**Complete Guide to A/B Testing Design, Implementation and Pitfalls**

End-to-end A/B testing for your Data Science experiments for non-technical and technical specialists with examples and Python implementation

A/B testing also called **split testing,**originated from the**randomized control trials**in Statistics, is one of the most popular ways for Businesses to test new UX features, new versions of a product, or an algorithm to decide whether your business should launch that new product/feature or not.

*“The world is a big A/B test”****Sean Ellis***

This article is dedicated to both technical and non-technical audiences where I will cover the following topics that one should consider and conduct when performing an A/B test:

**- What is A/B testing and when to use it?  
- Questions to clarify before any A/B test  
- Choice of Primary metric   
- Hypothesis of the test  
- Design of the test (Power Analysis)  
- Calculation of Sample Size, Test Duration   
- Statistical tests (T-test, Z-test, Chi-squared test)  
- Analysing A/B test results in Python   
- Bootstrapping and Bootstrap Quantile Method for SE and CI  
- Statistical Significance vs Practical Significance  
- Quality of A/B test (Reliability, Validity, Potency)  
- Common problems and pitfalls of A/B testing  
- Ethics and privacy in A/B testing**

*If you have no prior Statistical knowledge, you can simply skip the statistical derivations and formulas. However, if you want to learn or refresh your knowledge in the essential statistical concepts you can check this article:*[***Fundamentals of statistics for Data Scientists and Data Analysts***](https://towardsdatascience.com/fundamentals-of-statistics-for-data-scientists-and-data-analysts-69d93a05aae7)

**What is A/B testing and when to use it?**

The idea behind A/B testing is that you show the variated version of the product to a sample of customers **(the experimental group)** and the existing version of the product to another sample of customers **(the control group)**. Then, the difference in product performance in experimental/treatment versus the control group is tracked, to identify the effect of this new version(s) of the product on the performance of the product. So, the goal is then to track the metric during the test period and find out whether there is a *difference* in the performance of the product and what type of difference is it.

The motivation behind this test is to test new product variants that will improve the performance of the existing product and will make this product more successful and optimal, showing a positive treatment effect.

What makes this testing great is that businesses are getting direct feedback from their actual users by presenting them the existing versus variated product/feature options and in this way they can quickly test new ideas. In case the A/B test shows that the variated version/approach is not effective at least businesses can learn from it and decide whether they need to improve it or need to look for other ideas.

**Benefits of A/B testing**

* Allows to learn what works and what doesn’t in a quick manner
* You get feedback directly from actual/real product customers
* Since the users are not aware that they are being tested, the results will be unbiased

**Demerits of A/B testing**

* Presenting different content/price/features to different customers especially in the same geolocation might potentially be dangerous resulting in **Change Aversion**(we will discuss how this can be addressed later on)
* Requires a significant amount of Product, Engineering, and Data Science resources
* Might lead to wrong conclusions if not conducted properly

The motivation behind A/B test is to test new product variants that will improve the performance of the existing product and will make this product more successful and optimal, showing a positive treatment effect.

**Questions to ask before any A/B test**

Given that the A/B test requires a significant amount of resources and might result in product decisions with a significant impact, it’s highly important to ask yourself, the product, and the engineering teams, and other stakeholders involved in the experiment a few essential questions before jumping to running the test.

* What does a **sample population** look like and what are the **customer segments** for the target product?
* Can we find the answer to our business question using exploratory/historical data analysis (e.g. by using causal analysis)?
* Do we want to test single or multiple variants of the target product?
* Can we ensure **truly randomized**control and experimental groups s.t. both samples are an unbiased and true representation of the true user population?
* Can we ensure the **integrity** of the treatment vs control effects during the entire duration of the test?

The goal of A/B testing is to track the primary metric during the test period and find out whether there is a *difference* in the performance of the product and what type of difference is it.

**Choosing primary metric for the A/B test**

Choosing the metric is one of the most important parts of the A/B test since this metric will be used to measure the performance of the product or feature for the experimental ad control groups and will be used to identify whether there is a **statistically significant difference** between these two groups.

The choice of the *success* metric depends on the underlying hypothesis that is being tested with this A/B test. This is if not the most, one of the most important parts of the A/B test because it determines how the test will be designed and also how well the proposed ideas perform. Choosing poor metrics might disqualify a large amount of work or might result in wrong conclusions.

Revenue is not always the end goal, so for the A/B test, we need to tie up the primary metric to the direct and the higher-level goals of the product. The expectation is that if the product makes more money, then this suggests the content is great. But in achieving that goal, instead of improving the overall content of the material and writing, one can just optimize the conversion funnels. One way to test the accuracy of a metric you have chosen for your A/B test could be to go back to the exact problem you want to solve. You can ask yourself the following question:

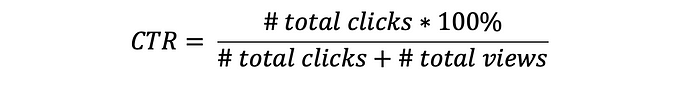
**Metric Validity Question: If this chosen metric were to increase significantly while everything else stays constant, would we achieve our goal and address the problem?**

Though you need to have a single primary metric for your A/B test, you still need to keep an eye on the remaining metrics to make sure all the metrics are showing change and not only the target one. Having multiple metrics in your A/B test will lead to false positives since you will identify many significant differences while there is no effect, which is something you want to avoid.

**Common A/B test metrics**

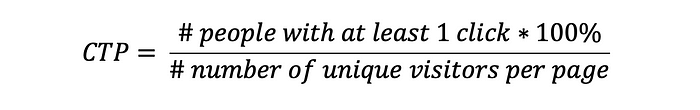
Popular performance metrics that are often used in A/B testing are the Click Through Rate, Click Through Probability and Conversion Rate.

**1: Click-Through Rate (CTR)** for usage



where the number of total views or sessions is taken into account. This number is the percentage of people who view the page (impressions) and then actually click on it (clicks).

**2: Click-Through Probability (CTP)** for impact

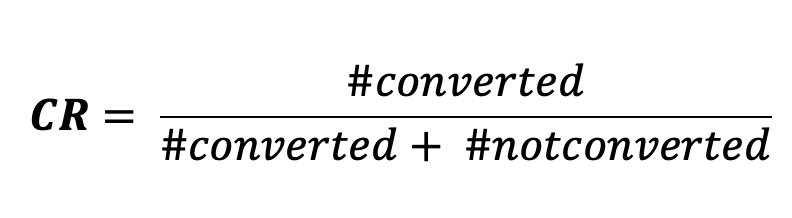


Unlike the CTR, the CTP does take into account the duplicate clicks which means that if a user for some reason has performed multiple clicks, in a single session, on the same item for some reason (e.g. because of impatience), this multiple clicks is counted as a single click in CTP.

For computing the CTP you need to work with engineers to modify the website s.t. for every page view you capture the view/impression event and you capture the click event, to match each page view with all the**child clicks** **per page** to make sure you count only 1 child click per unique page view.

**3: Conversion Rate**

Conversion rate, defined as the proportion of sessions ending up with a transaction.



