DAND P7: Design an A/B Test - Final Report.

Experiment Design

Metric Choice

List of Invariant Metrics:

- Number of cookies, It is an invariant vs. evaluation metric because:
  the experiment itself impose no statistical influence to this number, on the contrary this is an independent variable to the experiment in test. In addition, the number of unique cookies or page views should be kept close enough to controlled and experiment group.
- Number of clicks, It is an invariant vs. evaluation metric because:
  Since the click is an event prior to experimental manipulation, it will not have direct implications on the efficacy of the experiment, also if the Number of cookies are split evenly between control and experiment groups, the Number of clicks should also be close to even between groups when the split of population is in random.
- Click-through-probability
  For the same reason that this has no direct implication on the efficacy of the xperiment, it is a probability representation of above two events prior to experiment exposure. Even though this looks redundant (to the Number of clicks), it is still independent to the experiment, so I look it as an invariant vs. evaluation metrics to the test.

List of Evaluation Metrics:

- Gross conversion, It is an evaluation vs. invariant metrics because:
  It is the **Number of Enrolled User-ids divided by the Number of clicks**, and the Number of Enrolled User-ids is directly influenced by the implication of the experiment. Expectedly the Number of Enrolled User-ids would be just different btw. Control and experiment groups.So evaluate this metric would derive the efficacy of the experiment.
- Retention (optional, later know it is too long to run, also its unit of analysis is not cookie):
  It is the **Number of Sustained User-ids (who converted to paid user) divided by the Number of Enrolled User-ids**, both numbers have direct implication from the efficacy of the experiment, in that (hypothesis) “Start free trial” click users in experiment group will be presented with an extra questionnaire for committable time to the program, so the Number of Enrolled User-ids may be some reduced, but the retention rate should be higher for experiment group, due to enrolled users may be more determined there. Evaluate this metric may derive the efficacy of the experiment.
- Net conversion, it is an evaluation vs. invariant metrics because:
  It is the **Number of Sustained User-ids divided by the the Number of clicks**, and the Number of Sustained User-ids is indirectly influenced by the implication of experiment.

‘**Number of user-ids**’ is neither selected as invariant nor evaluation metrics, due to:

- It is not invariant as it is NOT independent to the test, i.e. experiment will have direct implication to this number.
- It is not one of evaluation metrics, because unit of diversion of experiment is a cookie.
- Also Number of user-ids along is not the best to present the experiment efficacy.

According to the design and hypothesis, I look forward to these **to launch the experiment**:

- Some **lower Gross conversion** with significance
- Some **Higher Retention** if possible
- **Net conversion** is at least not significant lower

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

In this calculation, we use the basic formula: $SD = \sqrt{\frac{p(1-p)}{N}}$

For each evaluation metric, p and N is given in baseline values in 40,000 unique pageview cookies per day. In this question, we are asked to calculate SD for each metric given 5,000 pageview, so we just need to scale SD of each metric (from baseline values) by $\sqrt{8}$ provided p for each is the same.

To calculate the standard error empirically, we need multiple A/A experiments, where we don't calculate the standard deviation of individual data points (days) within one experiment, instead we calculate the standard deviation of multiple experiments run over the same time period. Here we don't have granular data than what was provided in order to do bootstrapping, also the metrics that we are dealing with follow a binomial distribution, and the binomial distribution is well approximated by the normal distribution, so it is appropriate to use analytic estimates and think it be comparable to the empirical variability. Going empirically is a bit of a waste, which works well in cases no assumption of distributions can be made.

On the other hand, since the denominators (unit of analysis) of our metrics (except for 'Retention') are 'unique clicks', essentially it is the same as unit of diversion (cookie), so we can use the analytical estimate along, which is likely to match the empirical variability.

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.)

Using online resource: http://www.evanmiller.org/ab-testing/sample-size.html

Firstly I did choose **No Bonferroni correction** during my analysis, provided that it is over conservative in adjusting significant level to minimize false positive, but I'm look forward to all metrics with a significant change observed, so use Bonferroni correction seems unnecessary!

