A/B Testing for Data Science using Python – A Must-Read Guide for Data Scientists

Overview

- A/B testing is a popular way to test your products and is gaining steam in the data science field
- Here, we'll understand what A/B testing is and how you can leverage A/B testing in data science using Python

> Statistical analysis is our best tool for predicting outcomes we *don't* know, using the information we know.

Picture this scenario – You have made certain changes to your website recently. Unfortunately, you have no way of knowing with full accuracy how the next 100,000 people who visit your website will behave. That is the information we *cannot* know today, and if we were to wait until those 100,000 people visited our site, it would be too late to optimize their experience.

This seems to be a classic Catch-22 situation!

This is where a data scientist can take control. A data scientist collects and studies the data available to help optimize the website for a better consumer experience.

And for this, it is imperative to know how to use various statistical tools, especially the concept of A/B Testing.

A/B Testing is a widely used concept in most industries nowadays, and data scientists are at the forefront of implementing it. In this article, I will explain A/B testing in-depth and how a data scientist can leverage it to suggest changes in a product. In this Article you will get to understanding on AB testing on data science how data science ab testing works and hows the significance test happens and why we should use ab testing for data science.

This article will delve into A/B testing in data science utilizing Python. We will discuss the implementation of A/B testing in data science projects, with included free resources and practical examples. Furthermore, we will explore examples of A/B testing statistics to improve your grasp of the fundamental principles and methodologies. At the conclusion, you will have the expertise necessary to successfully utilize A/B testing in your data-based decision-making.

