A Summary of Udacity A/B Testing Course

Kelly Peng Follow
Nov 12, 2017 · 10 min read

Recently I finished the A/B testing course by Google on Udacity. The course has been highly recommended to people who want to learn about A/B testing. I think it would be very helpful to write a summary about what I've learned because my notes are very detailed, and I also heard my friends complaining that it's hard to follow the instructors. Hope this summary could help more people.

This course emphasizes heavily on the business application of A/B testing instead of the statistical aspect. However, if you want to learn more about statistical inference to build

a foundation for A/B testing, I recommend you find a stats book or watch a few videos as a complement to this course.

. . .

First, why do we do A/B tests?

The answer is testing takes the guesswork out of website optimization and enables data-informed decisions that shift business conversations from “we think” to “we know.” By measuring the impact that changes have on your metrics, you can ensure that every change produces positive results. Nowadays it’s very common for companies to do A/B tests on web page versions, personalized recommendations and new features.

Can we test everything?

NO. There are situations we cannot analyze through A/B test. For example, if you are adding a new experience and want to test it, old users may resist against the new version (change aversion), or old users may all go for the new experience, then the test set has everything (novelty effect). Two issues to consider when it comes to new experience: (1) what is the base of your comparison? (2) how much time you need in order for your users to adapt to the new experience, so that you can actually say what is the plateaued experience and make a robust decision? Except for new experience, long term effect is hard to test too. For example, a home rental website test its referral effect, but a customer may not return even in six months, it’s very hard to measure through A/B testing. If this is the case, what shall we do?

When A/B testing is not useful, we can:

- Analyze the user activity logs
- Conduct retrospective analysis
- Conduct user experience research
- Focus groups and surveys
- Human evaluation

Then, how to do an A/B test?

In practice, an A/B test can be summarized into the 5 steps below:

1. Choose and characterize metrics to evaluate your experiments, i.e. what do you care about, how do you want to measure the effect
2. Choose significance level (alpha), statistical power (1-beta) and practical significance level you really want to launch the change if the test is statistically significant
3. Calculate required sample size
4. Take sample for control/treatment groups and run the test
5. Analyze the results and draw valid conclusions

In this Udacity course, the five steps are expanded into detailed explanation with numerous real-world examples:

Step 1: Choose and characterize metrics for both sanity check and evaluation

The metrics we choose for sanity check are called as invariant metrics. They are not supposed to be affected by the experiment. They should not change across control and treatment groups. Otherwise, the experiment setup is incorrect.

The evaluation metrics we choose are used to measure which variation is better. For example, we could use daily active users (DAU) to measure user engagement, use click through rate (CTR) to measure a button design on a webpage, etc. In general, there are **four categories of metrics** that you should keep in mind:

- Sums and counts
- Distribution (mean, median, percentiles)
- Probability and rates (e.g. Click-through probability, Click-through rate)
- Ratios: any two numbers divide by each other

Other than choosing the category of metrics, you should also consider **sensitivity** and **robustness**. You want to choose a metric that is has high sensitivity, that means the metric can pick up the change you care about. You also want the metric to be robust against changes you don't care about. It means the metric doesn't change a lot when nothing you're interested happened. If a metric is too sensitive then it is not robust enough, thus there's a balance between these two and you need to look into the data to find out which metric to use.

How to measure the sensitivity and robustness?

- Run experiments
- Use A/A test to see if metrics pick up difference (if yes, then the metric is not robust)
- Retrospective analysis

Step 2: Choose significance level, statistical power and practical significance level

Usually the significance level is 0.05 and power is set as 0.8. Practical significance level varies depends on each individual tests, it tells you how much change the test detects that makes you really want to launch the change. You may not want to launch a change even if the test is statistically significant because you need to consider the business impact of the change, whether it is worthwhile to launch considering the engineering cost, customer support or sales issue, and opportunity costs.

Step 3: Calculate required sample size

Overview: Need to consider the choice of metric, the choice of unit of diversion, and the choice of population into account because they all affect the variability of your metrics. Then decide on the size of experiment.

- **Subject**: What is the subject (**unit of diversion**) of the test? I.e. what are the units you are going to run the test on and comparing. Unit of diversion can be event based (e.g. pageview) or anonymous ID(e.g. cookie id) or user ID. These are commonly used unit of diversion. For user visible changes, you want to use user_id or cookie to measure the change. If measuring latency change, other metrics like event level diversion might be enough.
- **Population**: What subjects are eligible for the test? Everyone? Only people in the US? Only people in certain industry?
- **How to reduce the size of an experiment to get it done faster?** You can increase significance level alpha, or reduce power (1-beta) which means increase beta, or change the unit of diversion if originally it is not the same with unit of analysis (unit of analysis: denominator of your evaluation metric) .