Secondly use common parameters ($\alpha = 0.05,\ \beta = 0.2$). For 3 evaluation metrics earlier selected:

- Gross conversion (Enrolls / Clicks)

  Baseline conversion rate = 20.625%

  Minimum Detectable Effect = d_{min} = 1%

⇒ Sample Size of Clicks = 25,835
⇒ Number of Pageviews (in each group) = 25,835 ÷ (3200 ÷ 40000) = 322,938
⇒ Total Pageviews (both control and experiment group) = 322,938 × 2 = 645,875

- Retention (Paid-enrolls / Enrolls)

  Baseline conversion rate = 53%

  Minimum Detectable Effect = d_{min} = 1%

  ⇒ Sample Size of Enrolls = 39,115
  ⇒ Number of Pageviews (in each group) = 39,115 ÷ (660 ÷ 40000) = 2,370,606
  ⇒ Total Pageviews (both control & experiment group) = 2,370,606 × 2 = 4,741,212

  ***N.B.** Since total pageviews required for 'Retention' is too big, and test is too long to run, there is NO real value to keep it in evaluation metrics. Forward I kept it an optional calculation only!

- Net conversion (Paid-enrolls / Clicks)

  Baseline conversion rate = 10.931%

  Minimum Detectable Effect = d_{min} = 0.75%

  ⇒ Sample Size of Clicks = 27,413
  ⇒ Number of Pageviews (in each group) = 27,413 ÷ (3200 ÷ 40000) = 342,663
  ⇒ Total Pageviews (both control and experiment group) = 342,663 × 2 = **685,325**

Ideally we should choose the maximum of three total pageviews. But given the total pageviews from '**Retention**' is way too big with a big risk of making the experiment too long to run (given 40,000 views / day, it is approximately 120 days!) to gain any practical feasibility.
So in the answer of this and next quiz, I **ignore 'Retention'** in evaluation. So this is my answer:

- Will you use the Bonferroni correction: **No**
- Which evaluation metrics did you select:
  - **Gross conversion**
  - **Net conversion**
- How many pageviews will you need: **685,325**

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.)

Answers:

- Number of pageviews: **685,325**
- Fraction of traffic exposed: **1.00**
- Length of experiment: **18** (685,325 ÷ (40,000 × **100%**))

In general we should keep an experiment length as short as possible, but sometimes we should also not to introduce huge user experience glitch due to the experiments introduced, so it is a balance call. **Particular to this experiment, we have**:

- Firstly the experiment is a low to **no risk** to participants, nobody can be hurt.
- Secondly the experiment exposes or deals with no sensitive data.

So I can choose to divert entire (**100%**) traffic to this A/B test. Given the average daily page views of 40,000 to Udacity, this test may run slightly more than 17 days. **18 days** is just slightly above 2 weeks free trial period, so the risk of this choice should be acceptable.

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.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 unless all sanity checks pass.**

Calculation:

- Use 95% confidence interval, with 2-tail z value
- Only use data from the experiment spreadsheet vs. baseline table
- **Number of cookies**:
  - Probability of success $p = 0.5$
  - Total Control pageviews: $NP_{ctrl} = 345543$
  - Total Experiment pageviews: $NP_{exp} = 344660$
  - Standard Error $SE = \sqrt{\frac{p(1-p)}{NP_{ctrl} + NP_{exp}}} = 0.000602$
  - Margin of Error $E_m = 1.96 \times SE = 0.0011796$
  - Lower CI bound $p - E_m = 0.4988$
  - Upper CI bound $p + E_m = 0.5012$
  - Observed (divert) $P_{ob} = \frac{NP_{ctrl}}{NP_{ctrl} + NP_{exp}} = 0.5006$
  - Sanity Pass: Yes, since P_{ob} is in CI range
- **Number of clicks**:
  - Probability of success $p = 0.5$
  - Total Control clicks: $NC_{ctrl} = 28378$
  - Total Experiment clicks: $NC_{exp} = 28325$
  - Standard Error $SE = \sqrt{\frac{p(1-p)}{NC_{ctrl} + NC_{exp}}} = 0.0020997$
  - Margin of Error $E_m = 1.96 \times SE = 0.0041155$
  - Lower CI bound $p - E_m = 0.4959$
  - Upper CI bound $p + E_m = 0.5041$
  - Observed (divert) $P_{ob} = \frac{NC_{ctrl}}{NC_{ctrl} + NC_{exp}} = 0.5005$
  - Sanity Pass: Yes, since P_{ob} is in CI range
- **Click-through-probability**:
  - Probability of click-through in control $p = \frac{NC_{ctrl}}{NP_{ctrl}} = 0.0821258$
  - Standard Error $SE = \sqrt{\frac{p(1-p)}{NP_{ctrl}}} = 0.0004671$
  - Margin of Error $E_m = 1.96 \times SE = 0.0009155$
  - Lower CI bound $p - E_m = 0.0812$
  - Upper CI bound $p + E_m = 0.0830$
  - Observed (divert) $P_{ob} = \frac{NC_{exp}}{NP_{exp}} = 0.0822$
  - Sanity Pass: Yes, since P_{ob} is in CI range

