UDACITY

EN-US ZH-CN

Design an A/B test

Criteria	Meets Specifications
Metric Choice	
Have good invariant and evaluation metrics been selected for the experiment?	A good set of metrics have been selected for the experiment, without missing any necessary or valuable metrics.
Has a well-reasoned justification of the choice of metrics been made?	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.
For which results would we wish to launch the experiment?	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.
Variability	
Have the standard deviation for all evaluation metrics been correctly calculated?	The standard deviations for all evaluation metrics have been correctly calculated.
Has reasoning been made whether each analytic standard deviation is likely to be accurate?	Each evaluation metric has a clear and correct explanation of whether the analytic variability is likely to match the empirical variability.
Sizing	
Does the number of pageviews correctly take into account the planned analysis?	The number of pageviews given is correct given the students choice of whether to use the Bonferroni correction.

Has an appropriate level of exposure for the experiment been chosen based on the risk?	A well-reasoned argument about how risky the experiment will be is made and a fraction of traffic to divert is chosen accordingly.
Does the duration of the experiment correctly take the exposure chosen into account?	The duration of the experiment is correctly calculated given the fraction of traffic to divert that was chosen.

Sanity Checks

Have sanity checks been performed correctly?	The sanity checks have been correctly calculated for all chosen invariant metrics.
Have the results of sanity checks been analyzed?	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.

Effect Size Tests

Have confidence intervals been calculated for the difference in all evaluation metrics?	Correctly calculated confidence intervals have been reported for the difference in all evaluation metrics.
Have statistical and practical significance been correctly evaluated?	Statistical and practical significance have been correctly reported for all evaluation metrics.

Sign Tests

Has a sign test p-value been reported for each evaluation metric with indications whether the sign test is statistically significant?	P-value and statistical significance have been correctly reported for all evaluation metrics.

Results Summary

Has the choice whether to use the Bonferroni correction been justified?	The report provides good justification for the choice of whether to use the Bonferroni correction.
Have all discrepancies between the effect size tests and the sign tests been analyzed?	A well-reasoned and plausible explanation for each discrepancy between the effect size tests and the sign tests has been provided.

Recommendation

Has a well-reasoned recommendation been made based on the results of the experiment?	A recommendation is made that is well-reasoned and supported by the data.

Follow-Up Experiment

Has a plausible experiment for the purpose given been made with a clearly stated hypothesis?	A plausible experiment that would be worth testing has been made. A hypothesis for results of the experiment is clearly stated.
Have good metrics to evaluate the proposed experiment been selected with good reasoning to support them?	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.
Has a well-reasoned unit of diversion for the experiment been selected?	The report describes a reasonable unit of diversion and gives good support for this choice.