Step 4: Take sample for control/treatment groups and run the test

Several things to keep in mind:

- **Duration**: What's the best time to run it? Students going back to college? Holidays? Weekend vs. weekdays?
- **Exposure**: What fraction of traffic you want to expose the experiment to? Suggestion is take a small fraction, run multiple tests at the same time (different days: weekend, weekday, holiday).
- **Learning effect**: When there's a new change, in the beginning users may against the change or use the change a lot. But overtime, user behavior becomes stable, which is called plateau stage. The key thing to measure learning effect is time, but in reality you don't have that much luxury of taking that much time to make a decision. Suggestion: run on a smaller group of users, for a longer period of time.

Step 5: Analyze the results and draw conclusions

First step, sanity check.

Before analyzing result the first step is to do sanity check — check if your invariant metrics have changed. If your sanity check failed, do not proceed. Instead, go analyze why your sanity check failed. You can do either: (1) retrospective analysis, or (2) look into if there's learning effect.

Second step, analyze the results.

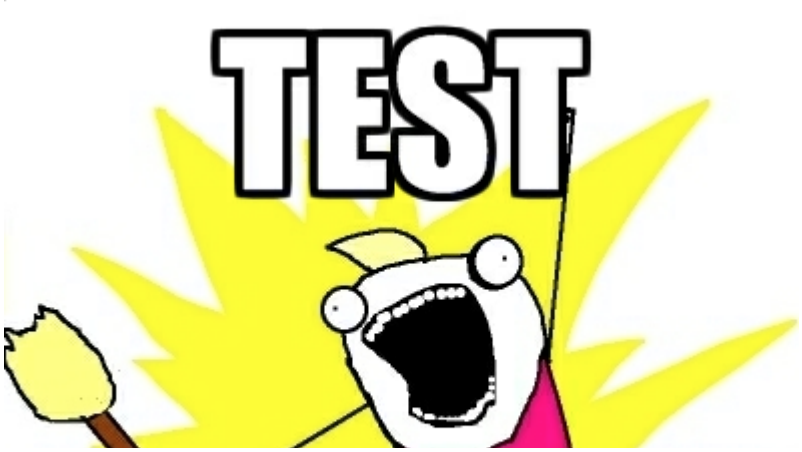
a. If you have one single metric for evaluation, and if it is not significant:

First round: Inspect result, to see if there is really not significant difference. e.g. Break down into different platforms, or day of the week. This may help you find out bug in the system, and may also help you find insight about how users react to your experiment.

Second round: Cross checking by using different methods. e.g. Compare with non-parametric sign test with parametric hypothesis test. What do you do if your hypothesis test and sign test does not agree? You should look into your data critically because you may be suffering from Simpson's paradox (a trend appears in different groups of data but disappears or reverses when these groups are combined). The reasons for Simpson's paradox happening could be: (1) The setup of your experiment is incorrect; (2) The change affects the new user and experienced users differently.

b. If you are measuring multiple metrics at the same time

(1) One potential problem is, you might see a significant result by chance. (Check out this xkcd: significant)



For example, you are running a tests with 20 variants, and you test each hypothesis separately:

$P(\text{one significant result}) = 1 - P(\text{no significant results})$

$P(\text{one significant result}) = 1 - (1 - 0.05)^{20} = 0.64$

There's very high chance you'll see a significant result by chance!! Luckily, there are **several ways to solve this problem**:

- **Bootstrap** and run experiments again and again, the significant metric should disappear if it occurred by chance.
- **Bonferroni correction**: divide the significance level 0.05 by the number of tests in the multiple testing. Say if you are measuring 20 tests, then your significance level for the test should be $0.05/20 = 0.0025$. The problem of Bonferroni correction is it tends to be too conservative. If many metrics are tested at the same time, maybe none of them turned out to be significant.
- Control **Familywise Error Rate (FWER)**: probability that any metric shows false positive.
- Control **false discovery rate (FDR)**: FDR = # false positives / # total rejections.