So, you can use CTR if you want to measure the usability of the site and use CTP if you want to measure the actual impact of the feature. CTR doesn’t take care of the duplicate clicks, so if the user has impatiently pushed the same button multiple times, this will not be corrected to be equal to 1.

**Stating the hypothesis of the test**

A/B test should always be based on a hypothesis that needs to be tested. This hypothesis is usually set as a result of brainstorming and collaboration of relevant people on the Product team and Data Science team. The idea behind this hypothesis is to decide how to ‘fix’ a potential issue in the product where a solution of these problems will influence the Key Performance Indicators (KPIs) of interest?

It’s also highly important to make prioritization out of the range of product problems and ideas to test, while you want to pick a problem s.t. fixing this problem would result in the biggest impact for the product.

For example, if the KPI of the product is to improve the quality of the recommender system’s recommendations and this can be done for example by adding **Impression Discounting** or building a **Re-ranker**modelfor the recommender. However, the impact of these two solutions will likely be different on the amount of improvement of the quality of the recommendations. Namely, the re-ranker model affects the ranks of the recommendations by potentially changing the set of recommendations presented to the user, unlike impression discounting which just makes sure that the user doesn’t see the recommendations that were previously viewed to the user.

For this particular example, we could decide to build a re-ranked model which we expect to improve the quality of the target recommender (let’s name this imaginary recommender system, RecSys). Additionally, we have performed research ad identified that [XGBoost](https://en.wikipedia.org/wiki/XGBoost) can be used as a re-ranker model to rerank RecSys recommendations.

Finally, we have performed an exploratory analysis/offline test where we have seen an increase in the quality of the recommendations (let’s say using [NDCG](https://www.youtube.com/watch?v=qG0wUgsEugw&t=31s&ab_channel=WayfairDataScience) as a performance measure for the quality of the recommendations) and have identified that significant impact. So, as a final check, we want to test the effectiveness of this XGBoost re-ranker on the quality of the RecSys recommendations compared to the existing version of the recommender. Hence, we can state the following hypothesis:

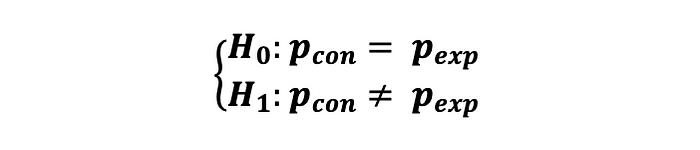
Hypothesis: Adding an XGBoost re-ranker model to the existing RecSys recommender will increase the CTR of the recommendations, that is, will improve the quality of the RecSys recommendations.

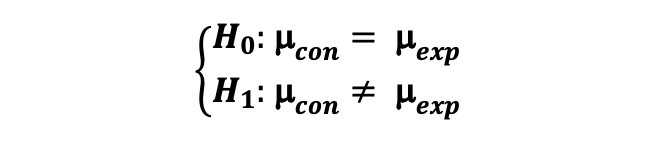
Do not merge multiple ideas into one hypothesis and also limit the variables introduced in the test so that you can understand their individual impact. Otherwise, you’ll be left with many questions and few answers at the end of your test.

**Designing the A/B test**

Some argue that A/B testing is an art and others state that it's a business-adjusted common statistical test. But the borderline is that to properly design this experiment, you need to be disciplined and intentional while keeping in mind that it’s not really about testing — it’s about *learning*. Following are the steps you need to take to have a solid design for your A/B test.

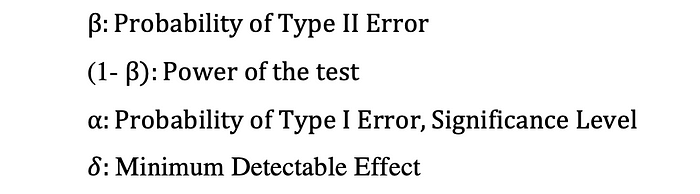
**Step 1: Statistical Hypothesis**





**Step 2: Power Analysis**

To make sure that our results are repeatable, robust, and can be generalized to the entire population we need to avoid [p-hacking](https://en.wikipedia.org/wiki/Data_dredging), to ensure real statistical significance and to avoid biased results, we want to make sure we collect “enough” amount of observations and we run the test for a minimum predetermined amount of time. Therefore, before running the test we need to determine the sample size of the control and experimental groups and for how long we need to run the test. This process is often referred to as Power Analysis and it includes 3 specific steps: determining the power of the test, determining the significance level of the test, and determining a Minimum Detectable Effect. A popular reference to the parameters involved in the Power Analysis for A/B testing is the following notation:



**Power of the test**

The power of the statistical test is the probability of correctly rejecting the null hypothesis. Power is the probability of making a correct decision (to reject the null hypothesis) when the null hypothesis is false.

The power, often defined by (1-beta) is equal to the probability of not making a type II error, where Type II error is the probability of not rejecting the null hypothesis while the null is false.

It’s common practice to pick 80% as the power of the A/B test, that is 20% Type II error which means that we are fine with not detecting (failing to reject the null) a treatment effect while there is an effect. However, the choice of a value of this parameter depends on the nature of the test and the business constraints.

**Significance level of the test**

The significance level which is also the probability of Type I error is the likelihood of rejecting the null, hence detecting a treatment effect, while the null is true and there is no statistically significant impact. This value, often defined by the Greek letter alpha, α, is the probability of making a false discovery often referred to as a false-positive rate.

Generally, we use the significance value of 5% which indicates that we have a 5% risk of concluding that there exists a statistically significant difference between the experimental and control variant performances when there is no actual difference. So, we are fine by having 5 out of 100 cases detecting a treatment effect while there is no effect. It also means that you have a significant result difference between the control and the experimental groups with a 95% *confidence*.

Like in the case of the power of the test, the choice of the value of alpha is dependent on the nature of the test and the business constraints. For instance, if running this A/B test is related to high engineering costs then the business might decide to pick high alpha such that it would be easier to detect a treatment effect. On the other hand, if the implementation costs of the proposed version in production are high you can pick a low significance level since this proposed feature should really have a big impact to justify the high implementation costs, so it should be harder to reject the null.

**Minimum Detectable Effect (delta)**

From the business point of view, what is the **substantive** to the statistical significance that the business wants to see as a minimum impact of the new version to find this variant investment-worthy?

The answer to this question is what form the amount of change we aim to observe in the new version’s metric compared to the existing one to make recommendations to the business that this feature should be launched in the production. An estimate of this parameter is what is known as the **Minimum Detectable Effect**, often defined by the Greek letter **delta,** which is also related to the practical significance of the test. MDE is a proxy that relates to **the smallest effect** that would matter in practice for the business and is usually set by stakeholders.

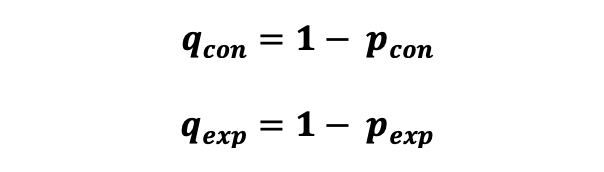
It’s common practice to pick 80% as the power and 5% as the significance level of the A/B test, that is 20% Type II error and 5% Type I error. However, the choice of a value of this parameter depends on the nature of the test and the business constraints.

