P7: Design an A/B Test

The full version of the project: https://olgabelitskaya.github.io/P7_Design_an_A_B_Test_Overview.html

Experiment Design

Metric Choice

Invariant metrics (expected to be unchanged in the control and experimental groups):
1) number of cookies (cannot be affected by the experiment: users made a decision to visit the page before they were asked the question);
2) number of clicks (cannot be affected by the experiment: users clicked the button before they were asked the question);
3) click-through probability (cannot be affected by the experiment: it equals to the number of clicks divided by the number of cookies).

Evaluation metrics (expected to be different in the control and experimental groups):
1) gross conversion (can be affected by the experiment / can decrease: users could make a decision to enroll in the free trial in the experimental group less than in the control group because they did not plan to learn 5+ hours per week);
2) retention (can be affected by the experiment / can increase: enrolled users could be disappointed in the learning process less and make more payments in the experimental group than in the control group because they paid attention to studying 5+ hours per week);
3) net conversion (can be affected by the experiment / can decrease: users could enroll in the free trial less in the experimental group than in the control group, thus could decrease the number of people who paid).

The goals of the experiment in the practical meaning:

- the number of payments should not be decreased;
- the number of students who were disappointed and had not paid because they could not study enough time should be reduced.

The goals of the experiment in terms of our metrics:

- the gross conversion should significantly decrease;
- the retention should significantly increase;
- the net conversion should not decrease.

An important remark: the number of user-ids is neither a good invariant metric nor a good evaluation metric.
From one side, the new pop-up message is likely to decrease the total number of user-ids who enrolled in the free trial, so it is not invariant; from the other side it is not normalized, the number of visitors may be different between the experiment and control groups, so it is not good for evaluation.

Measuring Standard Deviation

Number of cookies = 5000
Number of clicks on "Start free trial" = 5000 × 0.08 = 400
Number of enrollments = 5000 × 0.08 × 0.20625 = 82.5

$$SD\ Gross\ conversion = \sqrt{\frac{p*(1-p)}{n}} = \sqrt{\frac{0.20625*(1-0.20625)}{400}} = 0.0202$$

$$SD\ Retention = \sqrt{\frac{p*(1-p)}{n}} = \sqrt{\frac{0.53*(1-0.53)}{82.5}} = 0.0549$$

$$SD\ Net\ conversion = \sqrt{\frac{p*(1-p)}{n}} = \sqrt{\frac{0.1093125*(1-0.1093125)}{400}} = 0.0156$$

I would like to expect the analytical variance is close to the empirical variance for the gross conversion and for the net conversion: the denominator for these two indicators is the number of clicks, which is also the unit of diversion.

And it would be useful to collect an empirical estimate of the variability for the retention: the unit of diversion was not used in this case, the empirical variance of the retention is more likely to be higher than the analytical variance.

Sizing

Number of Samples vs. Power

I have used the online calculator (References, N5) for calculating the sample sizes and have not chosen the largest.
Gross conversion: 2 × 25835 × 40000 ÷ 3200 = 645875 pageviews
Retention: 2 × 39115 × 40000 ÷ 660 = 4741212 pageviews
Net conversion: 2 × 27413 × 40000 ÷ 3200 = 685325 pageviews

I did not use the Bonferroni correction. Number of pageviews: 685325.

Duration vs. Exposure

Number of pageviews: 4741212 (for the experiment with the gross conversion, the retention, and the net conversion)
Days (100% of the traffic) = 4741212 ÷ 40000 = 118.5303

Number of pageviews: 685325
(for the experiment with the gross conversion and the net conversion)
Days (100% of the traffic) = 685325 ÷ 40000 = 17.133125
Days (60% of the traffic) = 685325 ÷ 40000 ÷ 0.6 = 28.5552083333

Number of pageviews: 685325 Fraction of traffic exposed: 0.6 Length of experiment: 29

Even the use of 100 percent of the traffic does not allow testing in full for all three metrics (the gross conversion, the retention, and the net conversion). It turns out that we need to experiment 118 days. Of course, it is too long for tasks staged in the project. Percent reduction in traffic will increase this interval. Hence, it is necessary to reduce the number of test metrics and choose only two of them: the gross conversion and the net conversion.

I should reconsider an earlier decision about 4741212 needed pageviews and setup the number of pageviews equal to 685325. For the experiment with the gross conversion and the net conversion, we can use the period 17 days with the 100% traffic level. This interval is much better, but it gives us the result too quickly The behavior of people in the field of education is quite difficult to analyze and trends in this area could be rarely detected in a short time period for 2-3 weeks. To slightly increase the time interval, we will set the percentage of used traffic at 60 (fraction = 0.6) and it gives us the number: 29 days.

