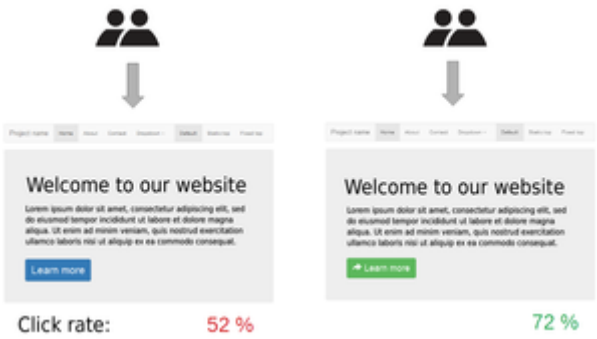
# A/B testing

**A/B testing** (also known as **bucket testing** or **split-run testing**) is a user experience research methodology.[1] A/B tests consist of a randomized experiment with two variants, A and B.[2][3] It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.[4]




Example of A/B testing on a website. By randomly serving visitors two versions of a website that differ only in the design of a single button element, the relative efficacy of the two designs can be measured.

## Contents



## Overview

A/B test is the shorthand for a simple controlled experiment.[5] As the name implies, two versions (A and B) of a single variable are compared, which are identical except for one variation that might affect a user's behavior. A/B tests are widely considered the simplest form of controlled experiment. However, by adding more variants to the test, this becomes more complex.[6]

A/B tests are useful for understanding user engagement and satisfaction of online features, such as a new feature or product.[7] Large social media sites like LinkedIn, Facebook, and Instagram use A/B testing to make user experiences more successful and as a way to streamline their services.[7]

Today, A/B tests are being used to run more complex experiments, such as network effects when users are offline, how online services affect user actions, and how users influence one another.[7] Many jobs use the data from A/B tests. This includes, data engineers, marketers, designers, software engineers, and entrepreneurs.[8] Many positions rely on the data from A/B tests, as they allow companies to understand growth, increase revenue, and optimize customer satisfaction.[8]

Version A might be the currently used version (control), while version B is modified in some respect (treatment). For instance, on an e-commerce website the purchase funnel is typically a good candidate for A/B testing, as even marginal decreases in drop-off rates can represent a significant gain in sales. Significant improvements can sometimes be seen through testing elements like copy text, layouts, images and colors,[9] but not always. In these tests, users only see one of two versions, as the goal is to discover which of the two versions is preferable.[10]

Multivariate testing or multinomial testing is similar to A/B testing, but may test more than two versions at the same time or use more controls. Simple A/B tests are not valid for observational, quasi-experimental or other non-experimental situations, as is common with survey data, offline data, and other, more complex phenomena.

A/B testing has been marketed by some as a change in philosophy and business strategy in certain niches, though the approach is identical to a between-subjects design, which is commonly used in a variety of research traditions.[11][12][13] A/B testing as a philosophy of web development brings the field into line with a broader movement toward evidence-based practice. The benefits of A/B testing are considered to be that it can be performed continuously on almost anything, especially since most marketing automation software now typically comes with the ability to run A/B tests on an ongoing basis.

### Common test statistics

"Two-sample hypothesis tests" are appropriate for comparing the two samples where the samples are divided by the two control cases in the experiment. Z-tests are appropriate for comparing means under stringent conditions regarding normality and a known standard deviation. Student's t-tests are appropriate for comparing means under relaxed conditions when less is assumed. Welch's t test assumes the least and is therefore the most commonly used test in a two-sample hypothesis test where the mean of a metric is to be optimized. While the mean of the variable to be optimized is the most common choice of estimator, others are regularly used.

For a comparison of two binomial distributions such as a click-through rate one would use Fisher's exact test.

| Assumed Distribution | Example Case | Standard Test | Alternative Test |
|---|---|---|---|
| Gaussian | Average revenue per user | Welch's t-test (Unpaired t-test) | Student's t-test |
| Binomial | Click-through rate | Fisher's exact test | Barnard's test |
| Poisson | Transactions per paying user | E-test[14] | C-test |
| Multinomial | Number of each product purchased | Chi-squared test | |
| Unknown | | Mann–Whitney U test | Gibbs sampling |

