A/B Testing

General Guidelines

Contact: jlinncsu@gmail.com

Firstly, it is important to understand how A/B testing, and statistical testing in general, needs to be approached to make sure you can trust the results.

1. **Everything related to the statistical test needs to be defined in advance, before running the test**. You cannot change your hypotheses by looking at the data after the A/B test is run and then, using those data, do a t-test to prove the new hypotheses. Statistical tests are meant to answer questions, but questions cannot change. If you develop new hypotheses, you will need to design a new A/B test to prove them.

2. Along the same lines, it is fine to test the effect of a change on multiple metrics, but these metrics need to be chosen in advance. You can't test metrics that were not defined before running the test. You need to define in advance which metrics you will look at as well as which thresholds will make you declare the test as a winner. I.e., my goal is to improve this metric by at least X% and I will also check this other metric which I expect to go down but am fine if it doesn't decrease by more than Z%.

3. Things can get very complicated quickly. Many key metrics are negatively correlated, so one goes up and the other one goes down. It is easy to increase conversion rate by reducing the denominator, which is the number of site visitors. But reducing the site visitor metric is not a great idea. The converse is true though, increasing site visitors is usually very good, even if it means reducing conversion rate. Or metrics can go up for the wrong reason. Example: time spent on site is a key social network metric and you want that to usually go up. But most changes that will simplify user experience might make time spent on-site decrease, at least short term.

4. To avoid the huge mess, it is good to have a "north star metric" that will be checked for any test you run. And every test should have a positive impact on this metric. North star metrics are

usually revenue for e-commerce sites, or engagement/retention for ads-based sites (which also means money, at the end of the day).

5. Say you run a test on Airbnb or FB. The change you are testing makes easier to upload a profile picture. The specific test metric would be something like percentage of users with a profile picture. But then you also want to test how that affects # of bookings or retention, respectively. Checking on the north star metric will avoid declaring bad tests as winners. For instance, you could run a test where you automatically upload a picture for all users. This will obviously improve percentage of users with a pic. And probably will also improve things like time spent on site, cause users will have to spend time to remove it or change it. But it will never improve # of bookings or retention. Again, all these things need to be defined before running the test.

6. Always keep in mind that an A/B test is not answering the question which site version is better in absolute terms. It is answering the question which site version is better for your current users. This could lead to a vicious cycle where you start from a very self-selected user base and you keep optimizing for them, alienating different user demographics. That's why it is useful to segment your test results by different user characteristics. But, again, the segments as well as the hypotheses on why certain segments might react differently need to be defined before running the test.

7. There is a sort of a paradox when it comes to A/B testing: as a product data scientist your main role is to deep dive into the data to uncover useful information, and useful often means unexpected. But in statistical testing, as we just said, you can't do that. You can only test hypotheses developed before collecting the data. However, that doesn't imply that you cannot deep dive into A/B test results. You certainly can and should. It simply means that the deep dive insights can only be used to develop new hypotheses for new A/B tests, and not be used to reach conclusions.

8. Example: you run a test with a new home page UI for your e-commerce site. Your hypothesis is that it will decrease number of bouncers (people who leave after visiting just one page) and, therefore, it will increase revenue. You run the test, but results are not significant. Then, you start looking at how users from different countries performed and notice that Asians are doing better with the new UI. You cannot now change the home page for Asians because that was not your test hypothesis when you designed the test. But, based on what you discovered, you develop a new hypothesis of why that's happening and decide to run a new A/B test just for Asians, with either the same new page or, most likely, an improved home page UI based on the new hypothesis you developed. And so on, testing -> check results -> develop new hypotheses -> test -> etc.

9. As a rule of thumb, do a t-test on the A/B test data. If t-test assumptions are not met, like for instance events are not independent, it means you have designed the test the wrong way and you should fix that, no point in looking for different statistical tests. Z-test will give you the same results as the t-test, but t-test is slightly more statistically sound. In any case, if you see any difference between z-test and t-test, it means you have way bigger problems than which one to choose between the two. It means the entire design was broken. Btw, if two things consistently give you the same result, it is irrelevant which one you choose.