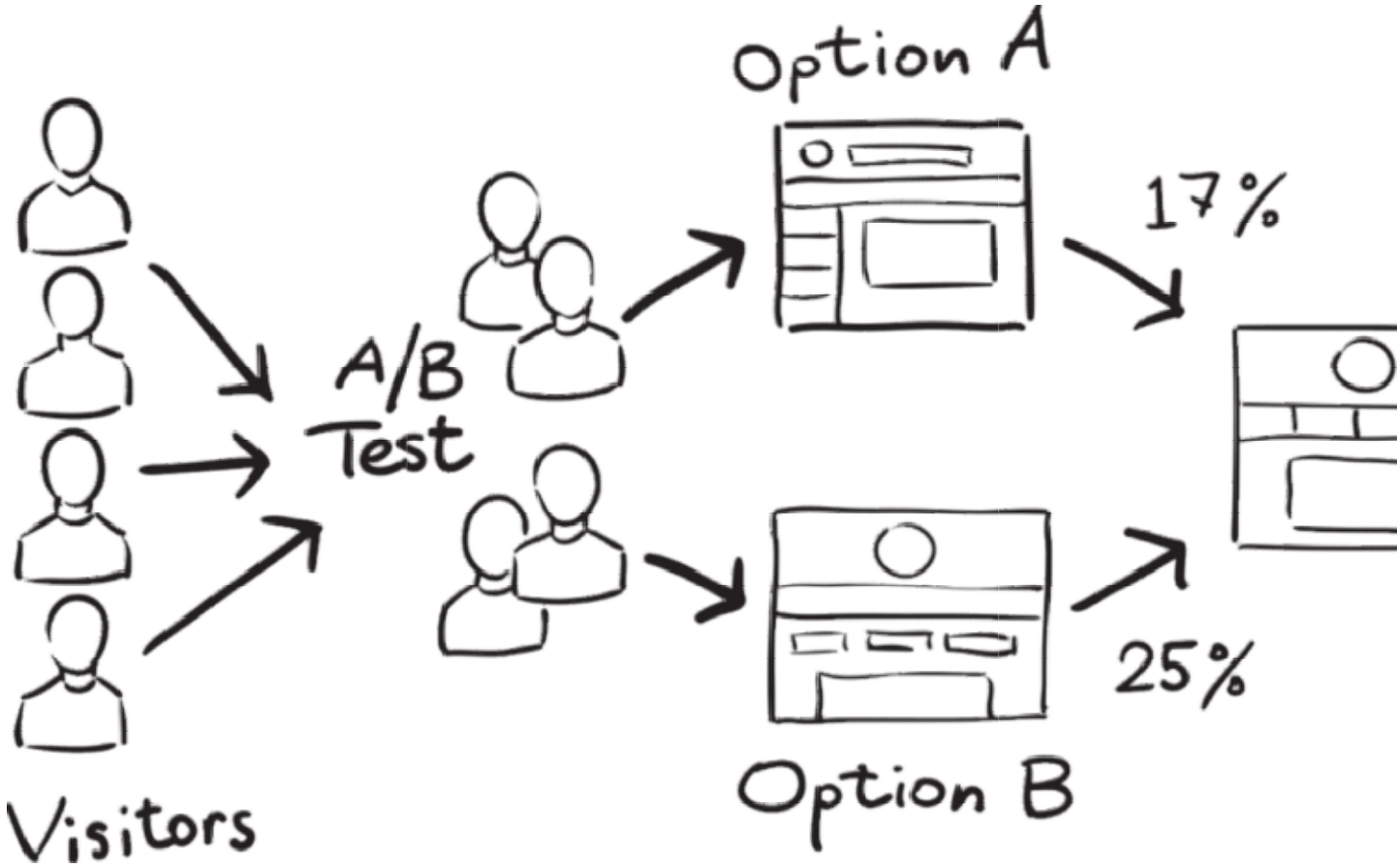
Table of contents

What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let's say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



Source

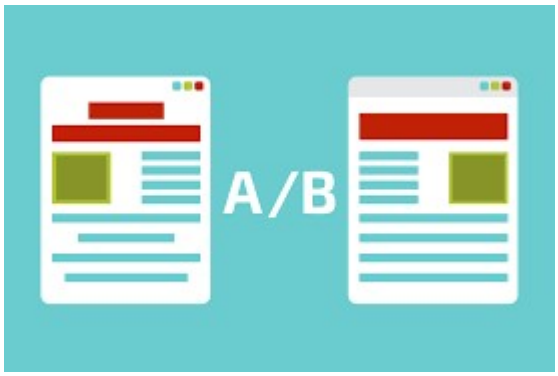
It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The **population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

How does A/B Testing Work?

The big question!

In this section, let's understand through an example the logic and methodology behind the concept of A/B testing.

Let's say there is an e-commerce company XYZ. It wants to make some changes in its newsletter format to increase the traffic on its website. It takes the original newsletter and marks it A and makes some changes in the language of A and calls it B. Both newsletters are otherwise the same in color, headlines, and format.



Objective

Our objective here is to check which newsletter brings higher traffic on the website i.e the conversion rate. We will use A/B testing and collect data to analyze which newsletter performs better.

Make a Hypothesis

Before making a hypothesis, let's first understand what is a hypothesis.

> A hypothesis is a tentative insight into the natural world; a concept that is not yet verified but if true would explain certain facts or phenomena.

It is an **educated guess** about something in the world around you. It should be testable, either by experiment or observation. In our example, the hypothesis can be "By making changes in the language of the newsletter, we can get more traffic on the website".

In **hypothesis testing**, we have to make two hypotheses i.e Null hypothesis and the alternative hypothesis. Let's have a look at both.

The alternative hypothesis challenges the null hypothesis and is basically a hypothesis that the researcher believes to be true. The alternative hypothesis is what you might hope that your A/B test will prove to be true.

In our example, the H_a is- "**the conversion rate of newsletter B is higher than those who receive newsletter A**".

Now, we have to collect enough evidence through our tests to **reject the null hypothesis**.

Create Control Group and Test Group

Once we are ready with our null and alternative hypothesis, the next step is to decide the group of customers that will participate in the test. Here we have two groups – **The Control group**, and **the Test (variant) group**.

The Control Group is the one that will receive newsletter A and the Test Group is the one that will receive newsletter B.

For this experiment, we randomly select 1000 customers – 500 each for our Control group and Test group.

Randomly selecting the sample from the population is called **random sampling**. It is a technique where 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.**

Another important aspect we must take care of is **the Sample size.** It is required that we determine the minimum sample size for our A/B test before conducting it so that we can eliminate **under coverage bias.** It is the bias from sampling too few observations.

Conduct the A/B Test and Collect the Data

One way to perform the test is to calculate **daily conversion rates** for both the treatment and the control groups. Since the conversion rate in a group on a certain

day represents a single data point, the sample size is actually the number of days. Thus, we will be testing the difference between the mean of daily conversion rates in each group across the testing period.