## History

Like most fields, setting a date for the advent of a new method is difficult. Experimentation with advertising campaigns, which has been compared to modern A/B testing, began in the early twentieth century.[15] The advertising pioneer Claude Hopkins used promotional coupons to test the effectiveness of his campaigns. However, this process, which Hopkins described in his Scientific Advertising, did not incorporate concepts such as statistical significance and the null hypothesis, which are used in statistical hypothesis testing.[16] Modern statistical methods for assessing the significance of sample data were developed separately in the same period. This work was done in 1908 by William Sealy Gosset when he altered the Z-test to create Student's t-test.[17][18]

With the growth of the internet, new ways to sample populations have become available. Google engineers ran their first A/B test in the year 2000 in an attempt to determine what the optimum number of results to display on its search engine results page would be.[4] The first test was unsuccessful due to glitches that resulted from slow loading times. Later A/B testing research would be more advanced, but the foundation and underlying principles generally remain the same, and in 2011, 11 years after Google's first test, Google ran over 7,000 different A/B tests.[4]

In 2012, a Microsoft employee working on the search engine Bing created an experiment to test different ways of displaying advertising headlines. Within hours, the alternative format produced a revenue increase of 12% with no impact on user-experience metrics.[3] Today, companies like Microsoft and Google each conduct over 10,000 A/B tests annually.[3]

Many companies now use the "designed experiment" approach to making marketing decisions, with the expectation that relevant sample results can improve positive conversion results. It is an increasingly common practice as the tools and expertise grow in this area.

# Examples

## Email marketing

A company with a customer database of 2,000 people decides to create an email campaign with a discount code in order to generate sales through its website. It creates two versions of the email with different call to action (the part of the copy which encourages customers to do something — in the case of a sales campaign, make a purchase) and identifying promotional code.

- To 1,000 people it sends the email with the call to action stating, "Offer ends this Saturday! Use code A1",
- and to another 1,000 people it sends the email with the call to action stating, "Offer ends soon! Use code B1".

All other elements of the emails' copy and layout are identical. The company then monitors which campaign has the higher success rate by analyzing the use of the promotional codes. The email using the code A1 has a 5% response rate (50 of the 1,000 people emailed used the code to buy a product), and the email using the code B1 has a 3% response rate (30 of the recipients used the code to buy a product). The company therefore determines that in this instance, the first Call To Action is more effective and will use it in future sales. A more nuanced approach would involve applying statistical testing to determine if the differences in response rates between A1 and B1 were statistically significant (that is, highly likely that the differences are real, repeatable, and not due to random chance).[19]

In the example above, the purpose of the test is to determine which is the more effective way to encourage customers to make a purchase. If, however, the aim of the test had been to see which email would generate the higher click-rate – that is, the number of people who actually click onto the website after receiving the email – then the results might have been different.

For example, even though more of the customers receiving the code B1 accessed the website, because the Call To Action didn't state the end-date of the promotion many of them may feel no urgency to make an immediate purchase. Consequently, if the purpose of the test had been simply to see which email would bring more traffic to the website, then the email containing code B1 might well have been more successful. An A/B test should have a defined outcome that is measurable such as number of sales made, click-rate conversion, or number of people signing up/registering.[20]

### Product Pricing A/B testing

A/B testing can be used to determine the right price for the product, as this is perhaps one of the most difficult tasks when a new product or service is launched.

A/B testing (especially valid for digital goods) is an excellent way to find out which price-point and offering maximize the total revenue.

### Political A/B testing

A/B tests are used for more than corporations, but are also driving political campaigns. In 2007, Barack Obama's presidential campaign used A/B testing as a way to garner online attraction and understand what voters wanted to see from the presidential candidate.[21] For example, Obama's team tested four distinct buttons on their website that led users to sign up for newsletters. Additionally, the team used six different accompanying images to draw in users. Through A/B testing, staffers were able to determine how to effectively draw in voters and garner additional interest.[21]

## Segmentation and targeting

A/B tests most commonly apply the same variant (e.g., user interface element) with equal probability to all users. However, in some circumstances, responses to variants may be heterogeneous. That is, while a variant A might have a higher response rate overall, variant B may have an even higher response rate within a specific segment of the customer base.[22]

For instance, in the above example, the breakdown of the response rates by gender could have been:

| Gender | Overall | Men | Women |
|---|---|---|---|
| **Total sends** | 2,000 | 1,000 | 1,000 |
| **Total responses** | 80 | 35 | 45 |
| **Variant A** | $\frac{50}{1{,}000}$ (5%) | $\frac{10}{500}$ (2%) | $\frac{40}{500}$ (8%) |
| **Variant B** | $\frac{30}{1{,}000}$ (3%) | $\frac{25}{500}$ (5%) | $\frac{5}{500}$ (1%) |

In this case, we can see that while variant A had a higher response rate overall, variant B actually had a higher response rate with men.

