P7: A/B Testing Final Project: Free Trial Screener

Lekhraj Sharma
Data Analyst Nanodegree
Sept 2016

This template can be used to organize your answers to the final project. Items that should be copied from your answers to the quizzes should be given in blue.

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.)

- **Number of cookies** (*That is, number of unique cookies to view the course overview page, d_{min}=3000*) : ***Invariant***
  This is a good invariant metric by design of our experiment. We must split the incoming traffic by unique cookies equally in two groups (control and experiment) to allow for proper comparison of our changes down the funnel. Therefore barring some goof up we should observe about the same total number of cookies being diverted to control and experiment groups. So by design of experiment, we want this to be invariant.
- **Number of user-ids** (*That is, number of users who enroll in the free trial, d_{min}=50*) : ***Neither Invariant nor Evaluation Metric***
  By nature of proposed change, we are expecting the enrollment to be impacted. We are asking users to reconsider their decision to enroll, therefore it is expected that in general we may see less number of enrollment. So this can **not** be invariant. The reason we do not want it to be an evaluation metric is because we can not draw good conclusions from change in absolute number of enrollment. We have better metrics like Gross Conversion which will help us to tack change in enrollment conversion.
- **Number of clicks** (That is, number of unique cookies to click the "Start free trial" button (which happens before the free trial screener is trigger, d_{min}=240): ***Invariant***
  We are sending same amount of traffic to top of our funnel (pageviews controlled by invariant number of cookies). We are not making any changes before users click “Start Free Trial” button. Therefore we expect number of users clicking “Start Free Trial” not to change because of our proposed change. So by design this metric should remain invariant unless there is some goof up, which we better know!
- **Click-through-probability** (That is, number of unique cookies to click the "Start free trial" button divided by number of unique cookies to view the course overview page, d_{min}=0.01) : ***Invariant***
  As both “Number of Cookies” and “Number of clicks” both are invariant, therefore this metric which is ratio of these two invariant metrics is also expected to remain invariant.
- **Gross conversion** (That 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, d_{min}= 0.01) : ***Evaluation Metric***
  By nature of proposed change, we are expecting the enrollment to be impacted. We are asking users to reconsider their decision to enroll, therefore it is expected that in general we may see less number of enrollment. While we are expecting the click-through probability to remain same, we do expect our enrollment conversion (i.e. Gross Conversion) to go down. Measuring the change in enrollment conversion allows us to evaluate how much impact did this change makes to first part of the clicks-enrollment-paid funnel.
- **Retention** (That 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, d_{min}=0.01) : ***Evaluation Metric***

Improving retention of students who become paid users after 14 days of enrollment in free trial is key business objective of proposed experiment. Measuring the change in enrollment to paid conversion allows us to evaluate how much impact did our change makes to second part of the clicks-enrollment-paid funnel. Therefore we should try to measure this if we can.

- **Net conversion** (That 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, d_{min}= 0.0075) : ***Evaluation Metric***
  The main business objective of proposed experiment is to improve the overall student experience. Hypothesis is that if we have more committed students, they will progress better and even improve coaches' capacity to support students, therefore improve their chance of successfully completing the course. Further hypothesis is that this change in funneling more committed students will not result in significant reduction in paid users (i.e. business does not suffer). Measuring the change in clicks to paid conversion allows us to see the impact on overall clicks-enrollment-paid funnel. Therefore this is a **key** evaluation metric for us. We need to find out if our experiment makes any significant impact to this metric. Our business objective is **not** to see this metric significantly reduce.

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.

Explanation on why a metric was chosen to be invariant or evaluation metric is already given above.

On result front, our business goal is to improve overall student experience (offer them better coaching support, improve their chances of successfully completing their course, etc.) by getting more committed students in the funnel (i.e. filtering out less committed students, which is likely to result in reduction in gross conversion) without significantly reducing the number of students who complete the course (i.e. who pay, so we do not want reduction in net conversion).

We have multiple evaluation metrics (Gross Conversion, Retention and Net Conversion). So we need to **match expectations on all of them** to launch. The key metric to watch would be "Net Conversion" since it encompasses both "Gross Conversion" and "Retention" (Net Conversion = Gross Conversion * Retention). Since we are expecting gross conversion to decrease, our retention rate may need to increase to support our goal of no decrease in net conversion. In order to launch, we would need to match all of our expectations (a decrease in gross conversion, increase in retention and a no decrease in the net conversion).