**Step 3: Calculating minimum sample size**

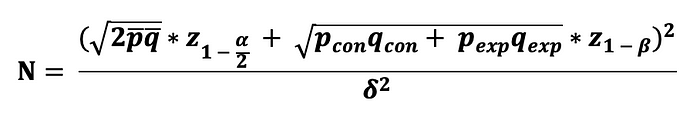
Another very important part of A/B testing is determining the minimum sample size of the control and experimental groups, which needs to be determined using the defined **power of the test**(1-beta), the **significance level**(alpha), **Minimum Detectable Effect** (MED), and the variances of the two [Normally Distributed](https://en.wikipedia.org/wiki/Normal_distribution)samples of equal size. Calculation of the sample size depends on the underlying primary metric that you have chosen for tracking the progress of the control and experimental versions. Here we distinguish two cases; case 1 where the primary metric of A/B testing is in the form of a binary variable (e.g. click or no click) and case 2 where the primary metric of the test is in the form of proportions or averages (e.g. mean order amount).

**Case 1: Sample Size Calculation with Binary Metric**

When we are dealing with a primary performance tracking metric that has two possible values such as the Click Through Rate where the user can either click (success) or not click (failure) and if the users’ responses to the product can be defined as “independent” events then we can consider this as Bernoulli Trials where the click event (success) occurs with probability **p\_con** in case of the Control Group and **p\_exp** in case of the Experimental Group. Moreover, the no click event (failure) occurs with probability **q\_con** in the case of the Control Group and **q\_exp** in the case of the Experimental Group, where:



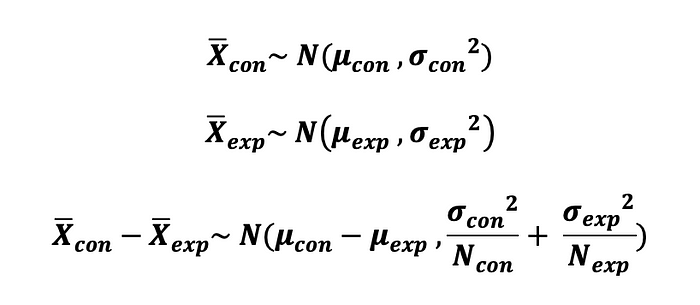
Consequently, the random variable describing the number of successes (clicks) received from the users during the test follows binomial distributions where the sample size is the number of times the feature/product was impressed to the users and the probability of success is **p\_con**and**p\_exp,**for the Control and Experimental Groups, respectively**.**Then the sample size needed to compare these two Binomial Proportions, using a two-sided test with prespecified significance level, power level, MID can be calculated as follows:



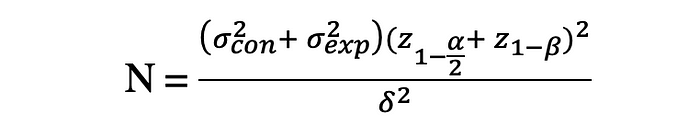
where we need to use **A/A testing**(an A/B test except assigning the same treatment to both groups)to obtain the estimates for p\_bar and q\_bar.

**Case 2: Sample Size Calculation with Continuous Metric**

When we are dealing with a primary performance tracking metric that is in the form of an average such as the *mean order amount* where we intend to compare the means of the Control and Experimental Group*s,* then we can use the Central Limit Theorem and state that the mean sampling distribution of both Control and Experimental Groups follow Normal Distribution. Consequently, the sampling distributions of difference of means of these two groups also follow Normal Distribution. That is:



Hence, the sample size needed to compare the means of two Normally Distributed Samples, using a two-sided test with prespecified significance level, power level, MID can be calculated as follows:



where we can run an A/A test to obtain the sample variances, σ²\_con, and σ²\_exp.

The random variable describing the number of successes (clicks) received from the users during the test follows binomial distributions where the sample size is the number of times the feature/product was impressed to the users and the probability of success is p\_con and p\_exp, for the Control and Experimental Groups, respectively.

**Step 4: Determining A/B test duration**

As mentioned before, this question needs to be answered before you run your experiment not during, by trying to stop the test when you detect statistical significance. To determine the baseline of a duration time, a common approach is to use the following formula:



For example, if this formula results in 14 this suggests running the test for 2 weeks. However, it’s highly important to take many business-specific aspects into account when choosing the time to run the test and for how long, and to use this formula with a grain of salt.

For instance, if one wanted to run an experiment at the beginning of January of 2020 when the COVID -19 pandemic shook the world and this had an impact on page usage, for some businesses this meant a high increase in page usage, and for some, a huge decrease in usability then running A/B test without taking this into account would results in inaccurate results since the activity period would not be a true representation of common page usage.

**Too small test duration: Novelty Effects**

Users tend to react quickly and positively to all types of changes independent of their nature. This positive effect to the experimental version that is entire since there is a change, regardless of what the change is, is referred to as *novelty effect* and it wears off in time and is thus considered “illusory”. So, it would be wrong to describe this effect to the experimental version itself and to expect that it will continue to persist after the novelty effect wears off.

Hence, when picking a test duration we need to make sure we don’t run the test for too short a time period otherwise we can have a novelty effect. Novelty effects can be a major threat to the external validity of an A/B test, so it's important to avoid it as much as possible.

**Too large test duration: Maturation Effects**

When planning an A/B test it is usually useful to consider a longer test duration for allowing users to get used to the new feature or product. In this way, one will be able to observe the real treatment effect by giving more time to returning users to cool down from an initial positive reaction or spike of interest due to a change that was introduced as part of a treatment. This should help to avoid novelty effect and thus better predictive value for the test outcome. However, the longer the test period the larger is the likelihood of external effects impacting the reaction of the users and possibly contaminating the test results, *maturation effect*. Therefore, running the A/B test for too long is is also not recommended and can better be avoided to increase the reliability of the results.

The longer the test period, the larger is the likelihood of external effects impacting the reaction of the users and possibly contaminating the test results.

**Running the A/B test**

Once the preparation work has been done, with the help of engineering you can start running the A/B test. Firstly, the engineering team needs to make sure that the integrity between Control and Experimental groups is kept. Secondly, the mechanism storing users’ responses to the treatment has to be accurate and the same across all users to avoid **systematic bias**. There are also few things you want to avoid doing such as stopping the test too early once you detect statistical significance (small p-value) while you have not reached the minimum sample size calculated before starting the test.

