous one is quite certainly an overestimate of the probability of a False Positive. The presence of some degree of correlation would as well render the Bonferroni correction even more stringent and therefore too conservative in this context. For the aforementioned reasons I will not use the Bonferroni correction in my analysis.

I'd be glad to receive the most detailed feedback possible regarding my conclusions on the Bonferroni correction as I'm aware of the fact that this is a delicate issue with a lot of aspects to be considered at the same time.

Measuring Standard Deviation

Gross Coversion	0.0202
Retention	0.0549
Net Conversion	0.0156

I would use the estimated variance in each case. In the case of Retention both the numerator and the denominator are expressed in the same unit so there is no reason to suspect for a variance increase. When the denominator differs from the numerator, in the Gross and Net Conversion I might be tempted not to use analytical variance as well as I don't think that the unique cookie in the denominator might introduce a consistent violation of independence: I find it not so likely that a user would systematically 'start the free trial' with several devices and /or cancelling their cookies. There might be exceptions but my hypothesis is that those would not be statistically relevant.
I therefore imagine that cookies might be logically compared to user IDs in this particular case.

I'm aware that practical experience and knowledge might disprove my guess. I'd be happy to be helped with some guidance and proper reasoning on this both if I'm right or wrong.

Sizing

Number of Samples vs. Power

Here's the process I've been following: No matter the Bonferroni correction (that would just increase the figure by reducing the alfa level) the number of pageviews required to power the Retention metric would be too high requiring about 120 days of 100% traffic.
I therefore excluded the Retention metric and did not apply the Bonferroni correction as from my initial reasoning.
The metric requiring more pageviews is the Net Conversion so I'm reporting just that one as the power required would obviously be sufficient for the metric requiring less power.
The actual number of pages for Net Conversion, multiplied by two to account for the need to have a control and an experiment group would be 685275.

Duration vs. Exposure

I would divert 100% of the traffic, the number of days (rounded up) required would be 18.

I would be ok diverting 100% of the traffic, 50% would anyway not change for the control group and I can't see risks associated to the introduction of a message: The risk is economic, there is potentially the possibility that as a consequence of the modification a significant drop in enrollment would materialize but the type of message and modification makes me think that the effect would be limited to the expectable increase or drop, shortening or lengthening the experiment would not change that.

Experiment Analysis

Sanity Checks

	Lower bound	Upper bound	Value
Number of cookies	0.4988	0.5012	0.5006
Number of clicks	0.4959	0.5041	0.5005
Click-through-probability	0.0812	0.0830	0.0822

All the sanity checks are passed; the values are in the confidence interval.

Result Analysis

Effect Size Tests

	Lower bound	Upper bound	Stat sig.	Pract. Sig
Gross conversion	-0.0291	-0.0119	Yes	Yes
Retention	0.0081	0.054	Yes	No
Net conversion	-0.0116	0.0018	No	No

Sign Tests

	P value	Stat sig.
Gross conversion	0.0026	Yes
Retention	0.6776	No
Net conversion	0.6776	No

Summary

I haven’t used the Bonferroni correction as from my initial setup and reasoning regarding it.

Recommendation

It seems that the only statistically and practically significant metric is the Gross Conversion. Though we used it to carry on our computations, the Retention metric has been excluded because of the power needed, it would anyway not have been practically significant and not even statistically significant according to the sign test. As for the Net Converstion it appears to be neither practically nor statistically significant, also in the sign test.

As for the Gross conversion it seems that the advice in the checkout process is reducing the number of people that would complete the checkout by something among 1,19% and 2.91%, quite a meaningful figure. The Retention and the Net Conversion do not provide practically significant results. The Retention seem to suggest that those who don't abandon the checkout are more motivated and actually manage to reach the first payment at a higher percentage, that percentage, though statistically significant, is lower than the practical threshold we set so it doesn't support the adoption of the change. An even more blurred picture emerges from the Net Conversion, our overall preferred metric: No statistically significant results are delivered.

I would therefore consider that there are no elements to support the launch of the change because the only effect we would have, and that is statistically meaningful, would be to reduce the number of students that complete checkout. We would not be able to support our hypothesis neither in the part where it states that we manage to reduce the students early leaving the free trial nor that we can do so without reducing the number of students that continue past the free trial.

Follow-Up Experiment

I'd perform some data mining using socio-demo-browsing data and amount of hours that the student is ready to devote to the course vs. drop-off rate before 14 days to identify the potentially dropping off students. A logistic regression would be a robust and helpful starting point.

I would then propose some sort of 'incentive' or 'challenge task' to those that have a probability higher than some threshold to drop-off according to my logistic regression. The incentive could be a discount, or a challenge with a prize like "complete in 6 weeks and get a 30% off on your course and on the next one". That option would be accessible via a button on the student course page, where the start a free trial button is now. It would be a separate data mining task to decide after which period that option would be available in the student course page, I would not discuss that now.

My **hypothesis** would be that the potentially dropping off students would, in a certain percentage, not drop off because of the incentive offered.

The **unit of diversion** would be the userID, therefore it would not represent a challenge in terms of variance as we would compute metrics that regard enrolled userIDs and calculate some metric based solely on userIDs.

The **metric** would be the drop off Conversion Rate: the number of (potentially dropping off) user IDs that drop off divided by the number of (potentially dropping off) user IDs that complete checkout. By comparing the results in the control and in the experiment group we would assess whether the change supports our hypothesis conducting the analysis in a similar fashion to the one implemented in this submission.