A PRACTICAL GUIDE FOR A/B TESTING

AUTHORS

Yoni Schamroth
Yuval Ben-Zion

CONTRIBUTORS

Eran Abikhzer-Agam
Tal Mizan
Liat Gilead
Matan Lindenbaum

fiverr

CONTENTS

01 SOME THEORETICAL BACKGROUND

1.1 QUICK INTRO

A/B testing is a randomized controlled experiment with two variants, A and B.
It is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective, in terms of the KPI we wish to improve. The test groups, A and B, are identical except for one variation that might affect a user's behavior. Usually group A reflects the current state ("control" group), while group B will posses the mentioned variation ("treatment" group).

The factor that is different between the control and treatment groups (e.g. a new feature developed) is known as the independent variable. This variable is independent because it does not depend on what happens in the experiment. Instead it is something that the experimenter applies or chooses.

In contrast, the dependent variable in an experiment is the response that is measured to see if the treatment had an effect. This is typically the main KPI of interest (e.g. revenue, conversion rate, etc.).

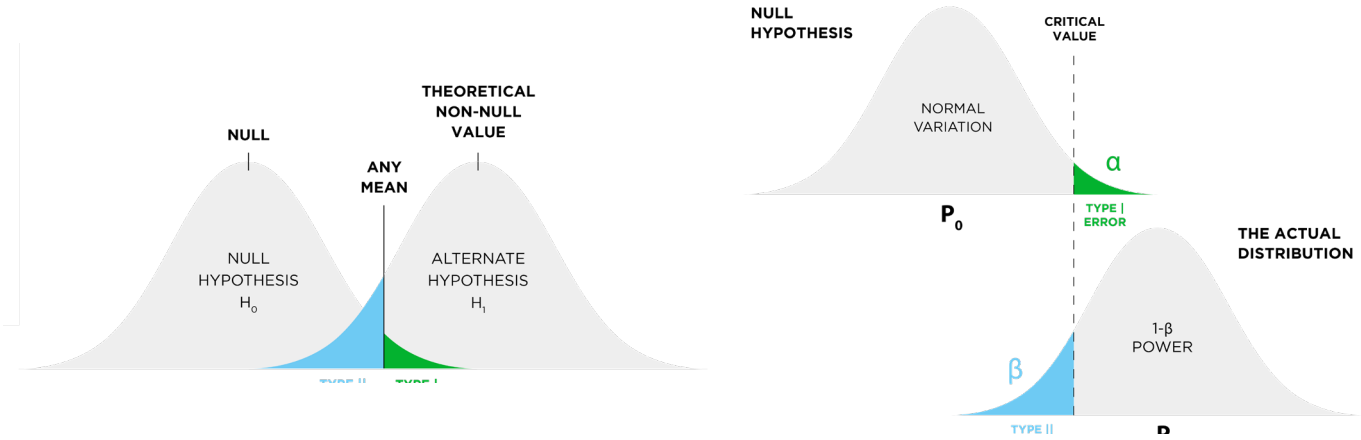
When planning a test (e.g. for a new feature) the hypothesis states that the treatment given to group B will help increase performance (in terms of the test's main KPI).

The way to statistically 'prove' this, is by rejecting the "null hypothesis" (or H0), as opposed to supporting it. For our purposes, the null hypothesis states that there's no difference between group A and group B, or that the metrics for the test's main KPI we observe in each of the groups comes from the same distribution.
The null hypothesis will be rejected in case the metrics observed in group B fall into the rejection region for a pre-specified level of significance, otherwise known as . In other words, they are too extreme/'out of the norm' in terms of group A's distribution. This is called the "alternative hypothesis" or H1.

The statistical power of a test, $1 - \beta$, is defined as the probability of rejecting the null hypothesis when the alternative hypothesis is true. It measures the ability of a test to reject the null hypothesis when it should be rejected. The statistical power ranges from 0 to 1, and as it increases, the probability of making a type 2 error, β (supporting the null hypothesis when it should be rejected; otherwise known as a false negative) decreases. At any given significance level, , the power of the test is increased by having a larger sample size.

The significance level affects type 2 error. It also affects the type 1 error - rejecting the null hypothesis when it shouldn't be rejected, otherwise known as a false positive.

The next diagrams illustrate the ideas mentioned above.



Here is an example for pregnancy test results versus the reality of the situation:

	TEST RESULT: PREGNANT	**TEST RESULT: NOT PREGNANT**
REALITY: NOT PREGNANT	TYPE 1 ERROR (PROBABILITY = α) **FALSE POSITIVE**	PROBABILITY = 1-α **CORRECT NEGATIVE**
REALITY: PREGNANT	POWER (1-β) **CORRECT POSITIVE**	TYPE 2 ERROR (PROBABILITY = β) **FALSE NEGATIVE**

All of the aforementioned, is also relevant for multivariate tests (A/B/C/D.. tests).

Interesting statistics:
- At Microsoft 60% of the experiments fail to show improvement.
- At Amazon 50% of the experiments fail to show improvement.

So, don't feel too bad if your test didn't "succeed"...

"You have to kiss a lot of frogs to find one prince. So how can you find your prince faster? By finding more frogs and kissing them faster and faster" (Moran, 2007).

