Final Analysis of A/B Test #1

Oliver Keyes

August 25, 2015

This report covers the first A/B test run by the Discovery team. In summary, we did not find an unambiguous improvement in outcome for users subject to the experimental condition, and recommend disabling the experiment and returning users to the default search experience - coupled with some pointers on future A/B test design.

Background

On Thursday, 6 August 2015 we launched an A/B test over the Wikimedia search system ("Cirrus") aimed at reducing the rate at which user queries would provide no results.

This test ran for 1 week over 10% of user queries. 5% of users were given our default search experience. 5% were given a system that had a reduced confidence - in other words, produced results when it was less certain they were *good* results - and a different smoothing algorithm to try and increase the quality of results returned with that reduced confidence.

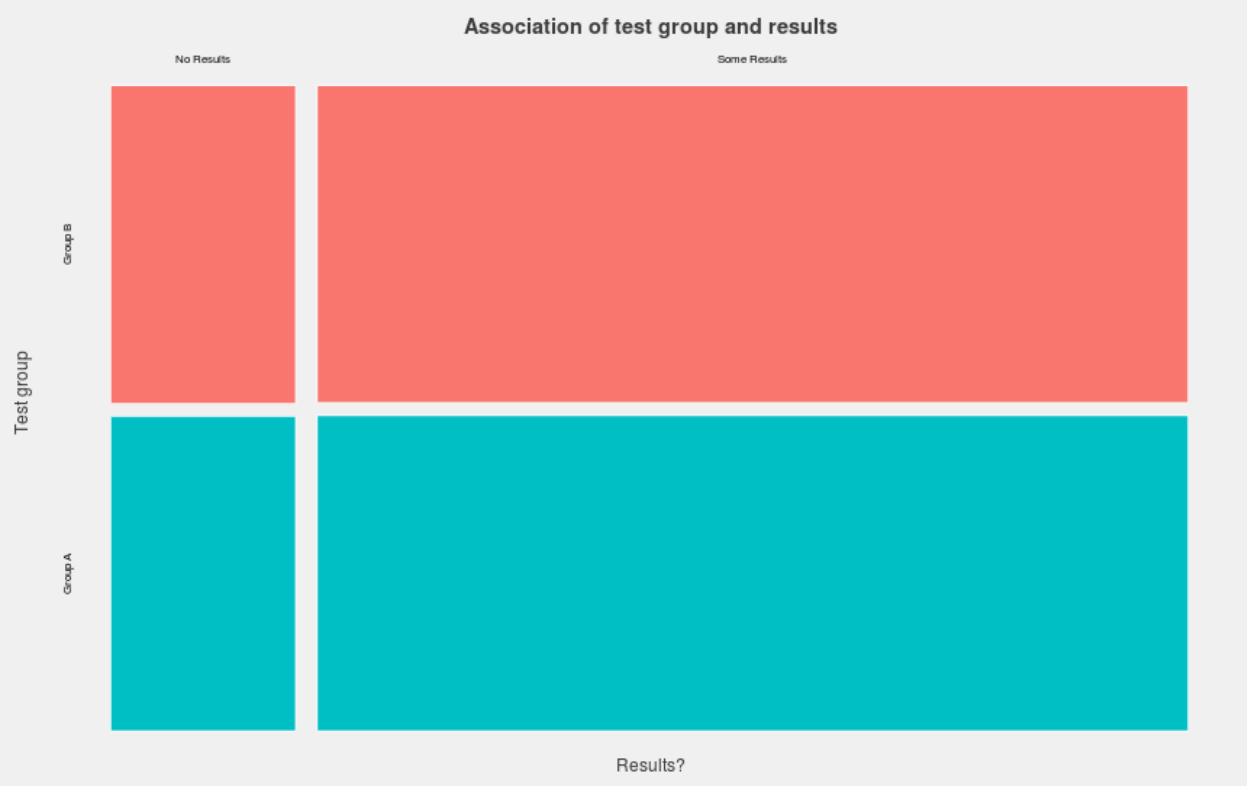
Our intent with this test was not only to impact the zero results rate but to identify problems with the testing process and resolve them so that future, larger tests would have more certain outcomes (and those outcomes would be easier to reach).

Initial results

The initial analysis we performed (which is documented separately) ran over 7 million events from the first day of A/B testing. It found statistically significant difference, with the test group showing a lower zero results rate than the control group, but this difference was small enough to be caused by sampling issues (discussed below), or to be entirely real but not worth deploying more widely.

Final results

For the final analysis we examined 7 days of logs, removing entries that only contained prefix search queries (which were not impacted by the test) or which came from known misusers of our systems. This resulted in a dataset of 13,450,869 events between 7 and 14 August 2015.



As with the initial test, there is no visually distinct difference between the outcome for Group A (the control group) or Group B (the test group). When we perform a Chi-Squared test for statistical significance we find a P-value of 0.00006, indicating that even if the difference in outcome is small, it *is* statistically significant and is unlikely to be from chance.

What that difference favours, however, is not Group B, as seen in the initial analysis - it's Group A. People given the default treatment are statistically significantly more likely to get results than those subject to the experimental treatment, albeit not enough to make much of a difference (the improvement is 0.08%)

So, the final analysis has produced the opposite outcome to the initial analysis. What does this mean?

The most probable reason for this is the size of the datasets. When dealing with 7m and 13m events, respectively, over two conditions, it is possible to see statistically significant outcomes - ones that vary wildly depending on *which* large sample is being used - where nothing is really taking place. It's simply the natural difference in the datasets on a large enough scale that significance is possible to achieve.

One solution for this problem is to perform power analyses before experiments to identify the number of results needed to see an effect, and using this to inform our sampling rate. That way we can be more certain that any effect we see is at least a real effect. We can also (instead of, or in addition to, power analysis) switch to a more Bayesian approach to A/B testing that is not as vulnerable to this class of issue.

Conclusion

Based on the tiny effect size seen, and the difference in the direction of this effect between the initial and final sample, we argue that the experiment failed to demonstrate an improvement in the zero results rate for searches in the testing group We recommend that the experiment be switched off, and that all users should recieve the existing, 'default' search experience.

In terms of how we approach future experiments, the conclusions of the initial analysis around experimental design stand, but are augmented with a recommendation that power analysis be performed prior to the setting of a sample rate, for each A/B test we deploy.