**Analyzing A/B test results with Python**

When it comes to interpreting the results of your A/B test, there is a set of values you should calculate to test the statistical hypothesis stated earlier (to test whether there is a statistically significant difference between control and experimental groups). This set includes:

* Choosing an appropriate statistical test
* Calculating the test statistics (T)
* Calculating the p-value of the test statistics
* Reject or fail to reject the statistical hypothesis (statistical significance)
* Calculate the margin of error (external validity of the experiment)
* Calculate confidence interval (external validity and practical significance of the experiment)

**Choosing an appropriate statistical test**

Once the interaction data of the Control and Experimental groups are collected, you can test the statistical hypothesis earlier by choosing an appropriate statistical test that is usually categorized in **parametric** and **non-parametric** tests. The choice of the test depends on the following factors:

* format of the primary metric (underlying pdf)
* sample size (for CLT)
* nature of the statistical hypothesis (show that a relationship between two groups merely exists or identify the type of relationship between the groups)

The most popular **parametric** tests that are used in A/B testing are:

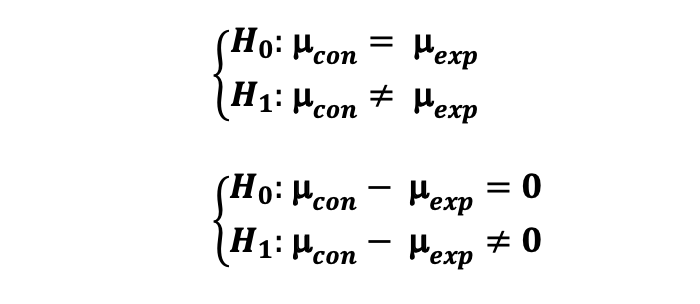
* **2 Sample T-test** (when N < 30, metric follows *student-t* distribution, and you want to identify whether there exist a relationship and *the type of relationship* between control and experimental groups)
* **2 Sample Z-test** (when N > 30, metric follows asymptotic *Normal* distribution and you want to identify whether there exist a relationship and *the* *type of relationship* between control and experimental groups)

The most popular **non-parametric** tests that are used in A/B testing are:

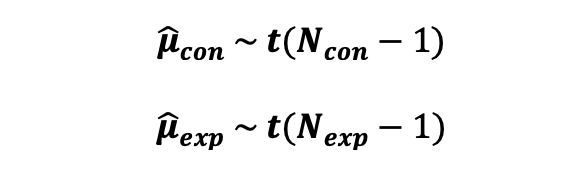
* **Fishers Exact test (**small N, identify and you want to identify whether there exist a relationship between control and experimental groups**)**
* **Chi-Squared test** (large N, identify and you want to identify whether there exist a relationship between control and experimental groups)
* **Wilcoxon Rank Sum/Mann Whitney test**(small N or large N, skewed sampling distributions, testing the difference in medians between control and experimental groups)

**2-sample T-test**

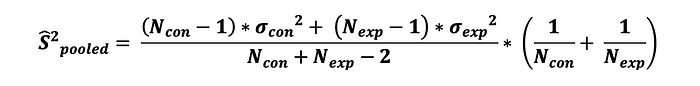
If you want to test whether there is a statistically significant difference between the control and experimental groups’ metrics that are in the form of averages (e.g. average purchase amount), metric follows *student-t* distribution and when the sample size is smaller than 30, you can use 2-sample T-test to test the following hypothesis:



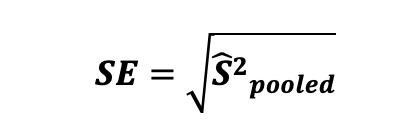
where the sampling distribution of means of Control group follows Student-t distribution with degrees of freedom N\_con-1. Moreover, the sampling distribution of means of the Experimental group also follows the Student-t distribution with degrees of freedom N\_exp-1. Note that, the N\_con and N\_exp are the number of users in the Control and Experimental groups, respectively.



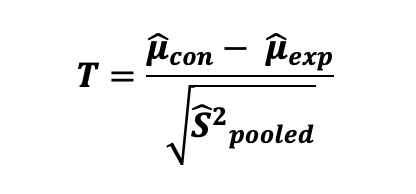
Then an estimate for the **pooled variance** of the two samples can be calculated as follows:



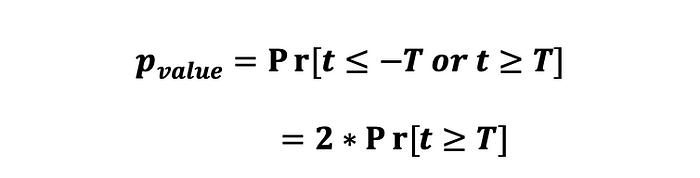
where σ²\_con and σ²\_exp are the sample variances of the Control and Experimental groups, respectively. Then the **Standard Error** is equal to the square root of the estimate of the pooled variance and can be defined as:



Consequently, the **test statistics** of the 2-sample T-test with the hypothesis stated earlier can be calculated as follows:

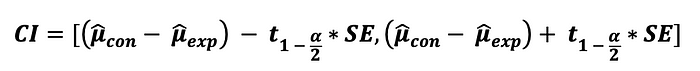


In order to test the **statistical significance** of the observed difference between sample means, we need to calculate the **p-value** of our test statistics. The p-value is the probability of observing values at least as extreme as the common value when this is due to a random chance. Stated differently, the p-value is the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the null hypothesis is true. Then the p-value of the test statistics can be calculated as follows:



The interpretation of a *p*-value is dependent on the chosen significance level, alpha, which was chosen before running the test during the *power analysis*. If the calculated *p*-value appears to be smaller than equal to alpha (e.g. 0.05 for 5% significance level) we can reject the null hypothesis and state that there is a statistically significant difference between the primary metrics of the Control and Experimental groups.

Finally, to determine how accurate the obtained results are and also to comment about the practical significance of the obtained results, you can compute the **Confidence Interval** of your test by using the following formula:



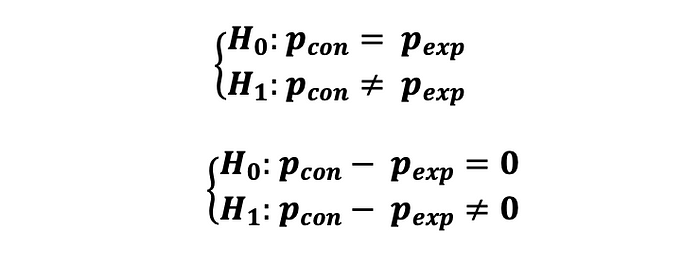
where the t\_(1-alpha/2) is the critical value of the test corresponding to the two-sided t-test with alpha significance level and can be found using the [t-table](https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf).

**2-sample Z-test**

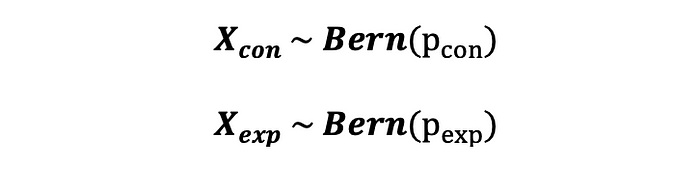
If you want to test whether there is a statistically significant difference between the control and experimental groups’ metrics that are in the form of averages (e.g. average purchase amount) or proportions (e.g. Click Through Rate), metric follows *Normal* distribution, or when the sample size is larger than 30 such that you can use Central Limit Theorem (CLT) to state that the sampling distributions of Control and Experimental groups are asymptotically Normal, you can use 2-sample Z-test. Here we will make a distinction between two cases: where the primary metric is in the form of proportions (e.g. Click Through Rate) and where the primary metric is in the form of averages (e.g. average purchase amount).

