UDACITY

Steps to Completing an A/B Test

Whether you're launching a new product, changing a website design, or testing a new price, an AB test can help you make a decision with confidence when you don't have a lot of data. Below is a useful, high level guide to designing and running an AB test.

Select performance metric

It's important to understand the metric used to evaluate the results of the test. Whether the goal is to increase sales, profit, conversion rate, etc., this should be specified at the upfront.

Select experimental design

Matched pair - when the sample size is small and/or the data is difficult to collect, a matched pair experiment should be used.

Randomized design - when the sample size is large and the data is easy to collect, then a randomized experiment should be used. Randomized experiments are very common for web-based AB tests.

Select the treatment and control units

Each individual in the test is considered a unit. The unit can be a person, store, etc. In a test, units are split into two groups, the treatment group and control group.

Treatment and control units are compared against each other

Select experimental and control variables

Experimental variable - The experimental, or treatment, variable, is the variable that is different between treatment and control units.

Control Variables - The control variables are the variables that should remain constant between test and control groups. These ensure that the treatment and control groups are representative of each other and that the results will apply to the population.

Determine sample size and test duration

These contribute most directly to statistical significance. You can improve statistical significance by either increasing the sample size or test duration.

Duration: Generally the duration of a test should be at least as long enough to capture a representative group. If users generally visit a store or website once a week, then the test duration should be a week.

Clean and prepare data

Clean and filter the data appropriately. This could mean filtering for the dates of the test, ensuring there are no duplicate records, removing records with incomplete data, etc.

Remove outliers. Make sure that any approach to remove outliers is unbiased.

Calculate lift

Compare the average performance between the two groups

It can also be useful to understand the distribution of the performance of the units

Determine statistical significance

Performing a t-test gives a p-value. If the p-value is below 0.05, the results are considered statistically significant.

Use a paired t-test for matched pair experiments.

Use an unpaired t-test for randomized experiments.

Estimate impact of broad implementation

In order to provide an expected impact of broad implementation of the treatment, apply the lift calculation to the entire population.