It is also possible to take into consideration that the commercial risk is low: the site offers students to better plan and to evaluate their training time. The assumption of a sharp decrease in payments as a result of the pilot warning is devoid of practical foundation. For investments, the decisive factor is the first free period as a way to assess their own abilities, and it does not change during the experiment. The number of payments theoretically may slightly decrease due to the lower number of subscriptions, but this is only an assumption. Students who spend less than 5 hours a week are hardly able to complete the program successfully. In most cases, we are talking about reducing the waste of time for students and mentors.

And there is no risk in terms of privacy violation at all. Users do not enter any additional information about themselves in the process of the experiment except a little piece of information about planning.

It should be noted that the use of the fraction 0.6 does not affect the commercial interests seriously. From one side, the time interval and therefore the risk will increase. From another side, we keep the certain number of pageviews, therefore the experimental audience and the risk will not increase. It means our experiment could not practically be risky for the business.

Removing one indicator in the main part of the project, I want to analyze the experimental data in the last section on my own initiative a little bit wider and to determine what happens to the metrics "Retention" as the most interesting in terms of the psychological effect and the most unpredictable.

Experiment Analysis
Sanity Checks

Control group: Clicks = 28378 Pageviews = 345543
Experimental group: Clicks = 28325 Pageviews = 344660

Number of cookies:
Standard error SE = 0.000601840740294
Margin of error ME = 0.00117960785098
Confidential interval CI = (0.49882039214902313, 0.5011796078509769)
$\hat{p}$ = 0.500639666881 ∈ (0.49882039214902313, 0.5011796078509769) ✓

Number of clicks on “Start free trial":
Standard error SE = 0.0020997470797

Margin of error ME = 0.00411550427621
Confidential interval CI = (0.49588449572378945, 0.5041155042762105)
$\hat{p}$ = 0.500467347407 ∈ (0.49588449572378945, 0.5041155042762105) ✓

Click-through-probability on "Start free trial":
Pooled probability p_pool = 0.0821540908979
Standard error SE = 0.000661060815639
Margin of error ME = 0.00129567919865
Difference $\hat{d}$ = 5.66270915869e-05
Confidential interval CI = (-0.0012956791986518956, 0.0012956791986518956)
$\hat{d}$ ∈ (-0.0012956791986518956, 0.0012956791986518956) ✓

Number of cookies
Lower bound = 0.4988; Upper bound = 0.5012; Observed = 0.5006; Passes = Yes

Number of clicks on “Start free trial"
Lower bound = 0.4959; Upper bound = 0.5041; Observed = 0.5005; Passes = Yes

Click-through-probability on "Start free trial" (Difference between the control and experimental groups)
Lower bound = -0.0013; Upper bound = 0.0013; Observed = 0.0001; Passes = Yes

All invariant metrics have stood the test successfully. This is a very predictable result. Selecting these metrics was based on the lack of experiment influence.

Result Analysis

Effect Size Tests

For our evaluation metrics, I gave a 95% confidence interval around the difference between the experiment and control groups and indicated whether each metric was statistically and practically significant.

Gross conversion
Pooled probability p_pool = 0.208607067404
Standard error SE = 0.00437167538523
Margin of error ME = 0.00856848375504
Difference d = -0.0205548745804
Confidential interval CI = (-0.0291233583354044, -0.01198639082531873)
(-0.01, 0, 0.01) ∉ (-0.0291233583354044, -0.01198639082531873)

Net conversion
Pooled probability p_pool = 0.115127485312
Standard error SE = 0.00343413351293
Margin of error ME = 0.00673090168535
Difference d = -0.00487372267454
Confidential interval CI = (-0.011604624359891718, 0.001857179010803383)
0 ∈ (-0.011604624359891718, 0.001857179010803383) ; dmin = -0.0075 ∈ (-0.011604624359891718, 0.001857179010803383)

I did not use the Bonferroni correction.

Gross conversion (Difference between the control and experimental groups)
Lower bound = -0.0291; Upper bound = -0.0120; Statistical significance = ✓; Practical significance = ✓

Net conversion (Difference between the control and experimental groups)
Lower bound = -0.0116; Upper bound = 0.0019; Statistical significance = ✘; Practical significance = ✘

Sign Tests

Gross conversion: success = 4 total = 23
Net conversion: success = 10 total = 23

I have used the online calculator (References, N6) for the sign tests.

I did not use the Bonferroni correction.

Gross conversion: p-value = 0.0026; Statistical significance = ✓

Net conversion: p-value = 0.6776; Statistical significance = ✘

Summary

Eventually, the effective size and sign tests show that the site change would statistically significantly reduce the gross conversion, but would not affect the net conversion in a statistically significant way. The effect size test states this in the practical meaning also.

We have measured two metrics in one experiment. Applying the Bonferroni correction means that the a-level for each hypothesis will be 2.5 % instead of 5% and confidential intervals will be significantly wider. It is too conservative for some reasons.