**Case 1: Z-test for comparing proportions (2-sided)**

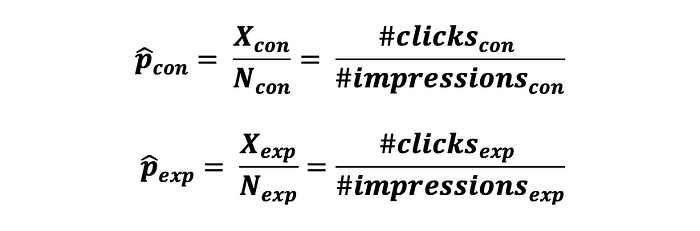
If you want to test whether there is a statistically significant difference between the Control and Experimental groups’ metrics that are in the form of proportions (e.g. CTR) and if the click event occurs independently, you can use a 2-sample Z-test to test the following hypothesis:



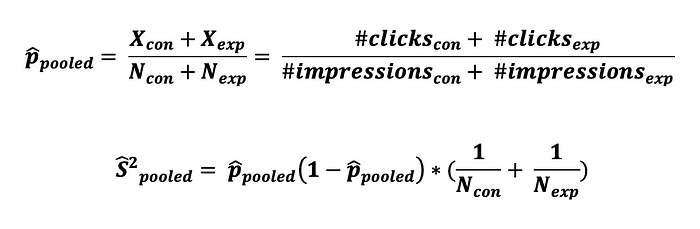
where each click event can be described by a random variable that can take two possible values 1 (success) and 0 (failure) that follows a Bernoulli distribution (click: success and no click: failure) with p\_con and p\_exp are the probabilities of clicking (probability of success) of Control and Experimental groups, respectively. That is:



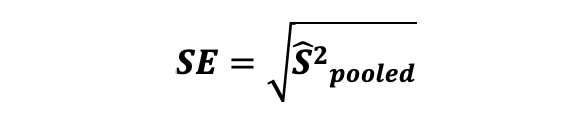
Hence, after collecting the interaction data of the Control and Experimental users, you can calculate the estimates of these two probabilities as follows:



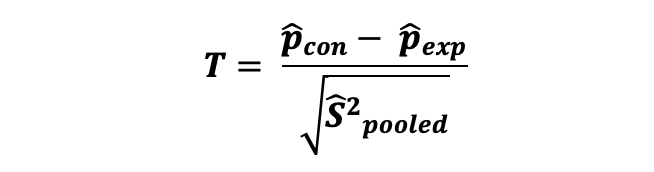
Since we are testing for the difference in these probabilities, we need to obtain an estimate for the pooled probability of success and an estimate for pooled variance, which can be done as follows:



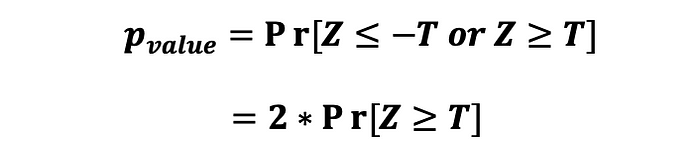
Then the **Standard Error** is equal to the square root of the estimate of the pooled variance and can be defined as:



Consequently, the **test statistics** of the 2-sample Z-test for the difference in proportions can be calculated as follows:



Then the p-value of this test statistics can be calculated as follows:



Finally, you can compute the **Confidence Interval** of the test as follows:

https://miro.medium.com/v2/resize:fit:700/1*ihNRWFh9q9XbC3u-jF-w1g.png

where the z\_(1-alpha/2) is the critical value of the test corresponding to the two-sided Z-test with alpha significance level and can be found using the [Z-table](http://www.z-table.com/). The rejection region of this two-sided 2-sample Z-test can be visualized by the following graph.

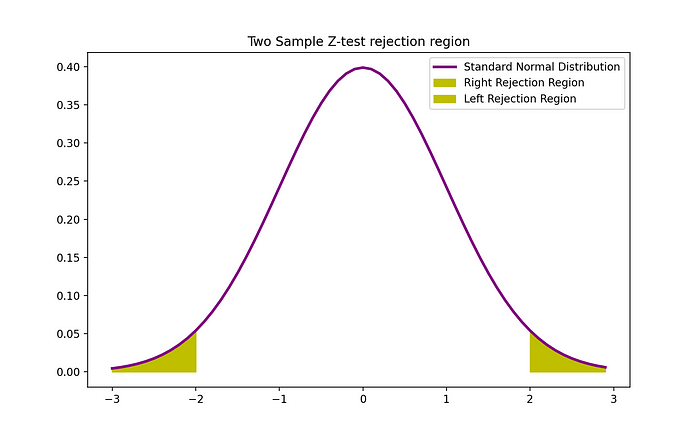
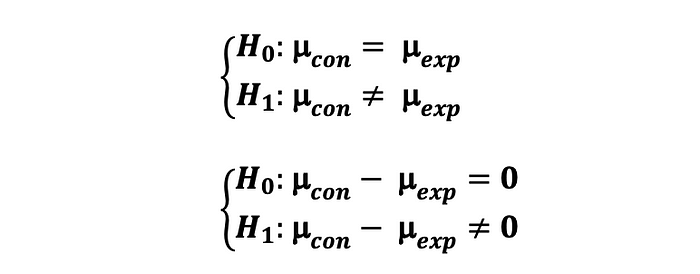


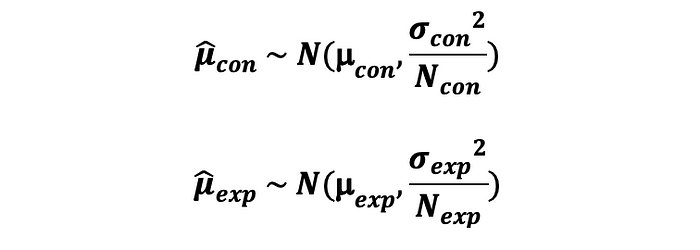
Image Source: The Author

**Case 2: Z-test for comparing means (2-sided)**

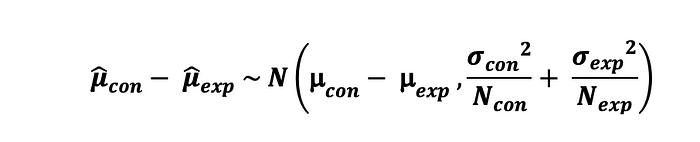
If you want to test whether there is a statistically significant difference between the Control and Experimental groups’ metrics that are in the form of averages (e.g. CTR) you can use a 2-sample Z-test to test the following hypothesis:



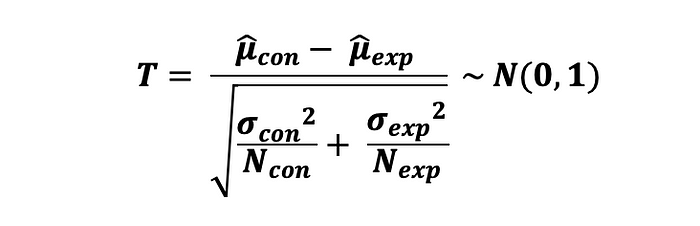
where the sampling distribution of means of Control group follows Normal distribution with mean mu\_con and σ²\_con/N\_con. Moreover, the sampling distribution of means of the Experimental group also follows the Normal distribution with mean mu\_exp and σ²\_exp/N\_exp.



