# A/B Testing Key Characteristics

* Controlled deliberate experiment.
* Based on randomized set of users.
* Defined hypothesis and ability to measure the success.
* There is a control group with default experience and at least one test group with new experience.

# Steps of the A/B Testing

1. Define a hypothesis.
2. Decide how to design the feature.
3. Build the feature.
4. Determine how to measure success of the feature.
5. Determine how to run the A/B test:
   1. For how long?
   2. How many users do we want to assign to the test?
   3. How do we monitor the performance of the test?
6. Analysing the results of the A/B Test.

# Define the Hypothesis

## What is a hypothesis?

* It is a statement that you will support or disprove with the A/B test.
* The base hypothesis (alternative hypothesis, H1) is usually phrased as what you expect would happen when you introduce the change you are testing.
* For the hypothesis testing, you cannot prove your base Hypothesis (H1), instead you need to disprove the Null Hypothesis (H0), being this one the opposite of the base hypothesis.

In this case, the hypothesis looks like this:

* **H0:** Tailored ads will have no effect on the user engagement with the ads, and will not affect CTR.
* **H1:** With the introduction of tailored ads, users will be more likely to click on the ads, meaning that the CTR per user will increase.

The **success metric** is based on the hypothesis, being the **dependent variable**. For this test, the dependent variable is the CTR. So, this will be our success metric.

# Defining GUARDRAIL Metrics

Apart from having a success metric, it is very common to have some guardrail metrics. We don’t only want to affect some metrics in a positive way, but we also want to make sure that some critical business metrics are not affected in a negative way. So, apart from the clear dependent variable, we can have other metrics that will also be impacted by the change we are introducing. Some of them could be very business critical.

Usually, we want to control the test impact for the business-critical metrics. We would define the non-inferiority margin for those critical metrics. How much can it decrease/increase depending on the variable so that we still consider the test a success?

It is very common to use retention and engagement metrics as the guardrail metrics as well as some more technical metrics as number of errors or crashes.

# How to know if an A/B test was successful? Defining KPIs and metrics

Metrics for digital products:

* Engagement based metrics:
  + Number of users.
  + Number of downloads.
  + Number of active users (daily, weekly, monthly)
  + User retention.
  + User engagement.
* Revenue and monetization metrics:
  + Ads and affiliate links.
  + Subscription-based.
  + In-app purchases.
  + Other revenue streams.
* Technical metrics:
  + Service level indicators (latency, downtime of the app, uptime of the app…)

# How do we set up the A/B test?

Deciding on group sizes and duration. For the decision on how long to run the A/B test, we need to know how many users should be exposed to the change to be able to reach statistical significance, so we can confirm that the results of the test are not an accident but deliberate and they represent the actual behaviour of the users.

To decide the test duration and group size (control and test group) we need to know:

* Minimum detectable effect
* Power (β) type 2 error rate
* Significance level (p-value) type 1 error rate

## Minimum Detectable Effect (MDE)

Knowing how many users we have, with the current mean and variance of our success metric (and guardrail metric), and what our business needs are, we can define the Minimum Detectable Effect (MDE).

* MDE is a mix of data analysis, statistics and business acumen (knowledge of the business).
* There’s often a trade-off between precision and practicality.
* We may want to have the smallest possible MDE, but with the restriction of our user base.

We want the MDE to be bigger than the standard deviation of the metric. Since the standard deviation indicates how much do values of the metric tend to deviate from the mean (average) value of the metric.

|  |
| --- |
| For example, we have 4 users with Click Through Rate (CTR) of: 30%, 32%, 32%, 34%.   * The average CTR is 32%. * The deviation of each value from the mean is:   This means that, at least, we want to obtain a difference of 1.5% between the test and control groups. |

For our project, we have the following data regarding Daily Active Users (DAU):  
A screenshot of a computer

Description automatically generated

This means that, at least we want to obtain a difference of 91 active users between the control and test groups, that is around 0.3% difference (91/30673.387).

About the second metric, the CTR:

A screenshot of a computer

Description automatically generated

Meaning that we are looking for a difference of 1.8 between the test and control groups, which can be translated in around 5.45%.

# Statistical Significance

How can we be sure that the test results will repeat?

We define the level of statistical significance (1-α) to set an acceptable level of mistakenly rejecting the Null Hypothesis – thinking we see the impact we expect when there’s none (False Positive).

