5 ways to Increase Statistical Power

Statistical Power in A/B testing visualized

Alison Yuhan Yao · Follow
Published in Towards Data Science · 9 min read · Nov 29, 2021

In Data Science, we often need to run A/B tests and interpret the results using statistical power. In the blog, I will explain what power is and how to increase power using visualization. And I emphasize that these methods have nothing to do with p-hacking.

Before we start, I'm assuming you have some basic knowledge about:

1. Hypothesis Testing (p-value)
2. Type I and Type II errors, or false negative and false positive

What is Power

In a binary hypothesis test, either the null hypothesis H0 is true, or the alternative hypothesis HA is true. Typically, we set the null as not having any effect or things staying the same, such as two means being the same (mu1 = mu2). And we set the alternative hypothesis as having some sort of effect or things change after you introduce a variable into the experiment, such as two means are not the same (mu1 ≠ mu2).

And for us testers, we don't want H0 to be true. We don't want to go through all the trouble implementing tests only to find out the variables we introduced are useless. Instead, we care about HA being true because we want to find significant variables, so we need to care about the probability that we are right: the probability that we reject H0 and therefore accept HA when HA is indeed true. And that is called statistical power, your ability to detect an effect when it is there.

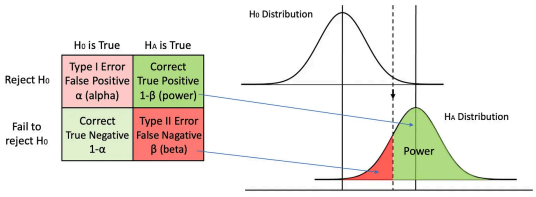


Image by Author

In this graph, we visualize power in darker green. In a hypothesis test, we always determine the alpha value beforehand, which is usually set at 0.05, so the Type I error rate is set in stone before we even begin the test. Then, we can calculate the minimum critical value we need to reject H0. We can draw a line from the null hypothesis distribution down to the alternative hypothesis distribution and separate the area under curve into two pieces. If our calculated t or z value falls on the left of the dashed line, we fail to reject H0 when HA is true and we make a Type II error. If the calculated value falls on the right, we reject H0 when HA is true and we make the right call. Therefore, the area on the right of the curve is our power.

Please note that we only talk about power when HA is true. If unfortunately, H0 is true and there is no effect whatsoever, no amount of power is going to help us. As you can see in the graph, when H0 is true, we deal with alpha only. But in real life, we have no idea if H0 or HA is true and we cannot change the ground truth. All we can control is to reject or not reject H0 while hoping HA is true. Despite that, we still want to increase our statistical power, so that we have the best chance of detecting an effect when it is indeed there.

How to Increase Power

Let's take a closer look at the visualization and understand what kind of information we know in the graph.

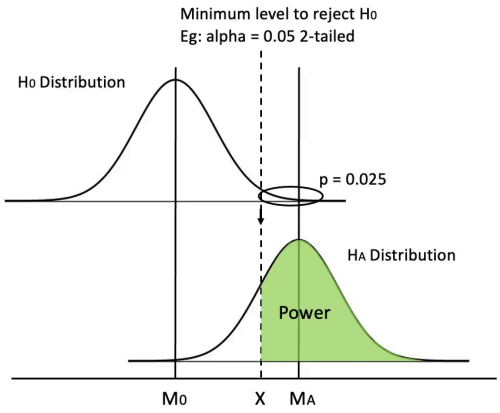


Image by Author

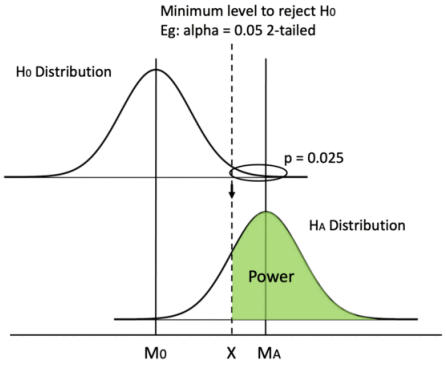
We know before start: (1) alpha level, (2) whether we use a 1-tailed or a 2-tailed test, (3) shape of the distribution (t or z). After doing the test, we know: (4) sample size, (5) means of the distributions M0 and MA (or the difference between them), (6) standard deviation of the distributions and therefore (7) standard error (SE) and (8) min critical value, aka the value of dashed line X (using M0, min critical value and SE).