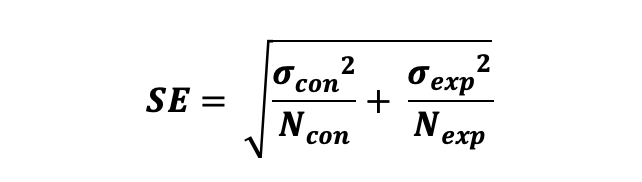
then the difference in the means of the control and experimental groups also follows Normal distributions with mean mu\_con-mu\_exp and variance σ²\_con/N\_con + σ²\_exp/N\_exp.



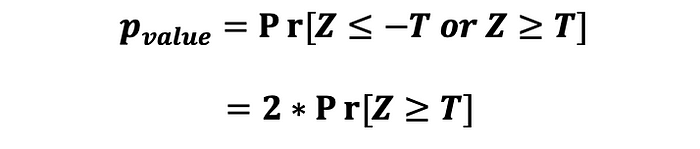
Consequently, the **test statistics** of the 2-sample Z-test for the difference in means can be calculated as follows:



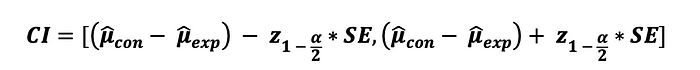
The **Standard Error** is equal to the square root of the estimate of the pooled variance and can be defined as:



Then the p-value of this test statistics can be calculated as follows:

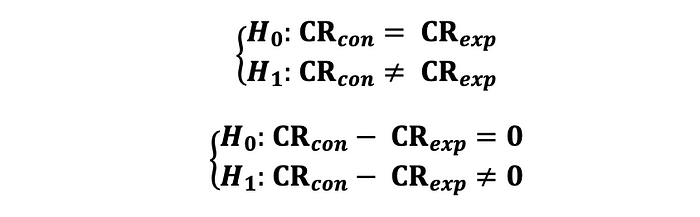


Finally, you can compute the **Confidence Interval** of the test as follows:

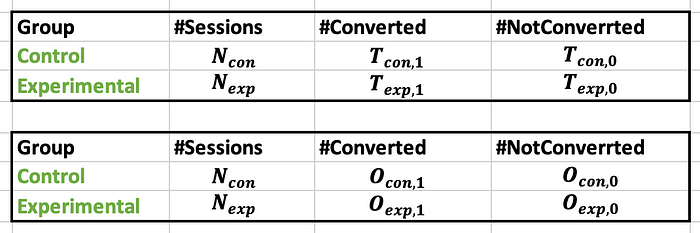


**Chi-Squared test**

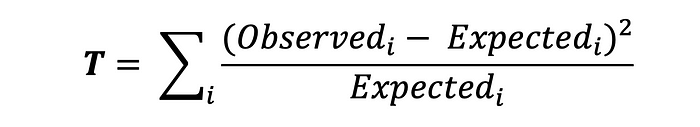
If you want to test whether there is a statistically significant difference between the Control and Experimental groups’ performance metrics (for example their conversions) and you don’t really want to know the nature of this relationship (which one is better) you can use a Chi-Square test to test the following hypothesis:



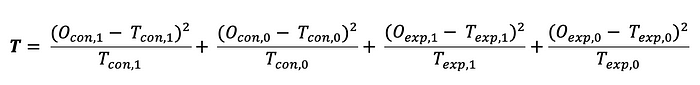
Note that the metric should be in the form of a binary variable (e.g. conversion or no conversion/click or no click). The data can then be represented in the form of the following table, where O and T correspond to observed and theoretical values, respectively.



Then the test statistics of the Chi-2 test can be expressed as follows:



where the *Observered*corresponds to the observed data and the *Expected* corresponds to the theoretical value, and i can take values 0 (no conversion) and 1(conversion). It’s important to see that each of these factors has a separate denominator. The formula for the test statistics when you have two groups only can be represented as follows:



The expected value is simply equal to the number of times each version of the product is viewed multiplied by the probability of it leading to conversion (or to a click in case of CTR).

Note that, since the Chi-2 test is not a parametric test, its Standard Error and Confidence Interval can't be calculated in a standard way as it was done in the parametric Z-test or T-test.

The rejection region of this two-sided 2-sample Z-test can be visualized by the following graph.

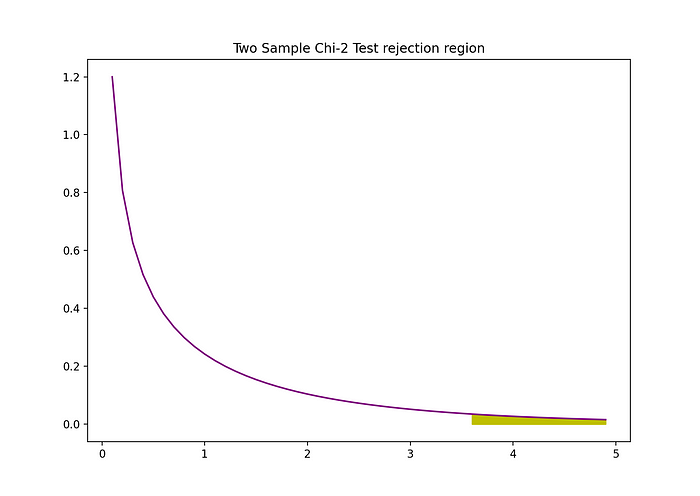
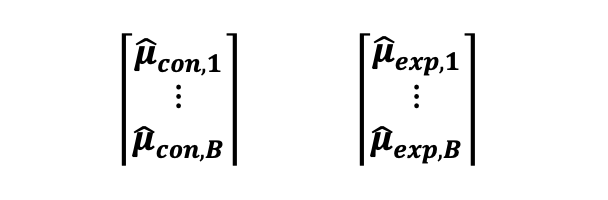


Image Source: The Author

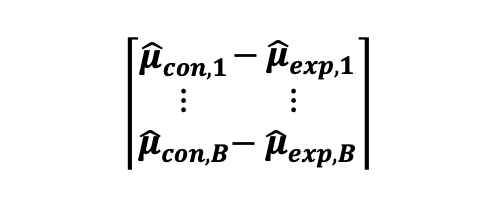
**Standard Error and Confidence Interval for Non-parametric Tests**

In the case of parametric tests, the calculation of Standard Error and Confidence Interval is straightforward. However, in the case of Non-parametric tests, the calculation is no longer straightforward. To calculate the Standard Error and the Confidence Interval of a non-parametric statistical test that aims to compare the sample means or sample medians of control and experimental groups, one needs to use resampling techniques such as ***Bootstrapping*** and ***Boostrap Quantile method,****respectively*.

