**The Joy of A/B Testing: Theory, Practice, and Pitfalls**

**How today’s tech companies make data-driven decisions in Machine Learning production**



Photo by [Marcelo Leal](https://unsplash.com/@marceloleal80?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

A/B testing is is deeply ingrained in modern tech companies, enabling them to continuously improve their product in order to stay on top of consumer preferences and beat the competition. A [Lyft](https://eng.lyft.com/challenges-in-experimentation-be9ab98a7ef4) article states that:

The norm is to test each and every product change, to build up evidence to drive large decisions, and to use causal data to support the strategic direction.

[DoorDash](https://doordash.engineering/2022/05/24/meet-dash-ab-the-statistics-engine-of-experimentation-at-doordash/) writes:

For any data-driven company, it’s key that every change is tested by experiments to ensure that it has a positive measurable impact on the key performance metrics.

Or consider this quote from [Netflix](https://medium.com/netflix-techblog/reimagining-experimentation-analysis-at-netflix-71356393af21):

Netflix runs on an A/B testing culture: nearly every decision we make about our product and business is guided by member behavior observed in test.

A/B testing is particularly important in **Machine Learning production** because the model performance can have enormous impact on the business metrics and user experience. For example, a tiny improvement in conversion rates in an ads ranking model can result in Millions of dollars in additional revenue. We need to A/B test these model changes because offline performance on a static test set is not a good indicator for production performance: sometimes a model may perform better in an offline but [worse in production](https://towardsdatascience.com/is-my-model-really-better-560e729f81d2).

In this post, we’ll take a deep-dive into the theory of A/B testing of ML models, best practices, and common pitfalls. You’ll learn:

* how to split your traffic for A/B testing,
* what it means for the test result to be statistically significant,
* the possible test results and the 4 types of test errors, and
* common pitfalls to avoid.

Let’s jump in.

**How to split your traffic for A/B testing**

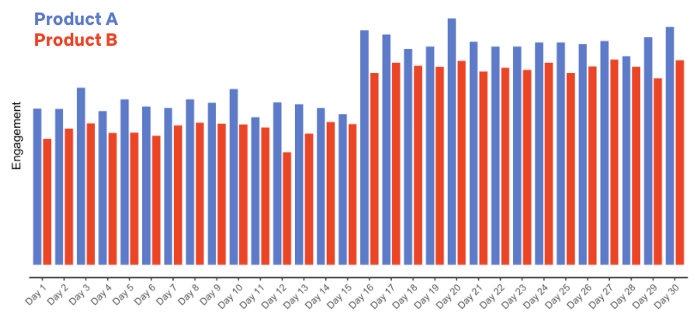
The basic idea behind A/B testing is to split the user population into a treatment group and a control group. For testing new ML production models, you’ll expose the treatment group to the new model, and the control group to the existing model.

A/B testing rests on 2 important statistical assumptions, namely **identity** and **independence**:

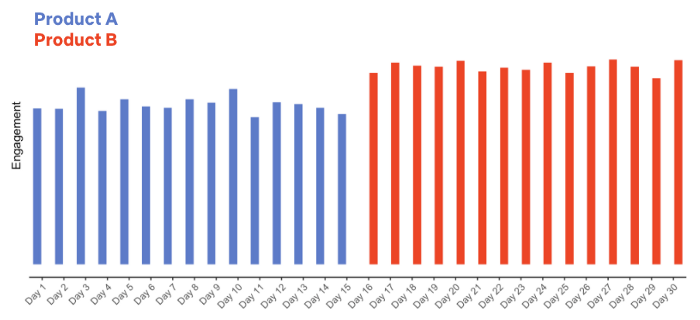
* **identity**: the statistical properties of the control and treatment groups are identical. When you have similar groups, then you can be confident that any change in the group’s metrics can be attributed to the treatment, and nothing else. In statistical terms, you have controlled for all **confounders**.
* **independence**: the two groups are independent from each other. This means that the user behavior in the control group has no impact on the user behavior in the treatment group, and vice versa. If it does, then we’ll have introduced an additional confounder that makes it harder to interpret the test results.

The most common approach in A/B testing is the **population split**, where we randomly split users into control and treatment. In practice, this can be done by hashing a user id into an integer that determines its group. Alternatively, we can use a **session split**, where we assign a user to either of the two groups at the start of each engagement session.

To see why we use population split, consider an alternative, namely a **time split**. Suppose we have two movie recommendation models as shown in red and blue below. A population split would clearly show us that the blue model drives more engagement compared to the red model, as shown in the top plot. However, if we had done a time split on Day 16, we might have falsely concluded that the red model was better. In statistical terms, the time split violates the identity assumption, and the test results are therefore inconclusive.In other words, population split allows us to determine **causality**, and not just correlation.



A/B testing over 30 days with a population split (source: [Netflix](https://netflixtechblog.com/what-is-an-a-b-test-b08cc1b57962))



A/B testing over the same 30 days with a time split (source: [Netflix](https://netflixtechblog.com/what-is-an-a-b-test-b08cc1b57962))

Lastly, how large should the treatment and control groups be? In practice, we can change the sizes over time. If we’re initally not certain that the new model will bring an improvement, we can start with a treatment group size of 1% or even lower, and then gradually ramp up the traffic if the test results look good.

