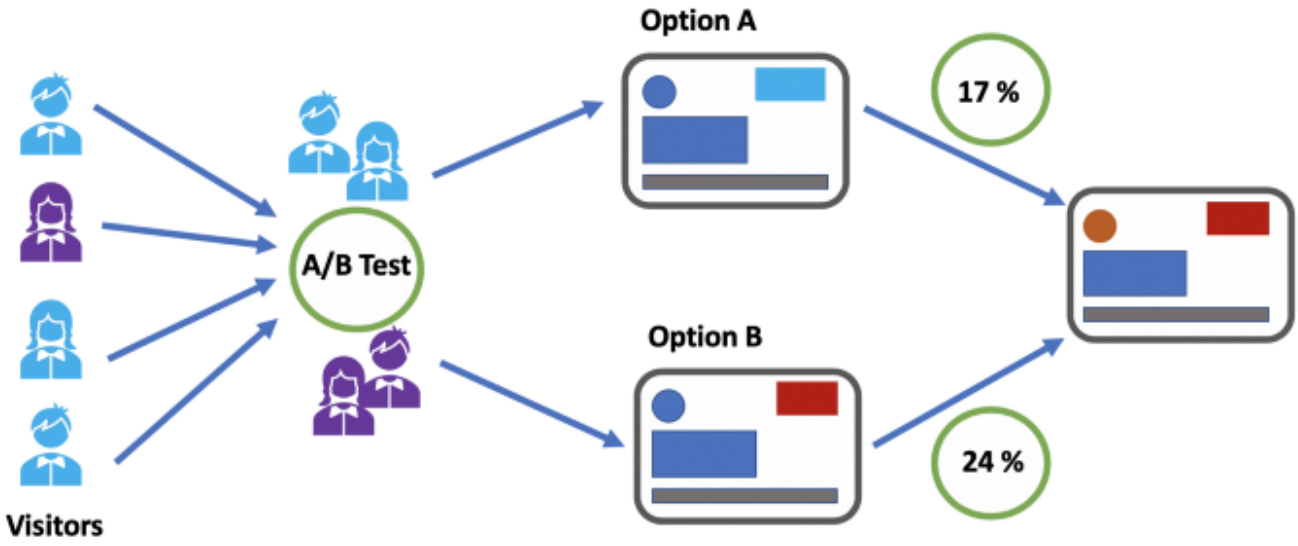
How to conduct A/B Testing?



The idea of A/B testing is to present different content to different variants (user groups), gather their reactions and user behaviour and use the results to build product or marketing strategies in the future.

> *A/B testing is a methodology of comparing multiple versions of a feature, a page, a button, headline, page structure, form, landing page, navigation and pricing etc. by showing the different versions to customers or prospective customers and assessing the quality of interaction by some metric (Click-through rate, purchase, following any call to action, etc.).*

This is becoming increasingly important in a data-driven world where business decisions need to be backed by facts and numbers.

How to conduct a standard A/B test

1. Formulate your Hypothesis
2. Deciding on Splitting and Evaluation Metrics
3. Create your Control group and Test group
4. Length of the A/B Test
5. Conduct the Test
6. Draw Conclusions

1. Formulate your hypothesis

Before conducting an A/B testing, you want to state your null hypothesis and alternative hypothesis:

*The **null hypothesis** is one that states that there is **no** difference between the control and variant group.The **alternative hypothesis** is one that states that there **is** a difference between the control and variant group.*

Imagine a software company that is looking for ways to increase the number of people **who pay for their software**. The way that the software is currently set up, users can download and use the software free of charge, for a 7-day trial. The company wants to change the layout of the homepage to emphasise with a red logo instead of blue logo that there is a 7-day trial available for the company's software.

Here is an example of hypothesis test:
Default action: Approve blue logo.
Alternative action: Approve red logo.
Null hypothesis: Blue logo ***does not cause*** at least 10% more license purchase than red logo.
Alternative hypothesis: Red logo ***does cause*** at least 10% more license purchase than blue logo.