An **example** can be testing if doing every SAT question in a practice book actually makes a difference. The population mean for SAT is M0 = 500 and population SD = 100. Since we know the population mean, we can use a z distribution. We can set the test as 2-tailed (alpha = 0.05) as usual. Now, we do our test and check the test data. Let's say N = 100, mean for the HA distribution is MA = 530. Since we assume the two distributions are the same, we can calculate that SE = SD/√N = 100/√100 = 10 and therefore X = z_crit * SE + M0 = 1.96 (we know that from a z table) * 10 + 500 = 519.6. Okay, now we are finally ready to calculate the power. The critical value for HA distribution is z_crit2 = (X-MA)/SE = (519.6-530)/10 = -1.04. We can check the z table and see that the corresponding p(x ≤ -1.04) = 0.1492, so the power is 1-0.1492 = 0.8508.

Okay, now the big question. How do we increase power?

I believe it's easier to **think visually**. How would you increase the area in green? We can shift the dashed line (method 0 & 1), move the mean MA (method 2), or change the shape of the distribution (method 3–5).

0. Raise significance level alpha (the WRONG way)

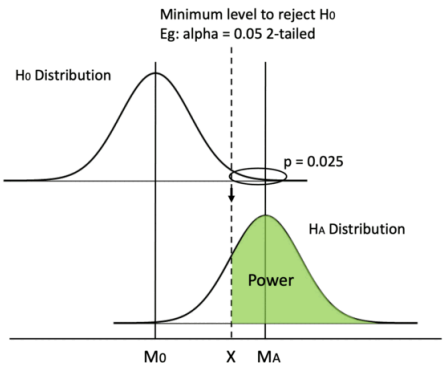


GIF by Author

Keeping everything else in place, one straightforward method is moving the dashed line to the left. By increasing alpha or switching from a 2-tailed test to a 1-tailed test, we can decrease the min critical value X.

Easy peasy, but one big **caveat** is that increasing alpha is going to result in a higher probability of making a type I error, so we never tamper with the value of alpha.

1. Switch from a 2-tailed test to a 1-tailed test



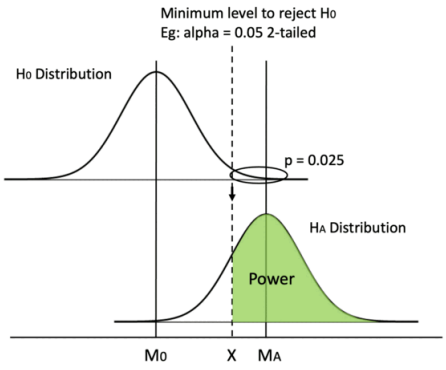
GIF by Author

Similarly, switching from a 2-tailed test to a 1-tailed test can move the dashed line to the left. In a 2-tailed test, the critical p value for each tail is half of alpha, while the critical p equals alpha in a 1-tailed test.

Whether we use a 2-tailed or a 1-tailed test depends on test design and is set before running the experiment, so we need to keep in mind from the beginning to favor a 1-tailed test.

2. Increase mean difference

Another way is to increase the difference between the two means. Since the H0 distribution is fixed because the null hypothesis generally doesn't change, we can only hope that the HA distribution can shift to the right.



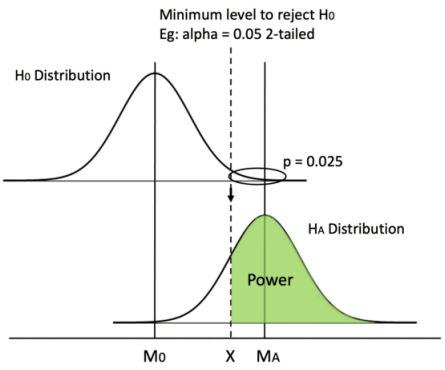
GIF by Author

However, we cannot manually achieve that. We need to report the distribution we get from the test data as it is (otherwise, it is data manipulation!). But this quantifies why we like a bigger mean difference, in that it gives us more power.