**Statistical significance in A/B testing**

Generally, the question we need to ask in any A/B test is whether the result (e.g., the new model causes a 1% improvement in ads click-through rate) could have been caused by chance. This is the **Null hypothesis**, i.e. the hypothesis that there’s no effect from the treatment at all. If the A/B test result is statistically significant, then we can confidently rule out the Null hypothesis.

Statistical significance is calculated using the **p-value**, the probability that the outcome we’re seeing is the result of chance. Prior to the experiment, we need to set a p-value threshold below which we would rule out the Null hypothesis, typically 0.05 or 0.01. For example, if the 1% improvement in click-through rate has a p-value of 0.02 and we’ve set a threshold of 0.05, then we can conclude that A/B test result is statistically significant (and rule out the Null hypothesis of no effect).

The p-value can be computed with a **permutation analysis**: after the A/B test is completed, shuffle the group assignments many times and measure the test result in each permutation. Then, calculate the ratio of permutations in which you’ve seen a result at least as extreme as the actual (unshuffled) A/B test result. This number is the p-value. In practice, you don’t have to do this yourself, you can use an A/B test [calculator](https://www.surveymonkey.com/mp/ab-testing-significance-calculator/) instead.

**Test results and the 4 types of test errors**

There can be 3 possible outcomes from the A/B test:

* you confirm that the new model is significantly better than the existing model. In this case, you’ll want to roll out the new model to the entire population, so that you can benefit from the model improvement.
* you find the opposite, namely that the new model is significantly worse than the existing model. In that case you’ll want to abort the experiment and go back to offline model development.
* the test result is inconclusive: the difference between the two models is not statistically significant. In that case you can either abort the experiment, or run it longer to collect more data.

Importantly, A/B tests aren’t perfect, and therefore can lead to erroneous results. Generally we distinguish between [4 types of errors](https://netflixtechblog.com/experimentation-is-a-major-focus-of-data-science-across-netflix-f67923f8e985):

**Type I error** (aka false positive): the models perform equally well, but the A/B test still produces a statistically significant result. As a consequence, you may roll out a new model that doesn’t really perform better. It’s a false positive. You can control the prevalence of this type of error with the p-value threshold. If your p-value threshold is 0.05, then you can expect a Type I error in about 1 in 20 experiments, but if it’s 0.01, then you only expect a Type I error in only about 1 in 100 experiments. The lower your p-value threshold, the fewer Type I errors you can expect.

**Type II error** (aka false negative): the new model is in fact better, but the A/B test result is not statistically significant. In statistical terms, your test is **underpowered**, and you should either collect more data, choose a more sensitive metric, or test on a population that’s more sensitive to the change.

**Type S error** (sign error): the A/B test shows that the new model is significantly better than the existing model, but in fact the new model is worse, and the test result is just a statistical fluke. This is the worst kind of error, as you may roll out a worse model into production which may hurt the business metrics.

**Type M error**: (magnitude error): the A/B test shows a much bigger performance boost from the new model than it can really provide, so you’ll over-estimate the impact that your new model will have on your business metrics.

**Common pitfalls**

Lastly, here are some common pitfalls in A/B testing you should be aware of.

**HARKing (hypothesizing after the results are known)**. Before running an A/B test, you should always have a hypothesis. This can be something simple like ‘*I hypothesize the new model to be better because it’s trained on more features*’. If you randomly test a large number of models without hypothesizing first, you’re exposing yourself to the [look-elsewhere effect](https://towardsdatascience.com/the-statistics-of-the-improbable-cec9a754e0ff), i.e. the risk that you may find a statistically significant result that is just the result of chance alone.

If you test A vs. B without a clear hypothesis, and B wins by 15%, that’s nice, but what have you learned? Nothing. — [Peep Laja](https://cxl.com/blog/12-ab-split-testing-mistakes-i-see-businesses-make-all-the-time/#early)

**Analysis paralysis.** It’s also important to agree on the success criterion in the beginning of the experiment, and avoid diverting into lengthy analysis after the A/B test is completed. For example, suppose you’ve trained a global e-commerce product classification model, and show in an A/B test that it works significantly better (in terms of precision at fixed recall) than the current production model. Great! But now the product team is asking how the comparison looks like on each of the 50 market segments the model operates in, and the 20 product types in each. How well are we doing for each of these 1,000 micro-segments? Examining long lists of metrics slows down decision making and increases the chance of finding statistical flukes in the data. After all, the A/B test result is just a noisy proxy for the true production performance.