Result - to the list of invariant metrics:

	Lower bound	Upper bound	Observed	Sanity check Pass?
Number of cookies	0.4988	0.5012	0.5006	Yes
Number of clicks	0.4959	0.5041	0.5005	Yes
Click-through-probability	0.0812	0.0830	0.0822	Yes

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.)

Calculation:

- For 95% confidence interval, use 2-tail **Z** value
- Use the experiment spreadsheet data vs. baseline table, range from **Sat, Oct 11** to **Sun, Nov 2**
- **No Bonferroni correction**
- **Gross conversion**:
  - Total Enrolled in Control: $NE_{ctrl} = 3785$
  - Total Enrolled in Experiment: $NE_{exp} = 3423$
  - Total Clicks in Control: $NC_{ctrl} = 17293$
  - Total Clicks in Experiment: $NC_{exp} = 17260$
  - Pooled Probability $p = \frac{NE_{ctrl} + NE_{exp}}{NC_{ctrl} + NC_{exp}} = 0.208607$
  - Pooled Standard Error $SE = \sqrt{p(1-p) \times (\frac{1}{NC_{ctrl}} + \frac{1}{NC_{exp}})} = 0.004372$
  - Margin of Error $E_m = 1.96 \times SE = 0.008569$
  - $\widehat{d} = GC_{exp} - GC_{ctrl} = \frac{NE_{exp}}{NC_{exp}} - \frac{NE_{ctrl}}{NC_{ctrl}} = -0.020555$
  - **Lower CI bound** $\widehat{d} - E_m = -0.029124 \approx -0.0291$
  - **Upper CI bound** $\widehat{d} + E_m = -0.011986 \approx -0.0120$
- **Retention** (an option):
  - Total Paid in Control: $NP_{ctrl} = 2033$
  - Total Paid in Experiment: $NP_{exp} = 1945$
  - Total Enrolled in Control: $NE_{ctrl} = 3785$
  - Total Enrolled in Experiment: $NE_{exp} = 3423$
  - Pooled Probability $p = \frac{NP_{ctrl} + NP_{exp}}{NE_{ctrl} + NE_{exp}} = 0.5518868$
  - Pooled Standard Error $SE = \sqrt{p(1-p) \times (\frac{1}{NE_{ctrl}} + \frac{1}{NE_{exp}})} = 0.01172978$
  - Margin of Error $E_m = 1.96 \times SE = 0.0229903688$
  - $\widehat{d} = GC_{exp} - GC_{ctrl} = \frac{NP_{exp}}{NE_{exp}} - \frac{NP_{ctrl}}{NE_{ctrl}} = 0.0310948$
  - **Lower CI bound** $\widehat{d} - E_m \approx 0.0081$
  - **Upper CI bound** $\widehat{d} + E_m \approx 0.0541$

- **Net conversion**:
  - Total Paid in Control: $NP_{ctrl} = 2033$
  - Total Paid in Experiment: $NP_{exp} = 1945$
  - Total Clicks in Control: $NC_{ctrl} = 17293$
  - Total Clicks in Experiment: $NC_{exp} = 17260$
  - Pooled Probability $p = \frac{NP_{ctrl} + NP_{exp}}{NC_{ctrl} + NC_{exp}} = 0.1151275$
  - Pooled Standard Error $SE = \sqrt{p(1-p) \times (\frac{1}{NC_{ctrl}} + \frac{1}{NC_{exp}})} = 0.003434$
  - Margin of Error $E_m = 1.96 \times SE = 0.00673064$
  - $\widehat{d} = GC_{exp} - GC_{ctrl} = \frac{NE_{exp}}{NC_{exp}} - \frac{NE_{ctrl}}{NC_{ctrl}} = -0.0048737$
  - **Lower CI bound** $\widehat{d} - E_m \approx -0.0116$
  - **Upper CI bound** $\widehat{d} + E_m \approx 0.0019$

