

PROJECT

Design an A/B test

A part of the Data Analyst Nanodegree Program

PROJECT REVIEW

NOTES

Requires Changes

9 SPECIFICATIONS REQUIRE CHANGES

Dear student,

well done with your submission, this is a fairly complex exam that involves challenging concepts and tricky decision nodes. It is designed to provide a flavour of what it means to deliver and interpret results as a data scientist and to design a real experiment. I hope that my comments might help you through this demanding journey. I've left as many comments as possible trying to provide actionable answers for you to proceed in your submission, I hope you might find them helpful.

Let's team up in building the best possible project!

Metric Choice

A good set of metrics have been selected for the experiment, without missing any necessary or valuable metrics.

SHARE YOUR ACCOMPLISHMENT



Rate this review

Each metric has a clear and well-reasoned explanation of why it was or was not chosen as an invariant metric and as an evaluation metric.

As for the number of user_ids are not sure I understand the rationale behind: "Udacity does not expect a significant modification of this metric. This means Udacity expects a behavioural change from its users."- I'm not sure why Udacity is mentioned, the assessment section is about discussing whether a metric is suitable as invariant or evaluation.

Please explicitly explain why you are not choosing the number of userids as an evaluation metric. Hint: it is not about correlation as you seem to imply in: "As the 'number of user id' metric and the 'Gross conversion' are highly correlated", in that case we could randomly pick one or the other, there is another issue, which of the two metric is normalised and which is a raw count?

The report clearly states what results we look for in order to launch the experiment and the stated results are aligned with the experiment goals.

The expectation section is a bit confusing, I'm not sure about what is meant, in each answer by "Udacity does or does not expect" something. This section is not about what you (or Udacity) thinks is likely to happen, this section is about explicitly describing how the metrics you have chosen should behave in order for you to launch the experiment. Some more detailed and specific expectations should be provided though. Our initial hypothesis is made of two parts, namely:

1. "...this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time"
2. "without significantly reducing the number of students to continue past the free trial and eventually complete the course."

How do we tell if we significantly reduced the number who left the free trial early? How do tell if we did not significantly reduce the number who continue past the free trial? Which metric is sufficient to inform about the first part of the hypothesis, which one informs about the second? How should the metrics behave for us to launch? (Should they increase, decrease, stay the same?). Please describe which part of the hypothesis is covered by each metric you have chosen and how you exactly expect that very metric to behave in order to launch the experiment.

Variability

The standard deviations for all evaluation metrics have been correctly calculated.

Rate this review

Each evaluation metric has a clear and correct explanation of whether the analytic variability is likely to match the empirical variability.

Sizing

The number of pageviews given is correct given the students choice of whether to use the Bonferroni correction.

A well-reasoned argument about how risky the experiment will be is made and a fraction of traffic to divert is chosen accordingly.

Well done assessing that the experiment is low risk; diverting 50% of the traffic will take the experiment a few weeks to run, actually more than a month, which is more than the 'few weeks' that are acceptable for this experiment. Running the experiment for more time involves expenses and prevents other experiments from being run, therefore there must be a solid justification for that otherwise a higher percentage of the traffic should be diverted in a low risk setup like ours. Regarding risk assessment you can refer to lesson 2 sections 4 and 5.

Some hints:

1. Is there a chance that anyone gets hurt because of the duration of our experiment?
2. Are we dealing with sensitive data? (Political attitudes, personal disease history, sexual preferences)

If not, this experiment is not risky and we should divert the entire traffic.

The duration of the experiment is correctly calculated given the fraction of traffic to divert that was chosen.

You might want to revise this answer once the above issues have been addressed.

Sanity Checks

The sanity checks have been correctly calculated for all chosen invariant metrics.

The passing or failure of all sanity checks have been evaluated. If sanity checks failed, analysis has been performed to discover why the sanity checks may have failed and the experiment has not been continued.

Rate this review

permitted to discover any and every errors they have made and the experiment has not been conducted.

Effect Size Tests

Correctly calculated confidence intervals have been reported for the difference in all evaluation metrics.



Statistical and practical significance have been correctly reported for all evaluation metrics.

Sign Tests

P-value and statistical significance have been correctly reported for all evaluation metrics.

Results Summary

The report provides good justification for the choice of whether to use the Bonferroni correction.

The answer is correct, this section does not meet specifications however because the rationale is not the proper one. Let me indulge in this as it is one of the most complex concepts involved in this exam: To propose your recommendations you will, correctly, consider *both* the net and gross conversion. That is because, in order to launch, you would need them *both* to match our expectations (we look for a decrease in gross conversion and for a no decrease in the net conversion). We are in the situation where more metrics need to be *all* matching what we expect in order to launch. The case where *all* metrics need to match the expectations in order to launch is not the same as the case where *any* metric needs to match the expectations. In fact it is the exact opposite: For the former the risk of a  increases as the number of metrics increases, for the latter the risk of a  increases.