It's important to note that all other variables need to be held constant when performing an A/B test.

2. Deciding on Splitting and Evaluation Metrics

We should consider two things: where and how we should split users into experiment groups when entering the website, and what metrics we will use to track the success or failure of the experimental manipulation. The choice of unit of diversion (the point at which we divide observations into groups) may affect what evaluation metrics we can use.

The control, or 'A' group, will see the old homepage, while the experimental, or 'B' group, will see the new homepage that emphasises the 7-day trial.

Three different splitting metric techniques:

a) Event-based diversion
b) Cookie-based diversion
c) Account-based diversion

An **event-based diversion** (like a pageview) can provide many observations to draw conclusions from, but if the condition changes on each pageview, then a visitor might get a different experience on each homepage visit. Event-based diversion is much better when the changes aren't as easily visible to users, to avoid disruption of experience.

In addition, event-based diversion would let us know how many times the download page was accessed from each condition, but can't go any further in tracking how many actual downloads were generated from each condition.

Account-based can be stable, but is not suitable in this case. Since visitors only register after getting to the download page, this is too late to introduce the new homepage to people who should be assigned to the experimental condition.

So this leaves the consideration of **cookie-based diversion**, which feels like the right choice. Cookies also allow tracking of each visitor hitting each page. The downside of cookie based diversion, is that it get some inconsistency in counts if users enter the site via incognito window, different browsers, or cookies that expire or get deleted before they make a download. As a simplification, however, we'll assume that this kind of assignment dilution will be small, and ignore its potential effects.

In terms of **evaluation metrics**, we should prefer using the **download rate** (# downloads / # cookies) and **purchase rate** (# licenses / # cookies) relative to the number of cookies as evaluation metrics.

Product usage statistics like the average time the software was used in the trial period are potentially interesting features, but aren't directly related to our experiment. Certainly, these statistics might help us dig deeper into the reasons for observed effects after an experiment is complete. But in terms of experiment success, product usage shouldn't be considered as an evaluation metric.

3. Create your control group and test group

Once you determine your null and alternative hypothesis, the next step is to create your control and test (variant) group. There are two important concepts to consider in this step, sampling and sample size.

Sampling

Random sampling is one most common sampling techniques. Each sample in a population has an equal chance of being chosen. Random sampling is important in hypothesis testing because it eliminates sampling bias, and **it's important to eliminate**

bias because you want the results of your A/B test to be representative of the entire population rather than the sample itself.

A problem of A/B tests is that if you haven't defined your target group properly or you're in the early stages of your product, you may not know a lot about your customers. If you're not sure who they are (try creating some user personas to get started!) then you might end up with misleading results. Important to understand which sampling method that suits your use case.

Sample Size

It's essential that you determine the minimum sample size for your A/B test prior to conducting it so that you can eliminate **under coverage bias**, bias from sampling too few observations.

4. Length of the A/B test

A calculator like this one can help you determine the length of time you need to get any real significance from your A/B tests.

Example Case

We will now walk you through an example. History data shows that there are about 3250 unique visitors per day. There are about 520 software downloads per day (a **.16 rate**) and about 65 licenses purchased each day (a **.02 rate**). In an ideal case, both the download rate and license purchase rate should increase with the new homepage; a statistically significant negative change should be a sign to not deploy the homepage change. However, if only one of our metrics shows a statistically significant positive change we should be happy enough to deploy the new homepage

Use the link above for the test days calculations:
Estimated existing conversion rate (%): 16%
Minimum improvement in conversion rate you want to detect (%): 50/520*100 %
Number of variations/combinations (including control): 2
Average number of daily visitors: 3250
Percent visitors included in test? 100% (3250)
Total number of days to run the test: 6 days

Estimated existing conversion rate (%): 2 %
Minimum improvement in conversion rate you want to detect (%): 10/65*100 %
Number of variations/combinations (including control): 2
Average number of daily visitors: 3250

Percent visitors included in test? 100% (3250)
Total number of days to run the test: 21 days

For an overall 5% Type I error rate with Bonferroni correction and 80% power, we should require 6 days to reliably detect a **50 download increase per day** and 21 days to detect an increase of **10 license purchases per day**. Performing both individual tests at a .05 error rate carries the risk of making too many Type I errors. As such, we'll apply the Bonferroni correction to run each test at a **.025 error rate** so as to protect against making too many errors.