Result - to the list of evaluation metrics, ***Bonferroni correction:*** **No**

	Lower bound	Upper bound	Statistical Significance?	Practical Significance?
Gross conversion	-0.0291	-0.0120	**Yes** (< 0)	**Yes** - ABS_min > d_{min} (0.01) - d_min is outside of CI ABS range
Retention	0.0081	0.0541	**Yes** (> 0)	No - d_{min} (0.01) falls in the CI range
Net conversion	-0.0116	0.0019	No (0 is in the CI range)	No - Not even Statistical Significance - ABS_min (0.0019) < *dmin* (0.0075)

Conclusion:

- Gross conversion:
  - Experiment is Statistical Significance, with negative implication (i.e. fewer enrollment)
  - Experiment is also practical significance, in that difference is notable
- Retention:
  - Experiment is statistical significance & positive implication (i.e. higher sustained rate of paid users)
  - But experiment is not practical significance at 95% CI, difference is not notable
- Net conversion:
  - Experiment is neither statistical significance nor practical significance
  - Experiment data can NOT show clear implication to sustained paid users in neither direction

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.)

For sign and binomial test, Using online resource to calculate p-value statistics:
http://graphpad.com/quickcalcs/binomial1.cfm

Sign Test:

- Use the experiment spreadsheet data vs. baseline table, range from **Sat, Oct 11** to **Sun, Nov 2**
- **No Bonferroni correction**
- For each evaluation metrics, table below has the $(exp. - ctrl.)$ delta of day-by-day data

Gross Conversion GC($exp. - ctrl.$) per day	Retention RE($exp. - ctrl.$) per day	Net Conversion NC($exp. - ctrl.$) per day
-0.04198972165	-0.1985785359	-0.05232960308
-0.04093276535	0.3082922824	0.02606477355
-0.01969122252	-0.02403468924	-0.01514393521
-0.01973467251	-0.00641025641	-0.01435262059
-0.02647389946	0.2787905346	0.0365172089
-0.003973638601	-0.1213346815	-0.02222431244
-0.03236665295	-0.1740912523	-0.04519402166
-0.02987885377	0.02321083172	-0.01566746913
-0.01741389083	0.1836513995	0.0236427775
-0.01373065392	0.05264408794	0.001293790347
-0.06055763809	-0.03921078921	-0.03893134051
-0.03351717275	-0.02102623457	-0.02239444132
-0.000946290961	0.1164321673	0.0217084768
-0.0485587777	-0.07000937207	-0.04641415875
-0.06486775302	-0.08947745168	-0.06416255119
-0.006621909164	0.01416798603	-0.001149509624
-0.03071828076	0.02649660481	-0.00902785254
0.01086956522	0.1498316498	0.03940217391
0.01125540481	0.03883643009	0.0156760409
0.05682022337	-0.1254302988	-0.01014686948
0.005618638814	0.09536440571	0.02730062072
-0.02475834311	0.108071506	0.007321015434
-0.04587708179	0.2406913475	0.04558409669

- The value of each cell in above table is the daily difference of the evaluation metric in experiment group minus control group, only sign (+ or -) is useful for sign test.

Sign Test summary (for each evaluation metrics)**:**

- **Gross conversion**:
  - Total samples: 23 (days)
  - Number of 'successes' samples: 4 (+)
  - Hypothetical probability for sign test: 0.5
  - Two-tail **p-value = 0.0026**
  - Statistical significance: **Yes** (p-value $\ll$ α at 95% confidence level)

- **Retention**:
  - Total samples: 23 (days)
  - Number of 'successes' samples: 13 (+)
  - Hypothetical probability for sign test: 0.5
  - Two-tail **p-value = 0.6776**
  - Statistical significance: **No** (p-value $\gg$ α at 95% confidence level)
- **Net conversion**:
  - Total samples: 23 (days)
  - Number of 'successes' samples: 10 (+)
  - Hypothetical probability for sign test: 0.5
  - Two-tail **p-value = 0.6776**
  - Statistical significance: **No** (p-value $\gg$ α at 95% confidence level)

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.