When we run our experiment for one month, we noticed that the mean conversion rate for the Control group is 16% whereas that for the test Group is 19%.

Data Science AB testing

A/B testing is a fundamental tool used by data scientists to optimize and improve various aspects of products and services. It's essentially a controlled experiment where two versions of something (A and B) are compared to see which performs better based on a predefined metric.

Here's is of how A/B testing works in data science:

Core Concept:

- Split your target audience or user base into two random groups.
- Show each group a different version (A or B) of the element you're testing. This could be a website layout, email format, product pricing, advertisement, etc.
- Collect data on user behavior and measure each version's impact on a specific metric (e.g., click-through rate, conversion rate, sales).
- Analyze the data statistically to determine if a performance difference exists between A and B.

Data Science Involvement:

Data scientists play a crucial role in various stages of A/B testing:

- **Designing the Test:**
    - They help formulate a clear hypothesis and define the metric for success.
    - They determine the sample size needed for statistical significance.
- **Building the Experiment:**
    - Data scientists may develop tools to randomly assign users to groups and ensure proper delivery of variations.
- **Data Analysis:**
    - They employ statistical methods to analyze the collected data and assess the validity of the results. This involves techniques like hypothesis testing and p-value calculations to determine if the observed difference is due to chance or a genuine effect of the variation.
- **Interpretation and Recommendation:**
    - Data scientists interpret the results by considering statistical significance and effect size.
    - They recommend keeping the winning variation, refining the test, or concluding the experiment.

Statistical significance of the Test

Now, the main question is – Can we conclude from here that the Test group is working better than the control group?

The answer to this is a simple No! For rejecting our null hypothesis we have to prove the **Statistical significance** of our test.

There are two types of errors that may occur in our hypothesis testing:

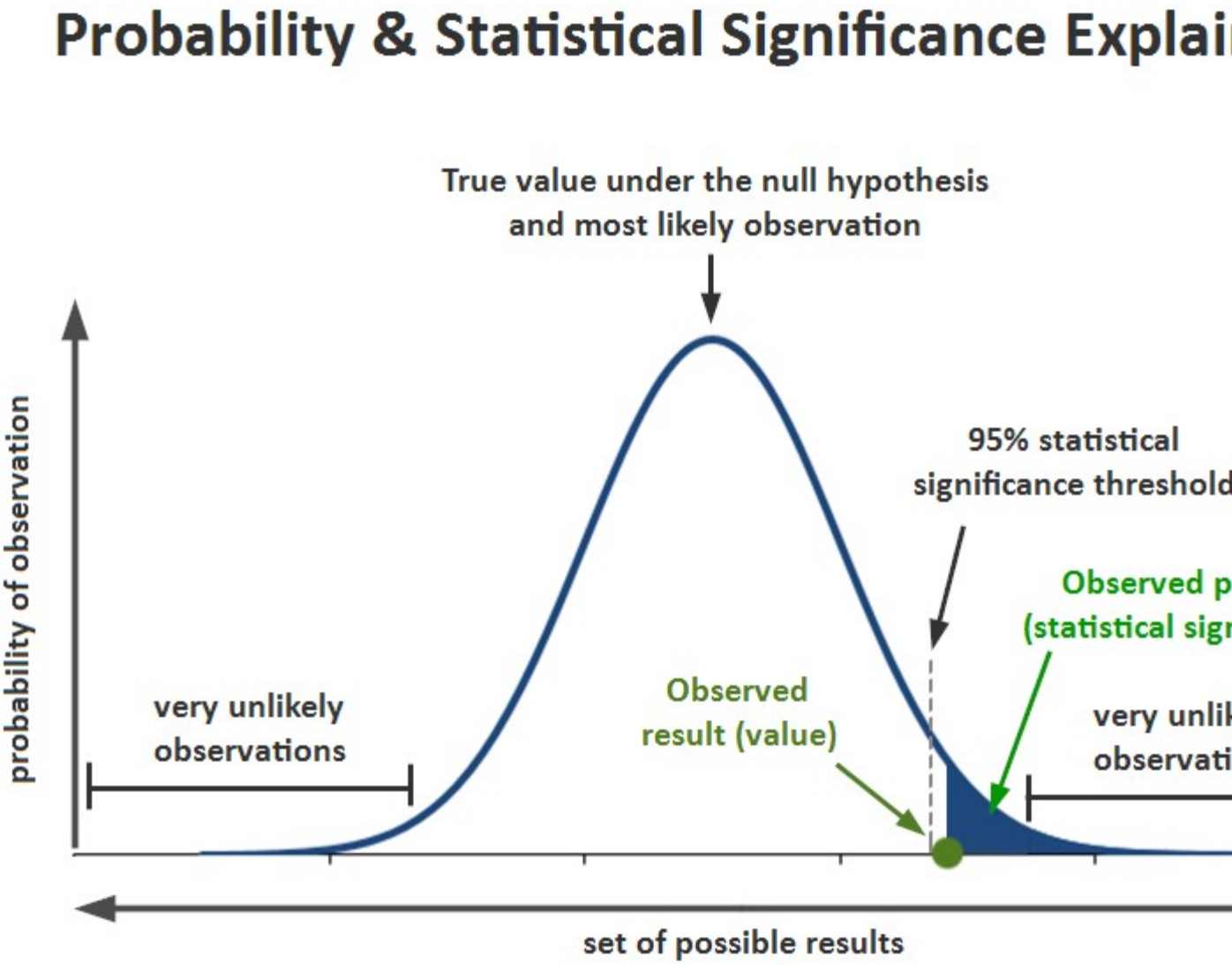
1. **Type I error**: We reject the null hypothesis when it is true. That is we accept the variant B when it is not performing better than A
2. **Type II error**: We failed to reject the null hypothesis when it is false. It means we conclude variant B is not good when it performs better than A