Special Note:
Mean difference is one type of effect size, so we can say increasing the effect size can increase power.

3. Use z distribution instead of t distribution

Starting from this method, we focus on modifying the shape of the distributions. Using a z distribution will make it easier to reach statistical significance. Why?



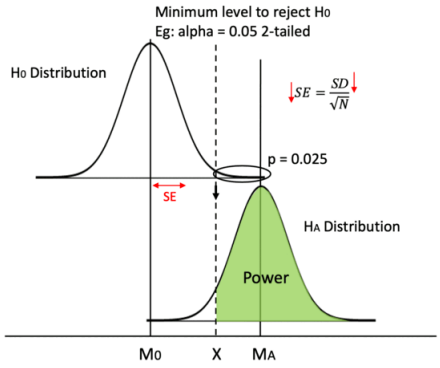
GIF by Author

Because a z distribution has a larger kurtosis (z is taller than t distribution) and has thinner tails. By changing the shape, X decreases as well, so the two effects increase power. In the previous example of the SAT score, using z distribution is appropriate because we know the population mean. Although this does not happen often in real life, we can use z to approximate a t distribution when the sample size is large.

This change in shape is not going to be very significant though. I exaggerated a bit in the GIF so that it is easier to see.

4. Decrease standard deviation

Another method to modify the shape of the distribution is to reduce SE by decreasing SD.

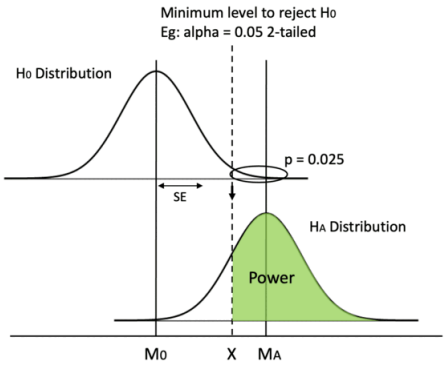


GIF by Author

We can use more precise measurements like asking clear questions instead of vague ones in user feedback surveys. That way, our data is going to have less error and less noise.

Also, we can try to run a paired samples t-test if it applies. The gist of why it works is that paired groups have a higher correlation and the difference in scores between paired samples are smaller.

5. Increase sample size (the most practical way)



GIF by Author

Finally, the one method we can always rely on is increasing sample size. If SD remains the same, an bigger sample size N increases the denominator, then SE will decrease.

This is also the most practical way in real business scenarios. It is much easier to collect data from more users or send out more surveys than following the methods listed above.

This is not p-hacking

If your first thought of reading this blog is p-hacking, congratulations! You are a well-trained statistics practitioner aware of experimental ethics.

Indeed, the word "increase" gives us the illusion that we are in control of the experiment outcome, while in fact we are not. Method (2) increase mean difference, for example, is outside of our control, but I think it is still great to understand the mechanism of why a bigger mean difference lead to a larger power. So I put it here, as do other posts.

Although we cannot control the test data, **we can control the experimental design**. Methods 0, 1, 3, 4 and 5 are all pre-determined **before** we conduct the tests and see the test data. We never ever manipulate test data!

For example, let's talk about (5) increasing sample size. Do we increase sample size during or after experiments? No! We calculate the sample size in advance based on a desired power (eg: power=0.85). If N=50, then let's get 50 people, conduct the experiment, check the test data and report the p-value as it is.

If I were to p-hack, I will first check my p-value = 0.053 after getting 50 people. Oh no. p> 0.05, so close. What do I do? I increase the sample size after conducting the test and find another 2 participants. I check my p-value again. Now p = 0.049 and I stop everything and report that I have found an effect. This is very wrong! If I check my p-value every time I add a new participant, anything has the probability of getting a $p < 0.05$ and I will always find an effect.

Therefore, the meaning of increasing sample size is using N=50 instead of 20 or 40 **before** conducting the test. The same applies to methods 1, 3 and 4.