Measuring Standard Deviation

List the standard deviation of each of your evaluation metrics. (These should be the answers from the "Calculating standard deviation" quiz.)

- **Gross conversion**: 0.0202
- **Retention**: 0.0549
- **Net conversion**: 0.0156

Following table outlines the step used to calculate these standard deviation with sample size of 5000 pageviews.

Unique cookies to view page per day:	5000	
Unique cookies to click "Start free trial" per day:	400	
Enrollments per day:	82.5	
Click-through-probability on "Start free trial":	0.08	STDDEV=sqrt(p*(1-p)/N)
Probability of enrolling, given click:	0.20625	0.0202
Probability of payment, given enroll:	0.53	0.0549
Probability of payment, given click	0.1093125	0.0156
Paid per day:	43.725	

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.

Since our metrics are probabilities that follow binomial distribution, and our baseline sample size of 5000 is large enough, our analytical estimation of standard deviation (sqrt(p * (1-p) / N)) should be reasonably accurate. Keeping in mind that our unit of diversion is cookie, below is specific analysis of variability per metric:

- **Gross conversion** (That 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)
  Here our unit of analysis (denominator of our metric) is cookie, which is same as unit of diversion. This means we should have less variability implying empirical estimate is likely to be closer to calculated analytical estimate of 0.0202.
- **Retention** (That 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)
  Here our unit of analysis (denominator of our metric) is user-ids, which is while not exactly same as unit of diversion (cookie), but they are quite comparable as most part these unique cookies will correspond to unique user-ids (although some user may end up trying multiple times). Plus our unit of diversion (cookies) is larger than unit of analysis (user-ids). This means we should have lower variability implying empirical estimate is more likely to be closer to calculated analytical estimate of 0.0549.
- **Net conversion** (That 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)
  Here our unit of analysis (denominator of our metric) is cookie, which is same as unit of diversion. This means we should have less variability implying empirical estimate is likely to be closer to calculated analytical estimate of 0.0156.

While we could always improve our estimate with more data if we have time, looking at low variability as described above, we should be good to go with our analytical estimates.

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.)

Will **not** use Bonferroni correction (see summary for further explanation).

Number of Pageviews: **685325**

Used online calculator available at http://www.evanmiller.org/ab-testing/sample-size.html to calculate sample size.

	Given Baseline Probability	Given d_{min}	Sample Size per group from Online Calculator (Alpha = 0.05, Beta = 0.2)	Number of Page Views per group	Total pageviews (2 * pageviews for each group)	
Gross Conversion: Probability of enrolling, given click	0.20625	0.01	25835 clicks	25835/0.08 = 322938	645875	
Retention: Probability of payment, given enroll	0.53	0.01	39115 enrolled	39115/0.20625/0.08 = 2370606	4741212	Too large, drop this eval metric
Net Conversion: Probability of payment, given click	0.1093125	0.0075	27413 clicks	27413/0.08 = 342663	**685325**	Larger of rest

Dropped “Retention” from list of evaluation metric chosen before since it was requiring around 4.7 million pageviews, which will be not be practical looking at daily traffic of forty thousand pagviews on the site (will require us to run experiment for months and months!). After dropping “Retention”, chose the 685325 as sample size since it was larger of remaining two sizes to give adequate power to remaining two evaluation metrics.

Final list of Evaluation Metric(s):

- Gross Conversion
- Net Conversion

Updated results expectation in order to launch

So now in order to launch, we would need **both** of these metrics to match our expectations (a decrease in gross conversion and a no decrease in the net conversion).

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.)

Final sample size	**685325**	Total number of pagviews needed across both control and experiment group
Fraction of Traffic	**0.5**	Will divert 50% of traffic. Effectively 25% (experiment group) would see the change.
Number of Days needed to run the experiment (Duration)	685325/40000/0.5 = **35**	Since given data is for 37 days, we have enough power in our tests.

Give your reasoning for the faction you chose to divert. How risky do you think this experiment would be for Udacity?