As a result, the company might select a segmented strategy as a result of the A/B test, sending variant B to men and variant A to women in the future. In this example, a segmented strategy would yield an increase in expected response rates from $5\% = \frac{40+10}{500+500}$ to $6.5\% = \frac{40+25}{500+500}$ – constituting a 30% increase.

It is important to note that if segmented results are expected from the A/B test, the test should be properly designed at the outset to be evenly distributed across key customer attributes, such as gender. That is, the test should both (a) contain a representative sample of men vs. women, and (b) assign men and women randomly to each "variant" (variant A vs. variant B). Failure to do so could lead to experiment bias and inaccurate conclusions to be drawn from the test.[23]

This segmentation and targeting approach can be further generalized to include multiple customer attributes rather than a single customer attribute – for example, customers' age *and* gender – to identify more nuanced patterns that may exist in the test results.

## See also

- Adaptive control
- Choice modelling
- Multi-armed bandit
- Multivariate testing
- Randomized controlled trial
- Scientific control
- Test statistic

## References


1. Young, Scott W. H. (2014). "Improving Library User Experience with A/B Testing: Principles and Process" (https://doi.org/10.3998%2Fweave.12535642.0001.101). *Weave: Journal of Library User Experience*. **1** (1). doi:10.3998/weave.12535642.0001.101 (https://doi.org/10.3998%2Fweave.12535642.0001.101). hdl:2027/spo.12535642.0001.101 (https://hdl.handle.net/2027%2Fspo.12535642.0001.101). ISSN 2333-3316 (https://www.worldcat.org/issn/2333-3316).
2. Kohavi, Ron; Longbotham, Roger (2017). "Online Controlled Experiments and A/B Tests" (http://www.exp-platform.com/Documents/2015%20Online%20Controlled%20Experiments_EncyclopediaOfMLDM.pdf) (PDF). In Sammut, Claude; Webb, Geoff (eds.). *Encyclopedia of Machine Learning and Data Mining*. Springer.
3. Kohavi, Ron; Thomke, Stefan (September 2017). "The Surprising Power of Online Experiments" (https://hbr.org/2017/09/the-surprising-power-of-online-experiments). *Harvard Business Review*: 74–82.
4. "The ABCs of A/B Testing - Pardot" (http://www.pardot.com/blog/abcs-ab-testing/). *Pardot*. Retrieved 2016-02-21.
5. Young, Scott W. H. (2014). "Improving Library User Experience with A/B Testing: Principles and Process" (https://doi.org/10.3998%2Fweave.12535642.0001.101). *Weave: Journal of Library User Experience*. **1** (1). doi:10.3998/weave.12535642.0001.101 (https://doi.org/10.3998%2Fweave.12535642.0001.101). hdl:2027/spo.12535642.0001.101 (https://hdl.handle.net/2027%2Fspo.12535642.0001.101). ISSN 2333-3316 (https://www.worldcat.org/issn/2333-3316).
6. Kohavi, Ron (2010). "Online Controlled Experiments and A/B Testing" (https://www.researchgate.net/publication/316116834).
7. "From Infrastructure to Culture | Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining". doi:10.1145/2783258.2788602 (https://doi.org/10.1145%2F2783258.2788602). S2CID 15847833 (https://api.semanticscholar.org/CorpusID:15847833).
8. Siroker, Dan; Koomen, Pete (2013-08-07). *A / B Testing: The Most Powerful Way to Turn Clicks Into Customers* (https://books.google.com/books?id=VfVvAAAAQBAJ&q=A/B%20Testing&pg=PT13). John Wiley & Sons. ISBN 978-1-118-65920-5.
9. "Split Testing Guide for Online Stores" (http://www.webics.com.au/blog/google-adwords/split-testing-guide-for-online-retailers/). webics.com.au. August 27, 2012. Retrieved 2012-08-28.
10. Kaufman, Emilie (2014). "On the Complexity of A/B Testing" (http://proceedings.mlr.press/v35/kaufmann14.pdf) (PDF). **35**. arXiv:1405.3224 (https://arxiv.org/abs/1405.3224). Bibcode:2014arXiv1405.3224K (https://ui.adsabs.harvard.edu/abs/2014arXiv1405.3224K) – via JMLR: Workshop and Conference Proceedings.
11. Christian, Brian (2000-02-27). "The A/B Test: Inside the Technology That's Changing the Rules of Business | Wired Business" (https://www.wired.com/business/2012/04/ff_abtesting/). *Wired.com*. Retrieved 2014-03-18.

