What is it?

A/B testing is also known as ***hypothesis testing***. It measures the degree of certainty we may have in the truth of an assumption made for the sake of an argument.

An A/B test should have a **defined outcome** that is measurable. A/B testing lets us compare several **alternate versions** of the same web page **simultaneously** and see which produces the **best outcome**.

What it is used for (benefits)?

Testing takes the **guesswork out** of website optimization and enables data-backed decisions that shift business conversations from "*we think*" to "*we know*."

It is a simple way to **check changes** to design of some implementations against the current design and determine which ones produce positive results. These tests are usually performed on **live sites with real users** who are completely unaware of the test.

Who is using it (examples) and for what purpose?

Hence, for example, A/B testing lets us compare several **alternate versions** (multinomial) or two versions of the same web page **simultaneously** and see which produces the **best outcome**. It is mainly used for web development and online marketing.

Outcome metrics for a website maybe:

- Number of sales made
- Increased click through
- Number of people signing up/registering/downloads

The statistics behind it

A/B Testing is a way of conducting an experiment where you compare a **control group** to the performance of one or more **experimental test groups** by randomly assigning each group a specific single-variable treatment.

For this scenario, we have a control treatment (current website), and an experimental treatment.

$\mathbf{p_c}$: conversion rate of the control; **p** : conversion rate of one of our experiments; **n** : sample size

$$\text{Conversion rate} = \frac{\text{Views that converted into positive conversions}}{\text{Total number of views}}$$

A ***null hypothesis*** is drawn: conversion rate of the control treatment is no less than the conversion rate of our experimental treatment (new treatment is worse/same as old). I.e.

$$H_0 : p_c \geq p$$

The ***alternative hypothesis*** is therefore that the experimental page has a higher conversion rate. This is what we want to see and quantify. I.e.

$$H_1 : p_c < p$$

Considering that the sample size is large(>30), central limit theorem specifies that the mean comes from normal distribution.

$$z = \frac{x - \mu}{\sigma/\sqrt{n}}$$

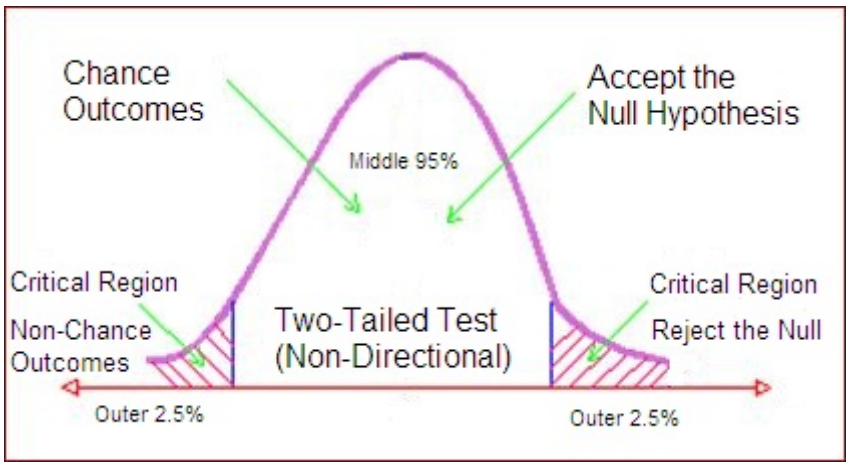
here,

$\mu = p$; $z = p_c - p / \sigma\sqrt{n}$

Finally, only two conclusions can be reached. H_0 is **false**, or H_0 is **plausible**.

The hypothesis test produces a number, (P-value) between 0 and 1 that measures the strength against of evidence against H_0. The smaller the P-value, the stronger is the evidence.

A rule of thumb suggests to reject H_0 whenever $P \leq 0.05$. (Convenient but no scientific basis.)



For a 95% confidence interval

Also, if segmented results are expected from A/B tests (results which depend on/or can vary significantly with customer attributes like gender), test should be **evenly distributed** across such key customer attributes. Failure to do so could lead to experiment bias and inaccurate conclusions to be drawn from the test. This can be achieved from a **completely randomized experiment.**