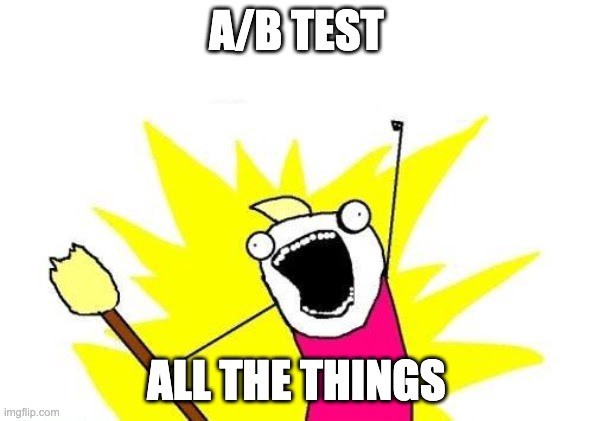
**Spillover effects.** Spillover effects, also known as cross-contamination, happen when the treatment group has a secondary effect on the control group, violating the independence assumption. For example, consider an ads ranking model that ranks a particular ad (let’s say for Nike shoes) much higher than the existing production model, and drives high conversion for that ad. Then, users in the treatment group may rapidly use up all of the advertiser’s budget (in this case, Nike’s) for ads impressions, leaving a smaller budget left for the control group. The treatment group therefore has an indirect impact on the behavior of the control group, causing a spillover effect.

**Random assignment may fail to distribute heavy users equally.** If your user population contains a few users that create a large amount of user activity, then a random A/B assignment is not guaranteed to distribute these users equally. This may violate the identity assumption and make the test results mode difficult to interpret.

**Unaligned trade-offs.** Your A/B test may result in a situation where one metric improves but another metric becomes worse, resulting in a conflict between different business goals. For example, suppose you’ve built a news ranking model that penalizes (down-ranks) clickbait titles. User satisfaction may be better as users find more engaging content, but overall click volume may go down.

**Treatment self-selection.** If users themselves can opt into a treatment group, the A/B test violates the identity assumption: the two groups are not identical, but instead the treatment group consists of a particular subset of users, namely those more willing to take part in experiments. In that case we will not be able to tell whether the difference between the groups is due to the treatment or due to the sample differences.

**Not A/B testing.** Lastly, another pitfall is to not A/B test at all. Suppose you work on a new credit card fraud detection model, and instead of A/B testing the new model, in production you simply evaluate the transactions first with the old model and then with the new model, and consider a transaction to be fraud if either model produces a high enough score. This may sound like a conservative approach, but it’s not scalable: what if you build a new version of the model in a month, and yet another one a month later, and so on? Over time all these models will create an overwhelming amount of maintenance and infrastructure costs.



Source: [Netflix](https://netflixtechblog.com/netflix-a-culture-of-learning-394bc7d0f94c)

**Conclusion**

To summarize:

* A/B testing is a critical part in ML production. Small changes in model performance can have massive impact on the business metrics, and we only want to roll out new models if we’re certain that they’re better. [Offline tests aren’t conclusive](https://towardsdatascience.com/is-my-model-really-better-560e729f81d2).
* Use a population split to divide your users into control and treatment groups, and use the existing model for the control and the new model for the treament. Time splits lead to inconclusive results because we cannot know whether the change is driven by the new model or by something else.
* The result from an A/B test is statistically significant if its p-value if below the threshold you’ve set prior to the experiment, e.g. 0.01. If the test result is statistically significant, you can be confident that the new model is truly better. The p-value can be computed with a permutation analysis.
* The test results can have 4 types of errors: Type I (false positive), Type II (false negative), Type S (sign error), and Type M (magnitude error). Of these, the Type S error is the most severe, as it may lead to an inferior model being rolled out into production.
* Be aware of common pitfalls such as HARKing, analysis paralysis, spillover effects, unequal distribution of heavy users, unaligned tradeoffs, treatment self-selection, and avoiding A/B tests in the first place.

**Before you go…**

*Enjoy this content? Follow/subscribe on Medium so you don’t miss new posts I write in the future.* [*Become a Medium member*](https://medium.com/@samuel.flender/membership) *so you can view unlimited articles. And make sure to follow me on* [*LinkedIn*](https://www.linkedin.com/in/sflender/) *and/or* [*Twitter*](https://twitter.com/samflender)*!*

**References for diving deeper:**

* [Challenges in experimentation](https://eng.lyft.com/challenges-in-experimentation-be9ab98a7ef4) (Lyft)
* [What is an A/B test?](https://netflixtechblog.com/what-is-an-a-b-test-b08cc1b57962) (Netflix)
* [Experimentation is a major focus of Data Science across Netflix](https://netflixtechblog.com/experimentation-is-a-major-focus-of-data-science-across-netflix-f67923f8e985) (Netflix)
* [Interpreting A/B test results: false positives and statistical significance](https://netflixtechblog.com/interpreting-a-b-test-results-false-positives-and-statistical-significance-c1522d0db27a) (Netflix)
* [Interpreting A/B test results: false negatives and power](https://netflixtechblog.com/interpreting-a-b-test-results-false-negatives-and-power-6943995cf3a8) (Netflix)
* [Statistical Significance in A/B testing (and How People Misinterpret Probability)](https://data36.com/statistical-significance-in-ab-testing/)
* [8 Common Pitfalls of Running A/B Tests](https://medium.com/towards-data-science/online-controlled-experiment-8-common-pitfalls-and-solutions-ea4488e5a82e)