Change does not seem to be super risky from technical perspective. Looking at low technical risk and no real privacy issues or handling of any sensitive data, we could divert the entire traffic. But we may still not want broader exposure without first understanding the impact on a smaller population. We typically should reduce the friction between user's ability to proceed in our funnel. Since user has already made up his mind about starting free trial, this change of asking him to reconsider while with a good intent can still be a drag in user wanting to move ahead. It is better that we first try this change on smaller population and understand its impact and any other unforeseen side effects before rolling out to broader population.

For adequate powered sample size of 685325 and looking at daily traffic of around 40k, we need to divert at least 50% of our traffic. Diverting less than that would extend our duration way beyond 35 days, making it unreasonably long and not practical. For example with only 25% traffic diverted to both groups, it will need 70 days to run this experiment, a way too long! With 50% traffic diverted, only 25% of our visitors to site would see the change (experiment group). Diverting 50% of traffic is a good tradeoff between risks of exposing unproven change to larger population and enough sample size needed for sufficient power to ensure we can rely on our results.

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.)

- **Number of cookies**: 0.4988 (lower bound), 0.5012 (upper bound), 0.5006 (observed), **Passes**
- **Number of clicks**: 0.4959 (lower bound), 0.5041 (upper bound), 0.5005 (observed), **Passes**
- **Click-through-probability**: 0.0812 (lower bound), 0.0830 (upper bound), 0.0821 (observed), **Passes**

Following table outlines details of how we went about doing these sanity checks. For simple count such as number of cookies and number clicks, we treated them as binomial test with even split. For click-through-probability, we first calculated standard error using pooled proportions method treating our baseline number as empirical estimate. This we than used to calculate a confidence interval and check whether our observed value (experiment) lies within this interval. As all our observed values lie within their corresponding confidence intervals, we can conclude that change is not statistically significant. Therefore our invariant metrics have remained statistically invariant, so our sanity check passes for all of them.

	p (split evenly)	control	experiment	p (control)	STDERR	m=1.96*STDERR	Interval. Min	Interval. Max	Is p(control) in Interval?
Number of cookies	0.5	345543	344660	0.5006	0.0006018407403	0.001179607851	0.4988	0.5012	YES
Number of Clicks	0.5	28378	28325	0.5005	0.00209974708	0.004115504276	0.4959	0.5041	YES
	Given empirical prob	STDERR_emp	Calculated prob (experiment)	Calculated prob (Control)	Calculated STDERR from portions	Delta(m) = 1.96 * STDERR	Interval. Min	Interval. Max	Is p(experiment) in interval?
Click-through-prob	0.08	0.001356465997	0.08218244067	0.0821258	0.000461812665	0.0009051528234	0.08122	0.08303	YES

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.***

Good news is that all our sanity checks passed for all of our invariant metrics. So we now have better confidence about our test setup, where as expected our invariants did not significantly change. Therefore we can now focus on analysing the real change in evaluation metrics.

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.)

Did **not** use Bonferroni correction (see summary for further explanation).

- **Gross Conversion**: -0.0291 (lower bound), -0.0120 (upper bound), **Statistically and Practically Significant**
- **Net Conversion**: -0.0016 (lower bound), 0.0019 (upper bound), **Statistically and Practically NOT Significant**

Following table outlines details of how we went about doing these effect size tests. We did not use Bonferroni correction as our evaluation metric are somewhat correlated and it would have meant being too conservative in measuring the change (see more about it in summary section).

	Clicks (Control)	Enrolled or Paid (Control)	Clicks (Experiment)	Enrolled or Paid (Experiment)	d_{min}	p (pool)	d = p_exp - p_cnt	STDERR(pool)	m=1.96*STDERR	Interval. Min	Interval. Max	Statistically Significant (interval does not contain zero?)	Practically Significant (interval does not contain d_{min} ?)
Gross Conversion	17293	3785	17260	3423	0.01	0.2086070674	-0.0205549	0.004371675385	0.008568483755	-0.0291	-0.0120	TRUE	TRUE
Net Conversion	17293	2033	17260	1945	0.0075	0.1151274853	-0.0048737	0.003434133513	0.006730901685	-0.0116	0.0019	FALSE	FALSE

Used pooled probability and proportions to calculate pooled standard error. This was then used to calculate a 95% confidence interval around difference between control and experiment groups. For group conversion metric, this interval neither contains zero or our practical significance boundary of given d_{min} (0.01). Therefore group conversion evaluation metric change is both statistically and practically significant. Whereas for net conversion metric, this interval contains zero as well as minimum of change is not larger than our practical significance boundary of given d_{min} (0.0075). Therefore net conversion evaluation metric change is neither statistically nor practically significant.