What Bootstrap does is that it takes the original training sample and resamples from it by replacement, resulting in B different samples. So, the idea behind Bootstrapping is to *resample with replacement (*the same observation can occur more than once in the bootstrap data set*)*from the existing data of the two groups, data that was collected during the experiment, B times which means you will end up with B samples for two groups. Then you need to calculate the sample means/medians for the control and experimental groups B times which can be presented by the following Bx1 vectors:



Consequently, you can calculate the difference in sample means for each pair of control and experimental groups resulting in B sample mean differences. You can then also draw the sampling distribution of this difference in a sample mean, which can be presented by the following Bx1 vector:



Then, if the B is large (for example B = 1000), we can make use of the Central Limit Theorem and assume that the sampling distribution of difference of means of Control and Experimental group follows Normal distribution, as it can be seen from the following graph.

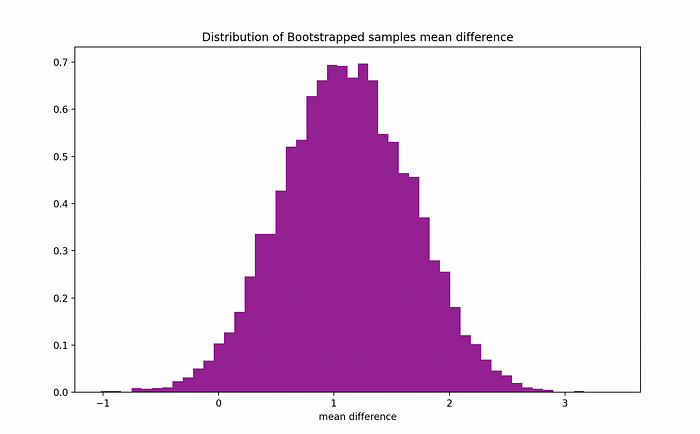
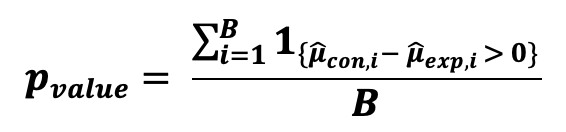


Image Source: The Author

We then need to count the number of times out of this B times, the difference in means is larger than 0 to obtain the p-value of this test as follows:



If the p-value is larger than the chosen significance level then we can state that we can’t reject the null. Hence, there is not enough evidence to state that there is a statistically significant difference in Control and Experimental sample means. *The same test with Bootstrapping can be performed for sample medians.*

To calculate the 95% Confidence Interval, one can use the **Percentile Method,**which uses the 2.5th and 97.5tth percentile of the bootstrap distribution of estimates as the lower and upper bounds of the interval.

If the p-value is larger than the chosen significance level then we can state that we can’t reject the null. Hence, there is not enough evidence to state that there is a statistically significant difference in Control and Experimental sample means.

**Statistical Significance vs Practical Significance**

During the statistical analysis phase of the A/B testing, when a small p-value is detected, then we speak about statistical significance. However, only statistical significance is not enough to make a recommendation about launching a feature or a product.

After statistical significance is detected, the next step is to understand whether there is a ***practical significance.***This will help us to understand whether the detected difference in the performances of the two groups is large enough to justify the investment or it's too small and making a launch decision is not worth the investment.

One way to determine whether the A/B test has practical significance is to use the Confidence Interval and compare its lower bound to the MDE (estimate of the economic significance). More specifically, if the lower bound of CI is larger than the MDE (delta), then you can state that you have a practical significance. For example, if the CI = [5%, 7.5%] and the MDE = 3% then you can conclude to have a practical significance since 5% > 3%.

Note that, you should also look at the width of the CI and make sure it’s not too big since too wide CI gives you an indication that the precision of your results is small and the results will not be generalizable to the entire population (External Validity).

**A/B Test Quality**

A/B testing is one example of an experimental design, and like with any other type of experiment, there are also 3factors needed to be satisfied to make solid conclusions and product decisions. Those factors are:

* **Reliability/Replicability**
* **Validity**
* **Potency**

**Reliability and Replicability**

The idea behind reliability is that the experimental results must be more than a one-off finding and be inherently reproducible and repeatable. Recently, the phenomenon of the *Replicability Crisis*has emerged in the research industry, since researchers are unable to recreate the experimental results. This can happen due to different reasons such as:

* when the original experiment was altered or there was a p-hacking
* when there was a measurement error in the original experiment
* when there was a systematic error in the original experiment
* lack of documentation or source code/data used to perform the experiment

Things you can do to increase the replicability of your A/B experiment:

* Store the source code with comments on a secure cloud
* Store the data with comments on a secure cloud
* Make detailed documentation of the process and results
* Check for the systematic errors (the way treatment responses are reported, so how you measure the impact of the treatment)
* Do the same analysis for another country
* In case you use sampling techniques or simulations, use a *random seed*

All these steps will make your work more reproducible.

**Validity**

The Validity encompasses the entire concept of your experiment and establishes whether the obtained results meet all of the requirements of the randomized control trials or not. In the case of Validity, we usually make a distinction between two types of Validity:

* **Internal Validity**
* **External Validity**

**Internal Validity** refers to observed data and results obtained with it. Are those results valid and reliable or they are inaccurate and biased? Are the changes in the dependent variable only due to the intervention (the independent variable) and not due to other factors? Following examples of problems that can negatively affect the internal validity of your A/B experiment:

* Omitted Variable Bias (use Heckman 2 Step Process)
* Reverse Causality (use IV or 2SLS Approaches)
* Spurious Variable (find control variable or instruments for target variable)
* Use of inappropriate Surrogate Variables (use the actual intervention variable)

**External Validity**refers to the degree to which your experimental results are generalizable to the entire population. It answers the question: Can the results be generalized to the wider population? External validity can be improved by replicating experiments that are repeating the experiment under similar conditions? Following examples of problems that can negatively affect the external validity of your A/B experiment:

* Biased Sample (use [**solid sampling**](https://github.com/TatevKaren/mathematics-statistics-for-data-science/tree/main/Sampling%20Techniques) technique to randomly sample unbiased sample)
* Unrepresentative Sample (use advanced statistical sampling techniques such as [**Weighted or Stratified Sampling**](https://github.com/TatevKaren/mathematics-statistics-for-data-science/tree/main/Sampling%20Techniques) to generate a sample that is not only unbiased but also representative of your population)

Especially, if your population is divided into several subpopulations that somehow differ, and the research requires each subpopulation to be equal in size, [**stratified sampling**](https://github.com/TatevKaren/mathematics-statistics-for-data-science/tree/main/Sampling%20Techniques) can be very useful. In this way, the units in each subpopulation are randomized, but not the whole sample. The results of an experiment are then generalized reliably from the experimental units to a larger population of units.