# Power of the test (β)

How can we know that we are not rejecting a successful test?

We define the test power (1-β) to set a level of mistakenly accepting the Null Hypothesis – thinking we don’t have an impact when there’s an impact (False Negative).

* Β usually takes values between 10 and 20

# Type I and Type II Errors

A chart with text overlay

Description automatically generated

A diagram of a scientific experiment

Description automatically generated with medium confidence

# Z-Test

* A z-test is a statistical test to determine whether two population means are different when the variances are known, and the sample size is large (over 30 samples). If the sample size is lower or the variance is unknown, another test will be implemented (t-test).
* A z-test is a hypothesis test in which the z-statistic follows a normal distribution. We will need the z-statistic to determine the sample size.
* We expect to have independent observations, on a randomly selected dataset, with a sample size over 30.

# Calculating the sample size

## Binomial metrics

We will use 2 tailed z-test too calculate the needed minimum sample size (N):

* p: pooled proportion =
* : control group mean.
* : test group mean (desired; ).
* in absolute value.

## Continuous metrics

|  |
| --- |
| For the CTR (binomial metric; a user either clicks or not), we would have:  🡪 I have rounded MDE from 0.018 to 0.02  The test would be run for 8797 users in each group, if groups are 50% split. |

|  |
| --- |
| For the number of daily active users (continuous metric), it would be:  This means that, we should run the test for 13 days for each group, if groups are 50% split.  The number of days is large, but this has been calculated for a difference of 100 users, which is a 0.3%. If we aim for a difference of 1%, which will be 307 users, this will be a much more realistic approach:  In this case, we would need to run the test for 2 days in each group, considering that the groups are 50% split. |

# Finalizing the decision on how to run the A/B test

1. Hypothesis is defined.
2. Success and non-inferiority metrics are set, and we can track them.
3. We agreed on a desired significance and power level for the test.
4. We can reach the significance and power in a reasonable time frame.
5. We have agreed on a halt criterion for the test during the monitoring stage.
6. Duration of the test includes the ability to measure long term metrics if needed.
7. We’ve made sure there are no other test/releases that could interfere with the test or the other way round.

# Performance monitoring

Monitor the test performance on critical metrics while it’s running to be able to prevent a potential negative impact on the users and the business.

NOTE: We don’t make decisions whether the test is a success or not (we must let it finish).

If we see alarming trends, like increase in crashes, drop in engagement, we may want to pause the test and investigate the issues.

There can also be a potential negative reaction from the users, so it’s a good idea to monitor Customer Support issues and social media.

# How does the assignment process work?

Usually, a randomised test assignment is done based on the user\_id. However, it is important to make sure there is no pretest bias between groups. Pre-test bias happens when users are not completely randomly shuffled between groups and one of the groups ends up with a significantly different mean of any tracked metric.

To avoid pre-test bias, companies run pre-assignment analysis, where they “assign” users randomly based on multiple different seeds and calculate the difference in the metrics. And then select the seed which has no potential bias.

## How to deal with an impact from PEEKING?

Peeking (or p-hacking) is an issue of multiple test result calculations, the more you check the significance of your success metric, the more likely you are to see a false positive. It can be seen when the p-value reaches the significant level before the end of the test, and we could be tempted to stop the test and take those results as the end result.

For instance, we could be introducing a new feature in our app and, because of the novelty effect, within the first 2 days, users are much more likely to engage with the new feature. P-hacking would occur if we only look at those days with the novelty effect and conclude that this is the impact for the rest of the duration and for the whole population.

# Analysis Steps

1. Look at the assignments. Does the number of users represent the correct precent of the assignment split? Are the number of daily/weekly/etc assignments expected?
2. Look at the pretest metrics. Are the groups evenly distributed across different types of users? Is there any bias?
3. Look at the critical performance metrics. UX critical (crashes, availability, etc) and business critical (revenue, engagement).
4. Calculate significance of the results on the Success Metric and Non-inferiority Margin.
5. Observe any temporal effects on the metrics. Is the impact stable or is there a novelty effect?
6. Summarise the results in an understandable, non-technical way, with the possibility to look at the assumptions made, and methods used for the analysis.
7. Summarise the suggestions for the next steps and potential additional research.
8. Share the results with stakeholders and insights community.