To avoid these errors we must calculate the statistical significance of our test.

An experiment is considered to be statistically significant when we have enough evidence to prove that the result we see in the sample also exists in the population.

That means the difference between your control version and the test version is not due to some error or random chance. To prove the statistical significance of our experiment we can use a **two-sample T-test**.

The **two–sample t–test** is one of the most commonly **used** hypothesis **tests**. It is applied to compare whether the average difference between **the two** groups.



Source

To understand this, we must be familiar with a few terms:

1. **Significance level (alpha):** The significance level, also denoted as alpha or α, is the probability of rejecting the null hypothesis when it is true. Generally, we use the significance value of 0.05

2. **P-Value:** It is the probability that the difference between the two values is just because of random chance. P-value is evidence against the null hypothesis. The smaller the p-value stronger the chances to reject the H_0. For the significance level of 0.05, if the p-value is lesser than it hence we can reject the null hypothesis
3. **Confidence interval:** The confidence interval is an observed range in which a given percentage of test outcomes fall. We manually select our desired confidence level at the beginning of our test. Generally, we take a 95% confidence interval

Next, we can calculate our t statistics using the below formula:

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

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

Let's Implement the Significance Test in Python

Let's see a python implementation of the significance test. Here, we have a dummy data having an experiment result of an A/B testing for 30 days. Now we will run a two-sample t-test on the data using Python to ensure the statistical significance of data.You can download the sample data **here**.

