Experiment Design

Metric Choice

We selected following metrics as invariant:-

- **Number of cookies**
  This metric specifies number of unique users to visit the page. This is independent of our change in behavior on clicking a button, so this metric would remain same between control and experiment.

- **Number of clicks**
  This metric specifies number of clicks on “Start Free Trial” button. Again this is independent of our change, so this metric is also invariant.

- **Click-through-probability**
  It is just the normalized version of last metric “Number of clicks”. It is more robust metric as it remains valid if number of users are different in controlled and experiment samples. Following the argument of last metric, it is also a invariant metric for sanity check.

We selected following metrics as evaluation metrics:-

- **Gross conversion**
  It is number of user-ids to complete checkout and enroll in the free trial divided by number of unique cookies to click the "Start free trial" button.
  Now this result would depend upon user experience on clicking button. We are trying to dissuade users with inadequate dedication. This would results in less number of enrollments.
  Using it as evaluation metric, we want to check if there is any statistical or practical significant change in this metric. If there is a significant decrease, then our experiment is effective.

- **Net conversion**
  It is number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by the number of unique cookies to click the "Start free trial" button.
  There should not be a significant statistical or practical difference in Net conversion metric, if we want to launch our experiment. If there is a significant decrease in this metric, then we might be missing rightful candidates in our experiment.

We selected following metrics as neither as invariant or evaluation.

- **Number of user-ids**
  It would get affected by our testing. We didn’t include it as evaluation metric because it is not normalized with respect to number of pageviews or number of clicks. If number of pageviews increase or decrease, corresponding high or low quantity of this metric can not be attributed to our change.

- **Retention**
  It is number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout. We expect to get increase in this metric. In fact we included it in our initial choice of evaluation metric. But this metric was dropped due to requirement of long duration(4 months) for large number of pageviews needed for getting specified power and significance.

Measuring Standard Deviation

- **Gross Conversion**
  Standard Deviation = 0.0202

- **Net Conversion**
  Standard Deviation = 0.0156

Both of evaluation metric above is estimated analytically considering binomial distribution using baseline values given in quiz spreadsheet. If our assumption of well defined binomial distribution is not valid, then our analytical estimate won't be correct. We should calculate variance empirically in that case. For empirical calculation, we would need to do multiple A/A experiments, which won't be possible with given data and duration. We can also use bootstrapping, but we would need more granular data than what is provided.
Also for these metrics, cookie as unit of diversion is matching with unit of analysis (denominator) so that analytical variability is likely to match empirical variability.

Sizing

Number of Samples vs. Power

We will not use Bonferroni correction during analysis phase as we want to satisfy both of our hypothesis metrics. We use Bonferroni correction, when we can satisfy either one of all metrics as in that case net significance probability for type-1 error is high.
Number of pageviews needed to satisfy given significance (alpha) and power(beta) is **685325**.

Duration vs. Exposure

As nobody would get hurt because of our experiment and there are no privacy issues, we would run our experiment on entire traffic by using fraction of 1.0.
With full fraction, duration of experiment is just 18 days, which is within given constraints.

Experiment Analysis

Sanity Checks

We do sanity check on invariant matrices as decided above.

- **Number of cookies**

CI : (0.4988, 0.5012)
Observed Value : 0.5006

- **Number of clicks**
CI : (0.4959, 0.5041)
Observed Value : 0.5005

For both metrics above, we consider binomial distribution with probability of either going in controlled or experiment as 0.5. We calculated standard deviation using formula sqrt(p(1-p)/n). With approximation of normal distribution, confidence interval is calculated as 1.96*SE in both direction of 0.5. As our observed values lies in interval, so above metrics pass sanity check.

- **Click-through-probability**
CI : (-0.0013, 0.0013)
Observed Value: -0.00005

For this metric, we find the interval for difference between control and experiment. We calculate pooled standard deviation and as hypothesis of zero difference in rates, we find interval around 0. As our observed value lies within the interval, so this metric also pass sanity check.

As all our metrics pass sanity checks, we continue with analysis of our result.

Result Analysis

Effect Size Tests

Effect size of both of our evaluation metric is as below

- **Gross conversion**
CI : (-0.0291, -0.0120)
As zero doesn't lie between the interval, so difference between control and experiment group is statistically significant for this metric.
Practical significance threshold (dmin) for Gross conversion metric is 0.01. As both positive and negative of this value lies outside interval, so this metric is practically significant.

- **Net conversion**
CI : (-0.0116 ,0.0019)
As zero lies between the interval, so difference between control and experiment is not statistically significant here.
Practical significance threshold (dmin) for Net conversion metric is 0.0075. As negative of this value lie in above confidence interval, so this metric is not practically significant.

Sign Tests

Sign test of both of our evaluation metric is as below

- **Gross conversion**
  P-value: 0.0026
  We found that in only 4 out of 23 cases, experiment group has high Gross conversion rate than control group. As P-value of 0.0026 corresponding to this ratio is very low, so result of sign test for this metric is statistically significant.

- **Net conversion**
  P-value: 0.6776
  We found that in 10 out of 23 cases, experiment group has has high Net conversion rate that control group. As p-value of 0.6776 corresponding to this ratio is high, so result of sign test for this metric is not statistically significant.

Summary

As described in one of section above, we didn't use Bonferroni correction in this experiment as we want to satisfy both of our hypothesis metrics. We use Bonferroni correction, when we can satisfy either one out of all metrics as in that case overall significance probability for type-1 error is high. When we want to statistically satisfy all matrices, overall significance probability for type-1 error become low.
Result of sign test is matching with effect size tests. For Gross conversion, we found significant decrease. For net conversion, there is no significant difference.

Recommendation

We were able to find statistically and practically significant differences in Gross conversion rate to confirm that there is decrease in enrollment of students who were not dedicated enough.
We also don't find statistically significant difference between control and experiment for net conversion rate but the confidence interval does include the negative of the practical significance boundary thus it is possible that decrease in net conversion rate might matter to business. Hence we don't recommend to launch the experiment as rightful students might avoid enrollment of course.

Follow-Up Experiment

For reducing number of dropouts i.e. students who enroll but withdraw without payments, we can try for engagement during initial phase of course. Early engagement could include introductory videos by coaches and preemptive resources like FAQ for expected student doubts. This change can give head start to students and encourage them to continue the course.
Our hypothesis is that initial engagement would significantly increase number of paying enrolled students out of total enrollments.
Invariant metric for this experiment would be number of enrollments as our change would affect only once student is enrolled

Evaluation metric for this experiment would be retention defined as number of user-ids to remain enrolled past the 14-day boundary (and thus make at least one payment) divided by number of user-ids to complete checkout. (dmin=0.01).
Unit of diversion and analysis for this experiment would be user-id as user need to sign-in for enrolling in course.