Since in the experiment design, I expect all evaluation metrics to show statistical significance to implement the change (launch the experiment), I don't need the Bonferroni correction. Normally you will need to use Bonferroni correction, if your experiment decision (to implement) can be made only on one or a few evaluation metrics of significant change from the experiment.

Fortunately the results (statistical significance) from sign tests matched up with the effect size hypothesis tests result, except for the **'Retention'** metric, which we know that we could not obtain enough samples in any reasonable time frame to support the power of experiment. So this discrepancy is fine, particularly the effect size hypothesis test also confirmed that 'Retention' metric is NOT practical significant.

Recommendation

Make a recommendation and briefly describe your reasoning.

Based on the given data, analytic estimate and experiment analysis, because not both 'Gross conversion' and 'Net conversion' show a significant change, also the confidence interval of the 'Net conversion' (-0.0116 ~ 0.0019) does include the negative boundary (-0.0075) of its practical significance, so it is possible to believe that the experiment may introduce a negative impact on the business (i.e. the number of continued paid-users). Base on these, I would not recommend Udacity to implement the change (i.e. launch the experiment).

Some follow-up experiments may help further.
Ignored 'Retention' for now. It is also insignificant even though it doesn't have enough power.

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.

From earlier on 'Number of Pageviews vs. Power' analysis, we know that if we could better reduce the number of frustrate students who enrolled but cancelled early (in 14 days) in the course, we could obtain a higher 'baseline conversion rate' of 'Retention', with everything else kept the same, it yields a much lower sample size (of pageview) required for metric 'Retention'.

On the other hand, if the 'drop-out' rate of enrolled students goes lower, the 'Net conversion' will move to closer to 'Gross conversion', this yields much stronger correlation that 'Net conversion' will be on the same course with 'Gross conversion', which helps to drive experiment decision.

So my follow-up experiment will be focus on drive down the 'drop-out' rate.

Follow-up experiment detail (with new metrics and unit of diversion)**:**

- To cut down drop-out rate, so that more enrolled students turn to be paid users, we can improve or change the simple 'Screen-take' questionnaire mechanism to 'enroll' a user. The 'Screen-take' questionnaire is easy but way too casual for the data quality and accuracy. In this proposal, I'd like to give each user who clicked 'Start free trial' a day or two to complete a well designed course prerequisite test/exercise prior to their confirmed enrollment. The time (days) for student to submit prerequisite test itself can be estimated initially, then adjusted (be another A/B test) later.

- Hypothesis: due to the above change from simple 'Screen-take', I would expect now confirmed enroll-users are more serious and even more suitable to a course, have them move to the end of 'free-trial', there will be less confusion, frustration, randomness, but more determination. So I'd expect the conversion rate from this stricter enroll to sustained paid-user be much higher, which improves metrics of both 'Retention' and 'Net conversion' statistically.

- For this follow-up test, we should change the **unit-of-diversion** to **'unique user_ID'**.
  This is because the new experiment takes place after enrolling with an associated user_ID, which uniquely identifies each test subject. On the contrary, cookie and click based metrics are not relevant, and less stable than the ones based on user_ID here.

- Invariant Metrics:
  - Number of user-ids (users clicked "Start trial" but enroll is yet confirmed): N_{uid}

- Evaluation Metrics:
  - Number of users finally enrolled (who passed prerequisite evaluation): N_{end}
    - It is different between control (without prerequisite) and experiment group
  - Enrollment-Probability (similar to Gross conversion): N_{end} / N_{uid}
  - Retention: N_{paid} / N_{end}, where N_{paid} is the number of user-ids to remain enrolled past 14-days free-trial boundary (i.e. paid users).
  - Net conversion: N_{paid} / N_{uid}

The End

Hai Xiao

List of web sites, books, forums, etc. that is used in this submission:

https://www.udacity.com

http://www.evanmiller.org/ab-testing/sample-size.html

http://graphpad.com/quickcalcs/binomial1.cfm

http://www.stat.ufl.edu/~athienit/Tables/Ztable.pdf