Let me help with an example:

Let's imagine we have 20 metrics, and let's imagine we decide we want *all* of them to match expectations (like in our case). Which is our risk? We risk not to launch because if at least one metric (out of 20) fail to reject the null, when the null is not the true effect (Type II error).

Conversely if we were to launch the experiment when *any* metric would match our expectations (so we would launch

Rate this review

if just one metric out of 20 does what we expect) then we would have to use Bonferroni: Out of 20 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.

<http://www.stat.berkeley.edu/~mgoldman/Section0402.pdf>

<http://www.utdallas.edu/~herve/Abdi-Bonferroni2007-pretty.pdf>

<http://onlinelibrary.wiley.com/doi/10.1111/opo.12131/full>

A well-reasoned and plausible explanation for each discrepancy between the effect size tests and the sign tests has been provided.

Recommendation

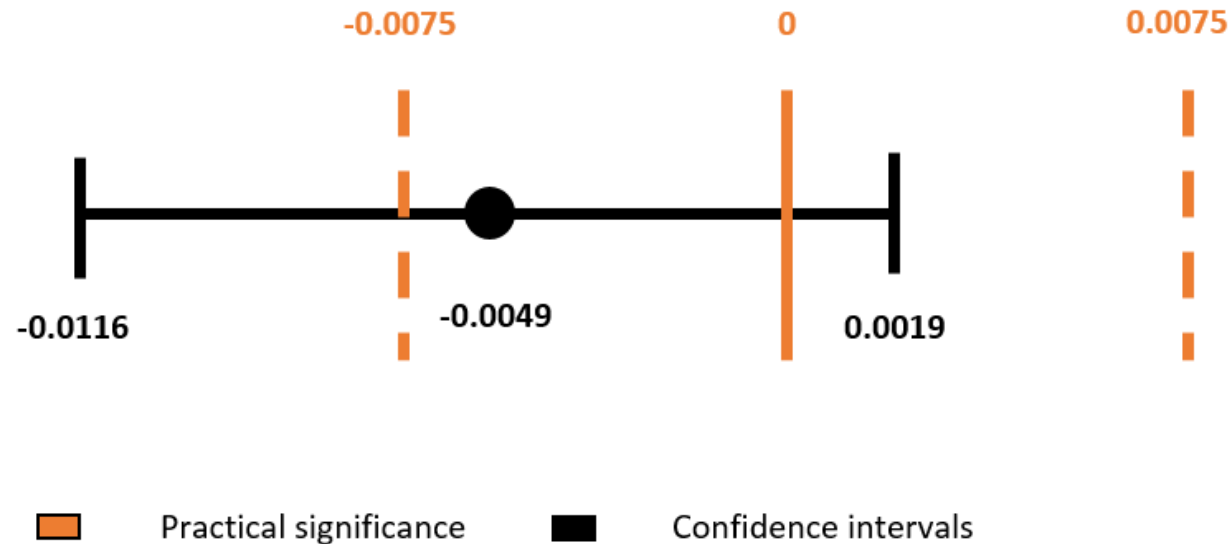
A recommendation is made that is well-reasoned and supported by the data.

The answer is correct though the rationale is not. Please update your expectation section, clearly stating what is expected from each metric in order to launch or not, and then contrast your expectations with the results to justify your choice:

1. Please note that the effect on the gross conversion is not adverse as you seem to imply in: "I saw a significant decrease of the gross conversion. Even if this decrease could be expected, this was not the Udacity desired effect." It is actually exactly what we were expecting and we were looking for in with the 1st part of our hypothesis, namely: "...this might set clearer expectations for students upfront, thus reducing the number of frustrated students who left the free trial because they didn't have enough time". A drop in the gross conversion means that we manage to reduce the number of frustrated students.
2. What happens with the net conversion? Is it behaving like we expected in order to launch? When discussing the net conversion you will notice that the confidence interval does include the negative of the practical significance boundary, that is, it's possible that this number went down by an amount that would matter to the business. Is this an acceptable risk in order to launch?

Rate this review

NET CONVERSION



Follow-Up Experiment

A plausible experiment that would be worth testing has been made. A hypothesis for results of the experiment is clearly stated.

The experiment logic is **identical** to the one proposed in the exam. To meet specifications you might need to design something more original and innovative. I'm sure you understand we cannot accept an experiment that is, logically, exactly the same as the one already proposed in the exam: A design effort is required here as this is the section that really proves that you understand how to:

What is required here is to:

1. Design and describe an experiment that helps reduce early cancellations.
2. Propose and clearly state an hypothesis.

Rate this review

3. Construct and explain proper metrics to evaluate it and explain how they inform the hypothesis.
4. Propose a coherent unit of diversion and invariant metric(s).