Sign Tests

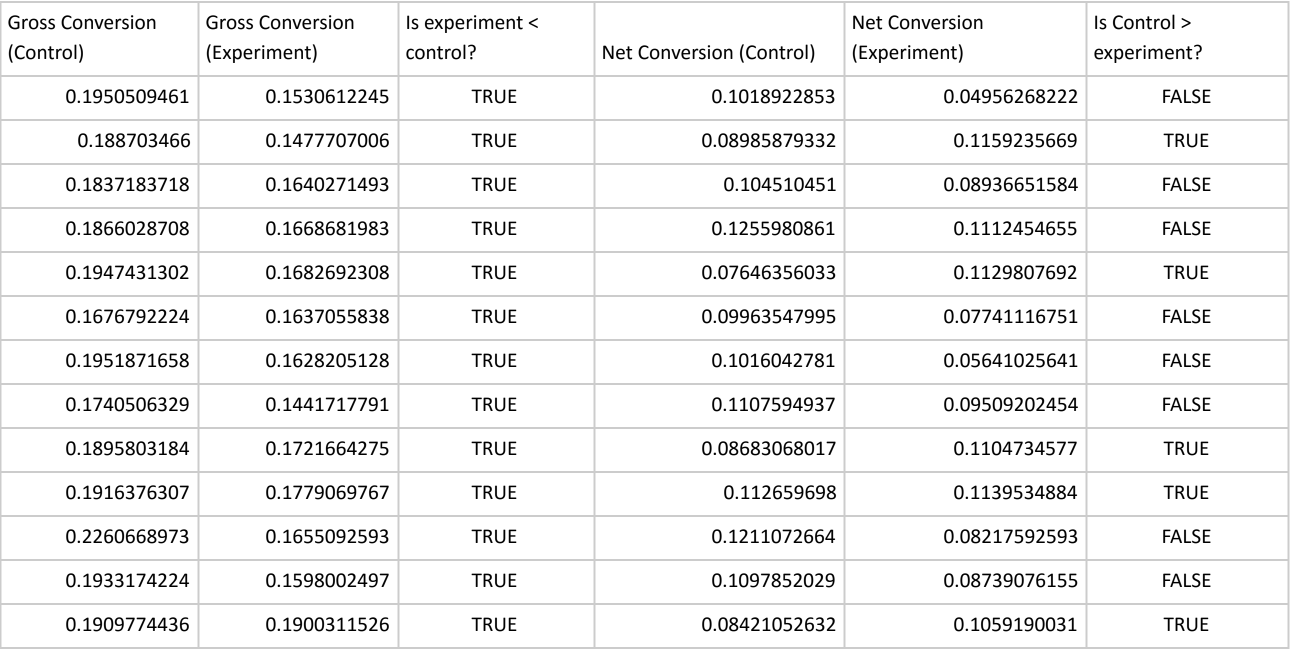
For each of your evaluation metrics, do a sign test using the day-by-day data, and report the p-value of the sign test and whether the result is statistically significant. (These should be the answers from the "Sign Tests" quiz.)

Did **not** use Bonferroni correction (see summary for further explanation).

- **Gross Conversion**: 0.0026 (p-value), **Statistically Significant**
- **Net Conversion**: 0.6776 (p-value), **Statistically NOT Significant**

Used online calculator available at http://graphpad.com/quickcalcs/binomial1.cfm to calculate probability.

Following table outlines the detailed mechanics of how these sign tests were done. Only considered first 23 days for which enrollment and paid data was available (since payment only happens after 14 days trial). For each day, calculated the conversion ratio (p=X/N) both for gross conversion (enrollments/clicks) and net conversion (payments/clicks) for both control and experiment groups. For sign tests on gross conversion metric, counted how many times the daily conversion of experiment group was larger than control group (since we know that normal trend in enrollment is downwards after the change). This turned out be 19 times out of 23 days. For net conversion, counted how many times experiment group daily conversion was larger than control group (we are not expecting a significant reduction here). This turned out to be 10 times out of 23 days. With these success and total trial numbers (19 out of 23 and 10 out of 23) with 0.5 probability of success, used online binomial test calculator to calculate the probability of these successes.