The use of the Bonferroni correction would really be needed if we test several metrics in one experiment and expect that at least one metrics will demonstrate the statistically significant change. In the set of metrics, this matching only for one indicator can be an absolutely random event, therefore the experiment will have a false positive result. It means we should increase the confidential intervals to avoid this situation and apply the Bonferroni correction.

But in the case of our experiment, we expect two metrics will have matched our criteria at the same time to proceed with the launch. It's a very strong condition without any correction. The positive results will be more likely to occur not by chance. Therefore, the Bonferroni correction could be the cause to approve the wrong null hypothesis and we should not use it this time.

Also, our metrics have a strong relationship between each other. If we know the outcome of one test of a difference between the control and experimental groups on one metrics, it would be easy to predict and to find the outcome of the other tests on the other metrics. It's absolutely natural to expect their behavior will be similar simultaneously.

Recommendation

The recommendation is not to launch the experiment change because the negative results have outweighed the positive ones.

Positive results of the experiment.
- The difference for the gross conversion is practically significant and negative. This is a good sign: the Udacity team can lower costs by a number of trial signups.
- The difference for the net conversion is not statistically significant. It means the absence of serious financial losses.

Negative results of the experiment:
- The interval for the difference in the case of the net conversion includes negative numbers. Therefore, the team has a risk to decrease incomes.
- We have not gathered enough data to draw conclusions about the retention and because of this we can not evaluate correctly the difference between the control and experimental groups for a number of students who were disappointed in studying during the free period. Consequently, we do not know enrolled users would be disappointed in the learning process less and make more payments or would not.

Follow-Up Experiment

Before scheduling the follow-up experiment, I would like to analyze the available data for the evaluation metrics "Retention".

Effect size test
Pooled probability p_pool = 0.551886792453
Standard error SE = 0.0117297800914
Margin of error ME = 0.0229903689791
Difference d = 0.0310948047071
Confidential interval CI = (0.008104435728019967, 0.05408517368626556)
$0 \notin (0.008104435728019967, 0.05408517368626556)$; dmin = $0.01 \in (0.008104435728019967, 0.05408517368626556)$
Retention (Difference between the control and experimental groups)
Lower bound = 0.0081; Upper bound = 0.0541; Statistical significance = ✓; Practical significance = ✘

Sign test
Retention: success = 13 total = 23
Retention: p-value = 0.6776; Statistical significance = ✘

Gross conversion difference	Retention difference	Net conversion difference
Median: -0.0247583431053	Median: 0.0232108317215	Median: -0.00902785253995
Mean: -0.0207845820293	Mean: 0.0333425074664	Mean: -0.00489685698981

Statistical significance of the difference between the control and experimental groups was checked by using the values of the mean (effect size tests) and the median (sign tests). As we can see, these values vary considerably for the retention and this is the cause that the test results also differ.

Even a cursory analysis of insufficient data in a certain way confirms our intuitive assumptions about the behavior of this metric. The ratio of payments to enrollments tends to increase in the presence of the experimental warning. However, it is not possible to confirm this with sufficient certainty in the borders of this experiment and it is necessary to redesign the research.

I would suggest these possible changes to the proposed experiment.

- Extend the duration of the experiment up to 2 months with constant monitoring of incomes for avoiding financial risks. If the decline in revenues becomes out of the certain limits, the study should be stopped immediately.
- Measure all three evaluation metrics (the gross conversion, the retention, and the net conversion) for the 100% level of traffic because the audience of this site is very different in education, age, nationality, and other characteristics. Any reduction in the percentage of participants can significantly distort the results.
- Replace the visualization message by the video with an explanation of successful learning strategies based on statistics of the particular site or by the input test for the course level recommendations exactly for this user.
- In order to avoid the negative psychological effect or cut the extremely talented part of the audience which is able to pass the course without spending a lot of time to study, all changes should be only informative and recommendatory.
- Recommendations in the videos or leveled tests should have the most practical character that is suitable for this course.
- The length of free trial period is unchanged.

I think the effect will be more detectable.

Now we can begin to define the technical details of the experiment. Selecting the unit of diversion, and invariant and evaluation metrics was quite reasonable. It is easy to obtain measurement results for decision making without the high level of costs or risks. Accordingly, I propose to leave them unchanged.

- Unit of diversion: the cookie.
- Invariant metrics: the number of cookies, the number of clicks, and the click-through-probability.
- Evaluation metrics: the gross conversion, the retention, and the net conversion.

The hypothesizes about the behavior of our metrics are also stayed the same:
- the gross conversion should significantly decrease;
- the retention should significantly increase;
- the net conversion should not decrease.

I expect the overcoming the negative results of the previous experiment and detecting the tendencies for all evaluation metrics.