It doesn't need to be complex. You might consider, for instance, some kind of incentive after enrollment and evaluate it, that would make things easier as, after enrollment, the unit of diversion can be a user_id which is much more stable than a cookie.

The metrics chosen in the report will be sufficient to evaluate the hypothesis of the experiment, would be possible to measure under most infrastructures, and are well-supported by reasoning in the report.

This section can be fully evaluated only in the light of valid follow up experiment.

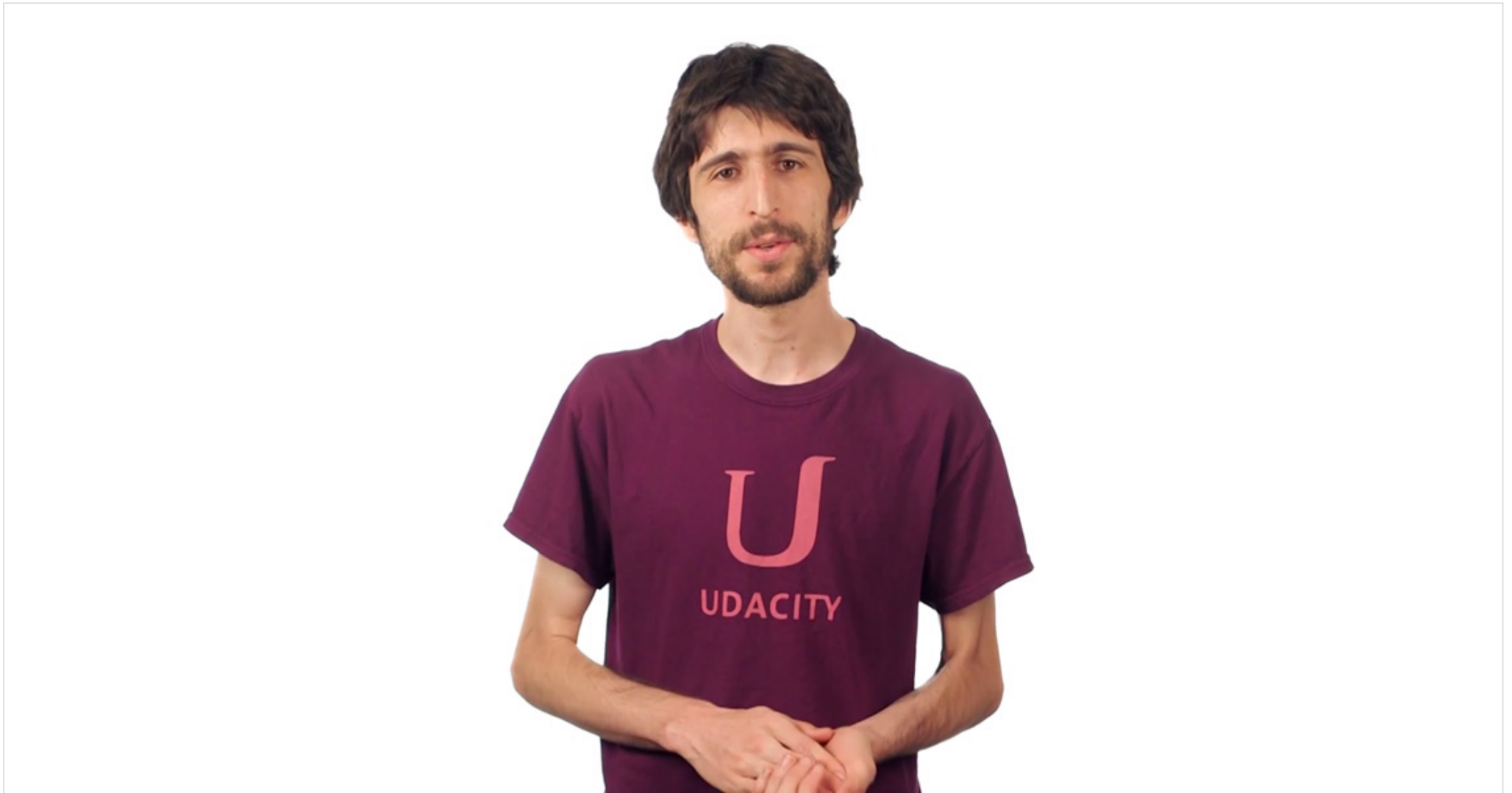
The report describes a reasonable unit of diversion and gives good support for this choice.

This section can be fully evaluated only in the light of valid follow up experiment.

 RESUBMIT PROJECT

 DOWNLOAD PROJECT

Rate this review



Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

[Watch Video](#) (3:01)

Rate this review

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.