A/B Testing

By Luca Taroni

Dear reviewer, please be inflexible, intolerant and merciless. I'd really like to learn properly every single aspect of this exam. I'll do my best to fulfill every requirement in my next submission in the likely event this one would need to be improved.

Experiment Design

Metric Choice

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.

I've chosen:

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

as invariant metrics because those are not affected by the change, they happen before the modification we want to test in the process and are therefore suitable to be used for sanity checks.

Conversely:

1. Gross conversion
2. Retention
3. Net conversion

are potentially affected by our modifications, they happen after the change in the process. Each of them might differ in the experiment group if our modification has an impact. And I guess this is exactly what we are about to discover.
'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.

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 then applied the correction for the 2 remaining metrics practically dividing by 2 the alfa. 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 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 791500.

Duration vs. Exposure

I would divert 100% of the traffic, the number of days required would be 20.

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 because I wanted to check first how many metrics would prove significant. In case more metric would have been statistically significant and practically significant I would have considered using the correction to adjust the alfa for the usage of multiple metrics.

Recommendation

It seems that the only statistically and practically significant metric is the Gross Conversion. Though we used it to carry on our computation, 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 be active subscribers after 14 days by something among 1,19% and 2.91%, quite a meaningful figure. I would therefore not implement that change if my horizon is purely financial and is to get the 14 days conversion. The matter gets trickier if the goal is not focused on the next 14 days subscription. Nonetheless under a financial perspective it would not be convenient to launch the change because, at worst, those who would leave anyway, would pay for two weeks more. Other factors though might influence the final decision like marketing and branding issues or more complex considerations over the Customer Lifetime Value and operating costs.

Follow-Up Experiment

Once we agree on what ‘cancelling early in the course’ means in terms of time, we set the threshold to X, to keep track of the number of people who leave the course before the threshold. 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 before X. 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 a probability: Number of user IDs likely to drop off that actually drop off divided by the number of user IDs likely to drop off that actually drop off but are presented with the ‘challenge’ or ‘incentive’ button.

The analysis could then be carried on in a similar fashion to the one implemented in this submission.

One caveat: I’m aware of the fact that splitting the students in potentially dropping off or not would inject some further variability when conducting the experiment as the machine learning task would bring its own uncertainty when creating control and experiment groups (we might unluckily assign to control all the false positives chosen by our algorithm) the matter is statistical though and could surely be addressed with the proper numbering.