It is important to know that there is more than one way to increase statistical power because sometimes increasing sample size is simply unaffordable. Speaking from the perspective of a university student, if my fundings have been approved and all I can afford is 50 test subjects, I will have to resort to a smarter experimental design in other ways to increase statistical power.

(This section was added 2 days after publishing the original blog. I see some confusion in the comment section, so hopefully this is clearer.)

Conclusion

In this post, we talked about what statistical power is using visualization, went through an example to understand the graph better, and talked about 5 ways (6 ways really) to increase power:

- Raise significance level alpha (the WRONG way)
- Switch from a 2-tailed test to a 1-tailed test
- Increase mean difference
- Use z distribution instead of t distribution
- Decrease standard deviation (by using precise measurements & paired samples t-test)
- Increase sample size (the most practical way)

. . .

Thank you for reading! I hope this blog has been helpful to you.

Data Science | Ab Testing | Statistical Power | Power Analysis | Editors Pick

Written by Alison Yuhan Yao

310 Followers · Writer for Towards Data Science

Data Science | GitHub: https://github.com/AlisonYao LinkedIn: https://www.linkedin.com/in/yuhanyao/

Follow

More from Alison Yuhan Yao and Towards Data Science

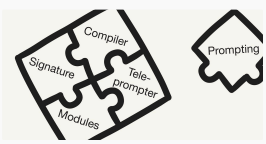
Alison Yuhan Yao in CodeX

Linear Regression for Causal Inference

A deeper dive into correlation vs causation.

6 min read · Feb 6, 2022

144



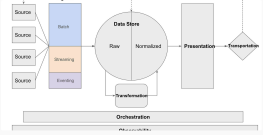
Leonie Monigatti in Towards Data Science

Intro to DSPy: Goodbye Prompting, Hello Programming!

How the DSPy framework solves the fragility problem in LLM-based applications by...

13 min read · Feb 27, 2024

3.3K 10



Dave Melillo in Towards Data Science

Building a Data Platform in 2024

How to build a modern, scalable data platform to power your analytics and data science...

9 min read · Feb 5, 2024

2.4K 33

Alison Yuhan Yao in Towards AI

Forecasting Time Series Data: Netflix Stock Price Prediction

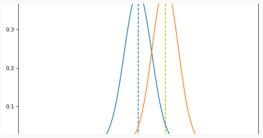
ARIMA-(G)ARCH models with MiniTab and R

8 min read · May 31, 2022

97 1

See all from Alison Yuhan Yao | See all from Towards Data Science

Recommended from Medium



Mark Eltsefon in Towards Data Science

Common Mistakes During A/B Testing

Our path to excellence is paved with mistakes. Let's make them!

· 5 min read · Apr 24, 2022

359 3

Oham Ugochukwu in Towards AI

End-to-End Experimental Design using A/B Test

A data-driven guide to decision-making and product design

· 6 min read · Nov 6, 2023

87 1

Lists

Predictive Modeling w/ Python

20 stories · 1013 saves

Practical Guides to Machine Learning

10 stories · 1219 saves

Coding & Development

11 stories · 515 saves

ChatGPT prompts

47 stories · 1291 saves

Thauri Dattadeen

Selecting the right Statistical Tests for Effective A/B Testing

A guide on ensuring your analysis is effective and trustworthy!

11 min read · Dec 28, 2023

1 1

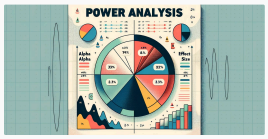
Wenkai Bao in Expedia Group Technology

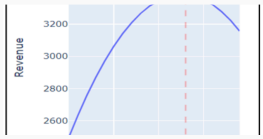
How to Size For Online Experiments With Ratio Metrics

Sizing Derivation and Evaluation

8 min read · Mar 28, 2023

170 4





Data Science Shower Thoughts

Behind the Curtain of Sample Size: Unmasking the Roles of Alpha,...

Introduction

8 min read · Oct 20, 2023

70

Ismetgocer in Academy Team

Price Optimization with Machine Learning: The Impact of Data...

Nowadays, the impact of data science and especially machine learning in the business...

10 min read · Dec 11, 2023

219

See more recommendations