Gross Conversion (Control)	Gross Conversion (Experiment)	Is experiment < control?	Net Conversion (Control)	Net Conversion (Experiment)	Is Control > experiment?
0.1950509461	0.1530612245	TRUE	0.1018922853	0.04956268222	FALSE
0.188703466	0.1477707006	TRUE	0.08985879332	0.1159235669	TRUE
0.1837183718	0.1640271493	TRUE	0.104510451	0.08936651584	FALSE
0.1866028708	0.1668681983	TRUE	0.1255980861	0.1112454655	FALSE
0.1947431302	0.1682692308	TRUE	0.07646356033	0.1129807692	TRUE
0.1676792224	0.1637055838	TRUE	0.09963547995	0.07741116751	FALSE
0.1951871658	0.1628205128	TRUE	0.1016042781	0.05641025641	FALSE
0.1740506329	0.1441717791	TRUE	0.1107594937	0.09509202454	FALSE
0.1895803184	0.1721664275	TRUE	0.08683068017	0.1104734577	TRUE
0.1916376307	0.1779069767	TRUE	0.112659698	0.1139534884	TRUE
0.2260668973	0.1655092593	TRUE	0.1211072664	0.08217592593	FALSE
0.1933174224	0.1598002497	TRUE	0.1097852029	0.08739076155	FALSE
0.1909774436	0.1900311526	TRUE	0.08421052632	0.1059190031	TRUE

0.3268945022	0.2783357245	TRUE	0.1812778603	0.1348637016	FALSE
0.2547033285	0.1898355755	TRUE	0.1852387844	0.1210762332	FALSE
0.2274011299	0.2207792208	TRUE	0.1468926554	0.1457431457	FALSE
0.3069828722	0.2762645914	TRUE	0.163372859	0.1543450065	FALSE
0.2092391304	0.2201086957	FALSE	0.1236413043	0.1630434783	TRUE
0.2652232747	0.2764786795	FALSE	0.1163734777	0.1320495186	TRUE
0.227520436	0.2843406593	FALSE	0.1021798365	0.09203296703	FALSE
0.2464589235	0.2520775623	FALSE	0.1430594901	0.1703601108	TRUE
0.2290748899	0.2043165468	TRUE	0.1365638767	0.1438848921	TRUE
0.2972582973	0.2513812155	TRUE	0.09668109668	0.1422651934	TRUE
Total counts	23	19		23	10
	p (Gross)	Is p < Alpha?	Alpha	p (Net)	Is p < Alpha?
Calculated Probability	0.0026	TRUE	0.05	0.6776	FALSE

For gross conversion, resultant probability of 0.0026 was less than our alpha of 0.05. Which meant gross conversion result is statistically significant. Whereas for net conversion, resultant probability of 0.6776 was larger than our alpha of 0.05. Which meant net conversion result is NOT statistically significant. Both these results agree with our earlier hypothesis tests as mentioned in effect size section.

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.

Did **not** use Bonferroni correction since we need **both** evaluation metric to match our expectations (we look for a decrease in gross conversion and for a no decrease in the net conversion). We may fail to launch if any one of these two metrics does not meet our expectation. We risk not to launch if at least one metric (out of 2) fail to reject the null, when the null is not the true effect (Type II error). Bonferroni correction is not designed to reduce type-II error risks. On the hand if we were launching with any one metric of our multiple metrics meeting our expectation (which we are not) than out of multiple metrics the risk that just one rejects the null by pure chance (Type I error), would be very high. Bonferroni is designed to reduce this type of risk. In our case of needing both metrics to meet our expectations, Bonferroni correction is not appropriate. In fact applying Bonferroni correction in our case may be counter-productive since its minimisation of type-I error is likely to come at the expense of increased type-II error, which is what we need to control in our case of needing both metrics to match our expectations!