One thing that isn't accounted for in the base experiment length calculations is that there is going to be a delay between when users download the software and when they actually purchase a license. That is, when we start the experiment, there could be about seven days before a user account associated with a cookie actually comes back to make their purchase. Any purchases observed within the first week might not be attributable to either experimental condition. As a way of accounting for this, we'll run the experiment for about one week longer to allow those users who come in during the third week a chance to come back and be counted in the license purchases tally.

As for biases, we don't expect users to come back to the homepage regularly. Downloading and license purchasing are actions we expect to only occur once per user, so there's no real 'return rate' to worry about. One possibility, however, is that if more people download the software under the new homepage, the expanded user base is qualitatively different from the people who came to the page under the original homepage. This might cause more homepage hits from people looking for the support pages on the site, causing the number of unique cookies under each condition to differ. If we do see something wrong or out of place in the invariant metric (number of cookies), then this might be an area to explore in further investigations.

5. Conduct the test

$$T - statistic = \frac{Observed\ value - hypothesized\ value}{Standard\ Error}$$

$$Stamdard\ Error = \sqrt{\frac{2 * Variance(sample)}{N}}$$

Once you conduct your experiment and collect your data, you want to determine if the difference between your control group and variant group is statistically significant. There are a few steps in determining this:

- First, you want to set your **alpha value. Alpha is** the probability of making a Type 1 error. Typically the alpha is set at 5% or 0.05
- Second, you want to determine the **p-value** (probability value) by first calculating the t-statistic using the formula above or using **z-score** (also called a standard **score**) gives you an idea of how far from the mean a data point is..).
- Lastly, compare the **p-value** to the **alpha**. *If the p-value is greater than the alpha, do not reject the null!*

5.1 Use actual statistics to compare the results

Do not rely on simple 1 on 1 comparison metrics to dictate what works and does not work. "*Version A* yields a 20 percent conversion rate and *Version B* yields a 22 percent conversion rate, therefore we should switch to Version B!" Please do not do this. Use actual confidence intervals, z-scores, and statistically significant data.

5.2 Product Growth

Changing colours and layout may have a marginal impact on your key performance metrics. However, these results seem to be very short-lived. Product growth does not result from changing a button from red to blue, it comes from building a product that people want to use.

> Instead of choosing feature that you think might work, you can use an A/B test to know what works.

5.3 Analyse Data

For the first evaluation metric, download rate, there was an extremely convincing effect. An absolute increase from 0.1612 to 0.1805 results in a z-score of 7.87 (z-score = 0.1805–0.1612/0.0025) and p-value < .00001, well beyond any standard significance bound. However, the second evaluation metric, license purchasing rate, only shows a small increase from 0.0210 to 0.0213 (following the assumption that only the first 21 days of cookies account for all purchases). This results in a p-value of 0.398 (z = 0.26).

6. Draw Conclusions

Despite the fact that statistical significance wasn't obtained for the number of licenses purchased, the new homepage appeared to have a strong effect on the number of downloads made. Based on our goals, this seems enough to suggest replacing the old homepage with the new homepage. Establishing whether there was a significant increase in the number of license purchases, either through the rate or the increase in the number of homepage visits, will need to wait for further experiments or data collection.

One inference we might like to make is that the new homepage attracted new users who would not normally try out the program, but that these new users didn't convert to purchases at the same rate as the existing user base. This is a nice story to tell, but we can't actually say that with the data as given. In order to make this inference, we would need more detailed information about individual visitors that isn't available. However, if the software did have the capability of reporting usage statistics, that might be a way of seeing if certain profiles are more likely to purchase a license. This might then open additional ideas for improving revenue.