A/B testing

Earlier, you learned that A/B testing is a way to compare two versions of something to find out which version performs better. For example, a data professional might use A/B testing to compare two versions of a web page or two versions of an online ad. You also learned that A/B testing utilizes statistical methods such as sampling and hypothesis testing.
In this reading, you'll learn more about the general purpose and design of an A/B test and how A/B testing uses statistical methods to analyze data.

Business context

Data professionals often use A/B testing to help stakeholders choose the best design for a website or app to optimize marketing, increase revenue, or enhance customer experience. In practice, A/B testing involves randomly selecting a sample of users and dividing them into two groups (A and B). The two groups visit different versions of a company's website. The two versions are identical except for a single design feature. For instance, the "Purchase" button on Group A's version might have a different size, shape, or color than the "Purchase" button on Group B's version. An A/B test uses statistical analysis to determine whether the change in the feature (e.g., a larger button) affects user behavior for a specific metric. A data professional might use an A/B test to analyze one of the following metrics:

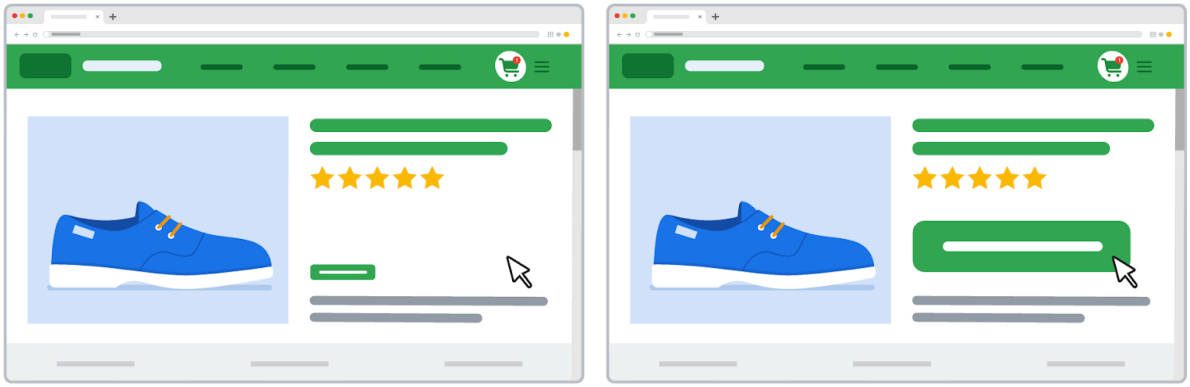
- *Average revenue per user*: How much revenue does a user generate for a website?
- *Average session duration*: How long does a user remain on a website?
- *Click rate:* If a user is shown an ad, does the user click on it?
- *Conversion rate*: If a user is shown an ad, will that user convert into a customer?

Let's explore an example to get a better understanding of how A/B testing works.

Example: Average revenue per user

Imagine you're a data professional who works for an online footwear retailer. The company is trying to grow its business and is researching the average revenue per user on its website. Your team leader asks you to conduct an A/B test to determine whether increasing the size of the "Purchase" button has any effect on average revenue. You randomly select a sample of users and divide them into two groups, A and B. Group A visits the standard version of the company website. Group B visits a version of the website that is identical to the standard version except for the larger "Purchase" button. You run the test online and collect your sample data. The results indicate that

average revenue per user is higher for Group B. Finally, you conduct a two-sample hypothesis test to determine whether the observed difference in average revenue is statistically significant or due to chance.



A typical A/B test has at least three main features:

1. Test design
2. Sampling
3. Hypothesis testing

Let's examine each feature in more detail using our example.

Test design

First, let's discuss the fundamental design of an A/B test.

Randomized controlled experiment

An A/B test is a basic version of what's known as a randomized controlled experiment. In a **randomized controlled experiment**, test subjects are randomly assigned to a control group and a treatment group. The **treatment** is the new change being tested in the experiment. The **control group** is not exposed to the treatment. The **treatment group** is exposed to the treatment. The difference in metric values between the two groups measures the treatment's effect on the test subjects.

Note: Ideally, exposure to the treatment is the only significant difference between the two groups. This test design allows researchers to control for other factors that might influence the test results and draw causal conclusions about the effect of the treatment.

In our example, group A is the control group, group B is the treatment group, and the treatment is displaying a larger "Purchase" button. Users in the control group (A) visit the standard version of the company's website. Users in the treatment group (B) visit an alternative version with a larger

“Purchase” button (i.e., are exposed to the treatment). By making the website versions for A and B identical except for the size of the “Purchase” button, you minimize the chance that any observed difference in average revenue is due to other features such as page layout or background. This allows you to measure the effect of the larger button by comparing the difference in average revenue per user for group A and group B.

Randomization, or randomly assigning test subjects to the control group or treatment group, also helps control the potential effect of other factors on the outcome of the experiment. In practice, many different factors might influence whether a user clicks the “Purchase” button or not. For example, perhaps super wealthy users are much more likely to make purchases in general, regardless of button size. If your treatment group consists *only* of super wealthy users, you won’t get valid test results. Any observed increase in average revenue might be due to wealth, not to the larger size of the “Purchase” button (the factor you’re interested in testing). Randomization helps minimize the chance that other factors, such as wealth, will significantly influence your results on average.

Sampling

Random selection helps you create a representative sample that reflects the characteristics of the overall user population. In our example, this is the population of online customers of the company you work for. Using a representative sample for your A/B test will give you valid results that are generalizable, or applicable to the overall population.

You’ll also need to choose a sample size that is appropriate for your A/B test. The larger the sample size, the more precise the results, and the more likely you’ll get results that are statistically significant when there is a difference between group A and group B. However, working with large samples can be expensive and time-consuming. Data professionals determine sample size based on both the goal of the analysis and their available budget.

Hypothesis testing

For the purpose of our example, let’s say you run the online test, collect your data, and discover that group B has a higher average revenue per user than group A. Recall that group B is the treatment group (larger “Purchase” button), and group A is the control group. The next step is to determine whether the observed difference in your data is statistically significant or due to chance. A/B tests use two-sample hypothesis tests to draw conclusions about statistical significance. To determine whether the observed difference in average revenue per user is statistically significant, you conduct a two-sample t-test. You formulate your hypotheses as follows:

- **H**0: There is no difference in average revenue per user between A and B

- **H**a: There is a difference in average revenue per user between A and B

Results

Based on the results of your t-test, you reject the null hypothesis and conclude that the observed increase in average revenue per user is statistically significant.
The results of your A/B test help you decide whether or not to recommend a design change for your company's website. In this case, when you present your results to company stakeholders, you suggest implementing the larger "Purchase" button to increase average revenue per user going forward.