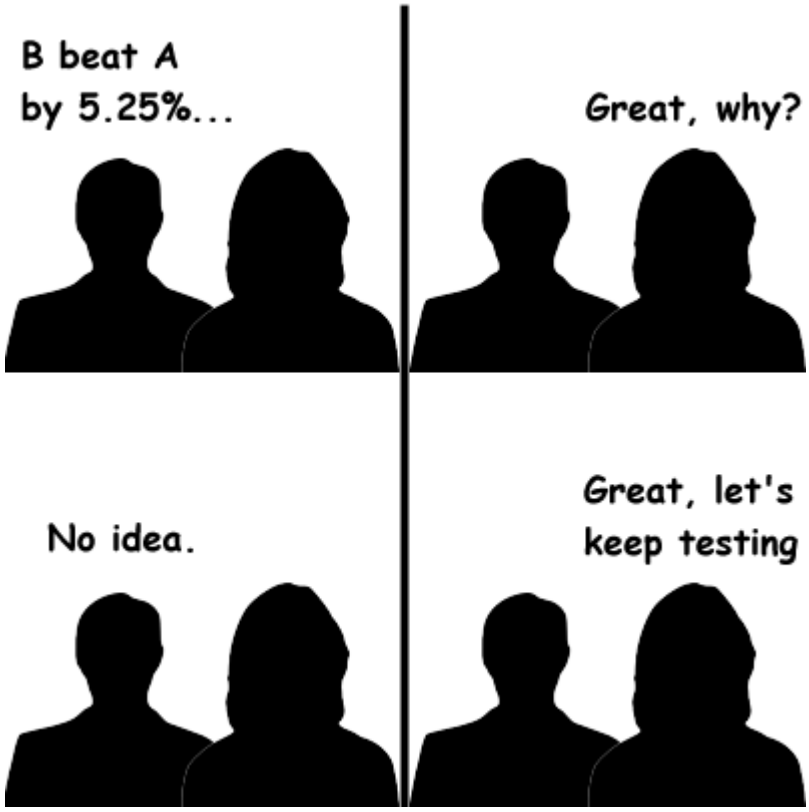
(2) Another potential problem is, what if metrics are **not moving at the same direction** as you thought? For example, you expect DAU and average length of time users use your app both increase. However you observe DAU decrease while average length of time increase. WTH??

You should **dive deeper, and figure out why**. And this is also why people usually want to have one OEC (Overall Evaluation Criterion). A good OEC gives you a balance between short-term and long-term goal, or the balance between different metrics. However you also need to keep in mind, having an OEC helps you understand what your

business cares about, and how do you balance metrics such as stay time and click, but it does not help you make a product change decision.

Last step, draw conclusions.

If you have a significant result from the test…There comes two questions: Do you understand the change? Do you want to launch the change? What if your change has a positive impact on one slice of users, but no impact or negative impact for other slices of users? Do you understand WHY?? (e.g. Bolded words in English vs. in Chinese have different test results, because bolded Chinese is hard to read).



Then how do I decide whether to launch the change or not?

Ask yourself a few questions: Do I have statistically significant and practically significant result in order to justify the change? Do I understand what the change actually done to our user experience? Last but not the least, is it worth it to launch?

. . .

Gotchas

Always do a ramp up when you want to launch a change after the A/B test. Because you want to know if there's any incidental impact to unaffected users that you didn't test in the original experiment.

When you are ramping up the change, you may see the effect flatten out. Thus making the tested effect not repeatable. There are many reasons for this phenomenon.

1. **Seasonality effect**: Social network platform user behavior changes a lot when students start summer vacation or going back to school. Holidays affect users' shopping behavior a lot. Solution: use hold-back method , launch the change to everyone except for one small hold-back group of users, and continue comparing their behavior to the control group.
2. **Novelty effect** or **change aversion**: cohort analysis may be helpful.

. . .

Lessons learned

1. Double check, triple check that your experiment is set up properly.
2. Not only think about statistically significant but also business impact. Think about the engineering cost, customer support or sales issue, what's the opportunity cost, etc.
3. If you are running your first experiment that have a big impact, you might want to run a couple of experiments and check the results to see if you are comfortable launching it.

. . .

Other thing to consider: Politics and ethics for experiments

Experiments involves real people, it is important to protect the users and follow the ethics. However, there were many problematic examples of experiments in the past. For

example, Tuskegee syphilis experiment, Milgram experiment in history and a recent Facebook emotion experiment. To conduct an A/B test in an ethical way, there are four principles to keep in mind:

1.Risk: What risk are the participants exposed to?

The main threshold is whether the risk exceeds that of "minimal risk". Minimal risk is defined as the probability and magnitude of harm that a participant would encounter in normal daily life. The harm considered encompasses physical, psychological and emotional, social, and economic concerns. If the risk exceeds minimal risk, then informed consent is required.

2. Benefit: What's the potential benefit of the outcome of the study?

It is important to be able to state what the benefit would be from completing the study.

3. Choice: What other choices do participants have?

In online experiments, the issues to consider are what the other alternative services that a user might have, and what the switching costs might be, in terms of time, money, information, etc.

4. Privacy: What privacy do participants have?

For new data being collected and stored, how sensitive is the data and what are the internal safeguards for handling that data? Then, for that data, how will it be used and how will participants' data be protected? How are participants guaranteed that their data, which was collected for use in the study, will not be used for some other purpose?

. . .

This summary is just a brief overview of the content of the course, if you are interested in A/B testing or preparing for interviews, I highly recommend you take this course, and do the final project. The most efficient way to master something is always learning by doing.