Our hypothesis tests and sign tests supported each other, so no discrepancy here. Therefore we saw that both using our confidence interval for evaluation metric and sign tests probability, gross conversion was statistically and practically significant but net conversion was not. As expected our enrollment came down significantly (Gross Conversion) but it did not result in any significantly better or worse conversion in terms of students who paid (Net Conversion).

Recommendation

Make a recommendation and briefly describe your reasoning.

Would **not** recommend going ahead with the proposed change. Our proposed change, asking users to re-consider opting for free access to course material if they may not be willing to put reasonable hours, while did bring down significantly the enrollment as expected (since we are now asking them to reconsider even after they opted to go ahead for free trial) but we are not able to establish that there is really no significant decrease in net conversion. While net conversion metric change not being statistically and practically significant may give us hope of no change in net conversion but we can also see that our confidence interval (-0.0116 to 0.0019) does include negative of our practical significance boundary (-0.0075). This implies there is a possibility that our net conversion may have gone down below our practical business expectation. This may not be the risk we are willing to live with since we do not want our net conversion to decrease. Therefore we may not be sure of meeting our expectation of no decrease in net conversion. Since we are **not** meeting **both** the expectations (decrease in gross conversion and no decrease in net conversion), we should **not launch**.

May be this change does help in decreasing the frustration level of enrolled population when they discover that they are not able to commit the required time and not make as much progress towards their goal. Therefore happiness index for enrolled students may go up who have been asked to seriously reconsider their commitment versus their goal of truly acquiring a new skill. This in long term may indeed result in more successful users. There may be other ways for us to improve overall happiness of students, for example improving quality of instructors and coaches, offering more personalized coaching, offering prep courses and guided exercises for those students who may be struggling, improving credibility/marketability of programs in industry, etc. We may be better off trying some of these changes to improve overall student experience.

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 may want to research main reasons students may be cancelling out early. While it could be a time commitment issues (as per our earlier experiment hypothesis) it could also be either frustration of not making desired progress or not getting good vibes about marketability of the course versus the effort they need to put in or just the pace of course is not matching their current understanding or something else.

Brief Description of follow-up experiment: Guided Exercise Site

Will like to select the reason of pace of course not matching their current level of understanding as the cause to go after for early cancellation. These users may be getting frustrated with not able to get timely help or guidance on concepts they are struggling to master. We could target this struggling population (low engagement) with a combination of personalized coaching, links to detailed answers to most common difficulties faced by past students, set of detailed guided examples/exercises which clarify most common applications of concepts being taught, a smart chatbot which answers your questions, etc. For our experiment we could pick sending them to another prep site which has easy to search and use set of detailed (step-by-step) additional exercises with solution for one to get out of their struggling period. We could do this as AB tests where we direct 50% of our chosen struggling population to this step-by-step guided exercise site and see if makes significant difference in reducing our early cancellation rate.

Hypothesis: Struggling enrolled students in trial period who are shown the guided exercise site will have significantly lower cancellation rate than others who are not.

Metric: Cancellation Rate of struggling enrolled students in their trial period. It can be measured as ratio of students who end up not paying divided by number of struggling students (already enrolled)

Population: Struggling students with low engagement scores (may be a machine learning based predictive score which identifies low engagement students. This could be combination of number of attempts in quizzes including not attempting at all, average number of days to progress in a course, large gaps in visiting course site, etc.). Or may be just pick-up a course having high historical early cancellation rate.

Unit of Diversion: User-id of struggling enrolled students. Since we are targeting already enrolled students but who may cancel in their trial period.

If we observe significant upside in this follow-up tests, we could see how this applies to reducing cancellation rate for already paid students but who later on cancel early without finishing the course.

References

1. AB Testing course at www.udacity.com for Data Abalyst Nanodegree
2. Sample size calculator at http://www.evanmiller.org/ab-testing/sample-size.html
3. Sign test calculator at http://graphpad.com/quickcalcs/binomial1.cfm
4. Statistical Inference course in R at https://github.com/swirldev/swirl_courses/tree/master/Statistical_Inference
5. https://en.wikipedia.org/wiki/A/B_testing
6. http://conversionxl.com/ab-testing-statistics/
7. http://onlinelibrary.wiley.com/doi/10.1111/opo.12131/full
8. Book: Data Science From Scratch by Joel Grus
9. Book: Statistics in Nutshell by Sarah Boslaugh