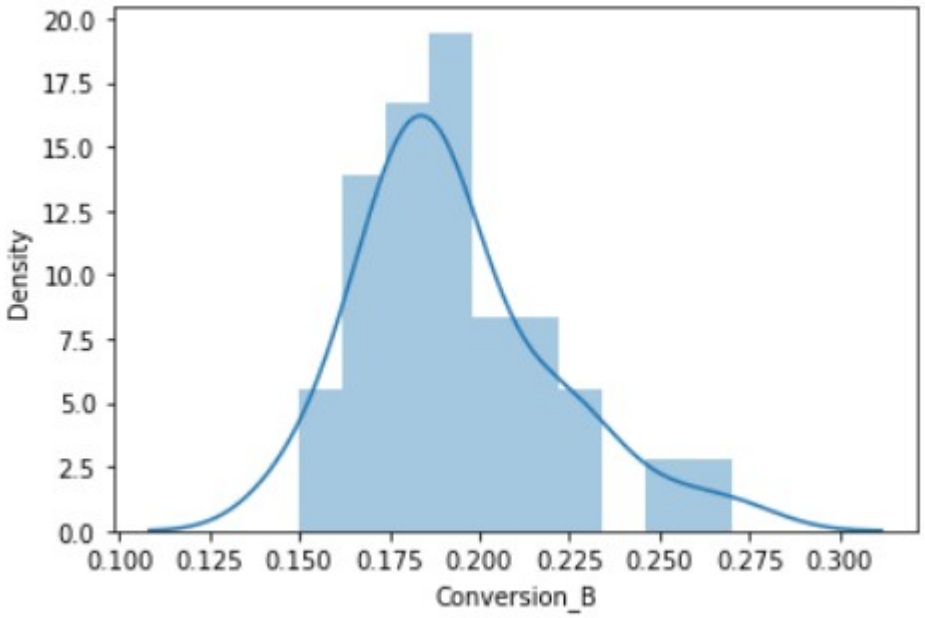
Python Code:

```
import pandas as pd
import numpy as np
import seaborn as sns
import scipy.stats as ss
import matplotlib.pyplot as plt
```

```
data= pd.read_csv("data.csv")
print(data.head())

# Let's plot the distribution of target and control group:

sns.distplot(data.Conversion_A)
plt.show()
<div>
<pre>sns.distplot(data.Conversion_B)</pre>
</div>
```



At last, we will perform the t-test:

```
<div>
<pre>t_stat, p_val= ss.ttest_ind(data.Conversion_B,data.Conversion_A)
t_stat , p_val</pre>
<pre>(3.78736793091929, 0.000363796012828762)</pre>
</div>
```

For our example, the observed value i.e the mean of the test group is 0.19. The hypothesized value (Mean of the control group) is 0.16. On the calculation of the t-score, we get the t-score as **.3787**. and the p-value is **0.00036**.

SO what does all this mean for our A/B Testing?

Here, our p-value is less than the significance level i.e 0.05. Hence, we can reject the null hypothesis. This means that in our A/B testing, newsletter B is performing better than newsletter A. So our recommendation would be to replace our current newsletter with B to bring more traffic on our website.

What Mistakes Should we Avoid While Conducting A/B Testing?

There are a few key mistakes I've seen data science professionals making. Let me clarify them for you here:

- **Invalid hypothesis**: The whole experiment depends on one thing i.e the hypothesis. What should be changed? Why should it be changed, what the expected outcome is, and so on? If you start with the wrong hypothesis, the probability of the test succeeding, decreases
- **Testing too Many Elements Together:** Industry experts caution against running too many tests at the same time. Testing too many elements together makes it difficult to pinpoint which element influenced the success or failure. Thus, prioritization of tests is indispensable for successful A/B testing
- **Ignoring Statistical Significance:** It doesn't matter what you feel about the test. Irrespective of everything, whether the test succeeds or fails, allow it to run through its entire course so that it reaches its statistical significance
- **Not considering the external factor:** Tests should be run in comparable periods to produce meaningful results. For example, it is unfair to compare website traffic on

the days when it gets the highest traffic to the days when it witnesses the lowest traffic because of external factors such as sale or holidays

When Should We Use A/B Testing?

A/B testing works best when testing incremental changes, such as UX changes, new features, ranking, and page load times. Here you may compare pre and post-modification results to decide whether the changes are working as desired or not.

A/B testing doesn’t work well when testing major changes, like new products, new branding, or completely new user experiences. In these cases, there may be effects that drive higher than normal engagement or emotional responses that may cause users to behave in a different manner.

End Notes

To summarize, A/B testing is at least a 100-year-old statistical methodology but in its current form, it comes in the 1990s. Now it has become more eminent with the online environment and availability for big data. It is easier for companies to conduct the test and utilize the results for better user experience and performance.

There are many tools available for conducting A/B testing but being a data scientist you must understand the factors working behind it. Also, you must be aware of the statistics in order to validate the test and prove it’s statistical significance.

We hope you like the article and understanding also on ab testing on data science and how data science ab testing works.

To know more about hypothesis testing, I will suggest you read the following article:

- ***Statistics for Analytics and Data Science: Hypothesis Testing and Z-Test vs. T-Test***