12. Christian, Brian. "Test Everything: Notes on the A/B Revolution | Wired Enterprise" (https://www.wired.com/wiredenterprise/2012/05/test-everything/). *Wired.com*. Retrieved 2014-03-18.
13. Cory Doctorow (2012-04-26). "A/B testing: the secret engine of creation and refinement for the 21st century" (https://boingboing.net/2012/04/26/ab-testing-the-secret-engine.html). Boing Boing. Retrieved 2014-03-18.
14. Krishnamoorthy, K.; Thomson, Jessica (2004). "A more powerful test for comparing two Poisson means". *Journal of Statistical Planning and Inference*. **119**: 23–35. doi:10.1016/S0378-3758(02)00408-1 (https://doi.org/10.1016%2FS0378-3758%2802%290040 8-1).
15. "What is A/B Testing." (https://www.convertize.com/what-is-ab-testing/) Convertize. Retrieved 2020-01-28.
16. "Claude Hopkins Turned Advertising Into A Science." (https://www.investors.com/news/management/leaders-and-success/claude-hopkins-scientific-advertising-bio/) Retrieved 2019-11-01.
17. "Brief history and background for the one sample t-test" (http://blog.gembaacademy.com/2007/06/20/how-beer-influenced-statistics/).
18. Box, Joan Fisher (1987). "Guinness, Gosset, Fisher, and Small Samples" (https://doi.org/10.1214%2Fss%2F1177013437). *Statistical Science*. **2** (1): 45–52. doi:10.1214/ss/1177013437 (https://doi.org/10.1214%2Fss%2F1177013437).
19. Amazon.com. "The Math Behind A/B Testing" (https://web.archive.org/web/20150921174256/https://developer.amazon.com/public/apis/manage/ab-testing/doc/math-behind-ab-testing). Archived from the original (https://developer.amazon.com/public/apis/manage/ab-testing/doc/math-behind-ab-testing) on 2015-09-21. Retrieved 2015-04-12.
20. Kohavi, Ron; Longbotham, Roger; Sommerfield, Dan; Henne, Randal M. (2009). "Controlled experiments on the web: survey and practical guide" (https://ai.stanford.edu/~ronnyk/2009controlledExperimentsOnTheWebSurvey.pdf) (PDF). *Data Mining and Knowledge Discovery*. Berlin: Springer. **18** (1): 140–181. doi:10.1007/s10618-008-0114-1 (https://doi.org/10.1007%2Fs10618-008-0114-1). ISSN 1384-5810 (https://www.worldcat.org/issn/1384-5810). S2CID 17165746 (https://api.semanticscholar.org/CorpusID:17165746).
21. Siroker, Dan; Koomen, Pete (2013-08-07). *A / B Testing: The Most Powerful Way to Turn Clicks Into Customers* (https://books.google.com/books?id=VfVvAAAAQBAJ&q=A/B%20Testing&pg=PT13). John Wiley & Sons. ISBN 978-1-118-65920-5.
22. "Advanced A/B Testing Tactics That You Should Know | Testing & Usability" (http://online-behavior.com/testing/advanced-ab-testing-tactics-1356). Online-behavior.com. Retrieved 2014-03-18.
23. "Eight Ways You've Misconfigured Your A/B Test" (http://drjasondavis.com/2013/09/12/eight-ways-youve-misconfigured-your-ab-test/). Dr. Jason Davis. 2013-09-12. Retrieved 2014-03-18.

---