You can also use Boostrapping to calculate the *Standard Error/Margin of Error* of your results and the width of the *Confidence Interval*. Namely, if the SE of your A/B test is large or CI is wide, then you can conclude that the precision of your results is low and your results will not generalize when applied to the entire population.

**Potency**

It is important to ensure that the intervention is of sufficient **potency** to produce a measurable change in the dependent variable otherwise you have assumed incorrectly that an intervention has no effect (*type II error*). Alternatively, this means that the dependent variable must be **sensitive** to the treatment. The sensitivity can be improved by reducing the noise (e.g. *measurement error*), by for example by making replicate measurements and averaging them (e.g. *Bootstrapping***)**

If the SE of your A/B test is large or CI is wide, then you can conclude that the precision of your results is low and your results will not generalize when applied to the entire population.

**Common problems and pitfalls of A/B tests**

In order to not fail your online experiment, it’s important to follow the specified guidelines and to patiently go through the list of actions that should occur to end up with a well-prepared and executed A/B experiment. Following will be presented common problems and pitfalls of A/B testing that is made frequently with their corresponding solutions.

**Confounding Effects**

It’s important to ensure that all other known possible factors that also have an impact on the dependent variable are held constant. Therefore, you need to control for as many unwanted or unequal factors (also called *extraneous variables*) as possible. Extraneous variables matter when they are associated with both the independent and the dependent variables. One special and extreme case of this problem occurs when the relationship between the independent and dependent variable completely changes/inverts when one takes into account certain spurious variables, this is often referred to by **Simpson’s Paradox**

The reason why one needs to control for these effects is that assigning units to treatments at random tends to mitigate confounding, which makes effects due to factors other than the treatment appear to result from the treatment. So, confounding effects threaten the Internal Validity of your A/B experiment. The following solutions might help you to avoid this problem.

* Control of confounding variables
* Reliable instruments (IV or 2SLS estimation)
* Appropriate choice of independent and dependent variables
* Generation of a random sample

**Selection Bias**

One of the fundamental assumptions of A/B testing is that your sample needs to be unbiased and every type of user needs to have equal probability to be included in that sample. If by some error, you have excluded a specific part of the population. (e.g. sampling for the average weight of the USA by only sampling one state: last time example about education) then we call this a Selection Bias.

*To check whether your sample is biased*while knowing the true population distribution, you create B bootstrapped samples from your sample and draw the distribution of sampling means. If this distribution is not centered around the true population mean then your sample is biased and you should use more [**solid sampling**](https://github.com/TatevKaren/mathematics-statistics-for-data-science/tree/main/Sampling%20Techniques) techniques to randomly sample an unbiased sample.

**Systematic Bias**

This problem relates to the way one measures the impact of the treatment (a new version of the product or a feature). Are you systematically making errors when measuring it? This type of error always affects measurements the same amount or by the same proportion, given that a reading is taken the same way each time, hence it is predictable. Unlike the random error which mainly affects the precision of the estimation results, the systematic error affects the *accuracy* of the results.

**Early Stopping or P-hacking**

A common mistake in an A/B experiment is to stop the experiment early once you observe a statistically significant result (e.g., a small p-value) while the significance level and all other model parameters are predetermined in the Power Analysis s stage of the A/B testing and assume that the experiment will run until the minimum sample size is achieved.

P-hacking or early stopping affects the Internal Validity of the results and makes them biased and it also leads to false positives.

**Spillover or Network Effects**

This problem usually occurs when an A/B test is performed on Social media platforms such as Facebook, Instagram, TikTok but also in other products where the users in the experimental and control group which connected, for example, are in the same group or community and influence each other’s response to the experimental and controlled product versions. This problem leads to biased results and wrong conclusions since it violates the integrity of the test and control effects.

To detect Network Effects, you can perform*Stratified Sampling* and then divide this into two groups. Then you can run an A/B test on one sample taking into account the clustered samples and the other one without. If there is a difference in the treatment effects, then there is a Network Effect problem.

**Change Aversion and Novelty Effects**

When you are testing significant changes on the product and the user doesn’t want that, at first the users might try it out just out of curiosity, even if the feature is not actually better than the controlled/current version, this is called *Novelty Effect* and it affects the internal validity of your results. Moreover, new features (experimental product version ) might also affect the overall user experience making some users churn since they don’t like this new version. This phenomenon is often referred to as *Change Aversion*.

One of the most popular used ways to check for the Novelty Effect is by segmenting users in *new vs old*. If the feature is liked by returning users, but not by new users, then most likely you are dealing with Novelty Effect.

**Sample Ratio Mismatch**

If it appears to you that the split between control and experimental looks suspicious suggesting that the treatment assignment process looks suspicious as more users are assigned to the control/experimental groups than to the experimental/control, then you can perform a Chi-square test. This test will help you to formally check for Sample Ratio Mismatch. You can read more about this test [here](https://www.searchdiscovery.com/blog/sample-mismatch-ratio/).

**Inadequate Choice of Test Period**

Another common mistake in A/B testing is the choice of the test period. As mentioned earlier, one of the fundamental assumptions of A/B testing is that every type of user needs to have equal probability to be included in that sample. However, if you run your test in a period that doesn’t take into account holidays, seasonality, weekends, and any other relevant events then the probability of the different types of users being selected is no longer the same (for example weekend shoppers, holiday shoppers, etc.). For example, running a test Sunday morning is different than running the same test on Tuesday at 11 pm.

**Running too many tests at the same time**

When you have more than 1 experimental variant for your product that you want to test such that you have multitasking where more than 2 variants are presented then you can no longer use the same significance level to test for the statistical significance. So, the p-value or the significance level that the results will be compared to needs to be adjusted.

In this case, one can use ***Bonferonni Correction*** to adjust that significance level based on the number of samples n. So, the significance level that needs to be used in multivariate testing should be alpha/N. For example, if the significance level is 5% then the new adjusted significance level should be 0.05/n.

One of the most popular ways to check for the Novelty Effect is by segmenting users in *new vs old*. If the feature is liked by returning users, but not by new users, then, most likely, you are dealing with Novelty Effect.

**Ethics and Privacy In A/B testing**

With the rise of the popularity of A/B testing, there have also been raising concerns with privacy and ethics behind A/B testing. Namely, the following questions come in handy.

* Are the users informed in terms of conditions and risks?
* What user identifiers are attached to the data?
* What type of data is collected? (personal, voluntarily consent)
* What is the level of confidentiality and security of the test and gathered data, does everyone know this?

When conducting an A/B experiment, you want to make sure that you don’t extract the rights of the users of the treatment (better product/better feature) when providing it o one set of users (experimental) and not providing it to another sample of users(control). Moreover, other issues to consider are what the other alternative services that a user might have, and what the switching costs might be, in terms of time, money, information, etc.

For example, if you are testing changes to a search engine, participants always have the choice to use another search engine. The main issue is that the fewer alternatives that participants have, the more issue that there is around coercion and whether participants really have a choice in whether to participate or not and how this balances against the risks and benefits.