1.2 THE FIRST A/B TEST

For centuries bloodletting, the withdrawal of blood from a patient to prevent or cure illness and disease, was the accepted form of medical treatment to heal a multitude of ailments. In fact, one British medical journal recommended bloodletting to help heal a list of over one hundred forms of diseases.

In 1836 Pierre Louis decided to test the effects of bloodletting scientifically. He conducted what is now recognized as one of the first A/B tests or randomized controlled experiment in history.
He randomly treated patients suffering with pneumonia either with bloodletting or a less aggressive treatment. At the end of the experiment, he was able to better understand what method was a success, what worked and what didn't.

02 MAIN GUIDELINES

2.1 CONVENTIONS AND DEFINITIONS

- Target population - the population of users being tested..
- Test KPI - a KPI to test the change in. Each test will have a main KPI (and optionally, secondary KPIs).
- Control group - a group that stays “untouchable” in terms of the change being tested (e.g. new feature dev).
- Treatment group - a group which exposed to a new change (e.g. new feature dev).
- *N* - sample size (for each test group).
- Proportional KPIs - CTR, conversion rate, registrations rate, etc.
- Continuous KPIs - orders, revenue, buyers, etc.

2.2 PLANNING A TEST

2.2.1 Test definitions

1. **Target population**

   1.1 Agree on the test’s target population in advance i.e.run the test on new users based only in the U.S. The allocation of users to test groups should be performed only on the target population (in this case only U.S. based users).

2. **KPIs**

   2.1 Agree on the test’s main KPI in advance. This will affect the rest of the test’s planning, running and post analysis.

   2.2 Optional - define the test’s secondary KPIs - those KPIs that can tolerate/not tolerate a decrease in.

   2.3 When launching a test, declare its goal i.e. improve the main KPI, while not damaging the secondary KPIs.

   2.4 When analyzing the test, refer to the chosen main KPI, as the indicator for a “successful test”. Reference the chosen secondary KPIs (if any), to check for any increase or decrease and relate to their part in the test goal.

2.2.2 Sample size

TL;DR

1. Calculate N and "disaster N" using and μ of the target population's main KPI (or highest variance KPI. see 3.1 in "the long version" section below).

   1.1 Minimum sample size (N) = $\left(\frac{1.645 * \sigma_{target\ population}}{0.02 * \mu_{target\ population}}\right)^2$

   1.2 "Disaster" sample size ("disaster N") = $\left(\frac{1.645 * \sigma_{target\ population}}{0.1 * \mu_{target\ population}}\right)^2$

2. Calculate the test's needed period of time according to: (1) Traffic getting to the test area. (2) Minimum sample size. (3) %Rollout planned for the test.

   2.1 Note - reasonable period of time for a test is **at most 2 months**.

3. See also 2.1, 2.2 in "the long version" section below.

The long version

1. Sample size (N) formula

   1.1 $N = \left(\frac{z * \sigma}{E}\right)^2$, where:

   1.2 z- critical value (z-score) (In practice use t-score, although z and t scores are very similar as degrees of freedom increase).

      1.2.1 For 1-tail tests with 0.95 significance level, use **1.645**.

   1.3 σ- standard deviation (In practice, use standard error - an estimation for the standard deviation).

   1.4 E-margin of error. The effect/change we wish to be able to identify.

      1.4.1 Need to translate %change we wish to be able to identify to absolute number. Meaning E = target population's mean * desired %change.

2. "Disaster N"

   2.1 This is the minimum sample size for identifying "disaster" meaning, at least a 10% change (margin of error) between treatment and control.

   2.2 Note - will be smaller than minimum sample size for reaching test significant. Before reaching "disaster N", do not present metrics comparison in A/B tests analysis tool.

   2.3 "Disaster N's" should be calculated on-demand according to the test's domain categories and according to the KPI with the highest variance, in order to get the highest N for being conservative per all KPIs.

3. N

   3.1 N's should be calculated on-demand according to the test's domain categories and according to the test's main KPI (note - usually better to choose revenue, unless it's not interesting in this test's scope, in order to "cover" all desired KPIs). **Note - the greater the KPI's variance, the larger the sample size needed.

   3.2 It is recommended to use an X% depending on traffic volume as the 'margin of error' (treatment-control difference). **Note - the smaller the 'margin of error' %, the larger the sample size needed.

4. Calculate the test's needed period of time according to: (1) Traffic getting to the test area. (2) Minimum sample size. (3) %Rollout planned for the test. **Note - reasonable period of time for a test is **at most 2 months**.

2.2.3 What should you do when you're not supposed to see significance in the near (reasonable) future?

*** This section refers to a situation in which in order to reach the minimum sample size, we need to run a test for a non-reasonable period of time.*

TL;DR
See 1 (and 3, if relevant) in "the long version" section below.

The long version

1. Reasonable period of time for tests is at most 2 months.

2. Minimum sample size is calculated based on the test KPIs:
   - 2.1 When based on the main KPI, the N is required for identifying a significant change for this specific KPI (if exist).
   - 2.2 When based on the KPI that has the highest variance, the largest N is required in order to identify significant change for all KPIs (as the variance of the KPI increases, so does the required minimum sample size. See formula in 'sample size' section).

3. Recommendations for when the above needed period of time > "reasonable period of time (2 months)":
   - 3.1 Find the other main KPI with less variance that's more sensitive, and start the test planning all over again.
     - 3.1.1 Verify that the new main KPI is a proxy for the original main KPI.
   - 3.2 Run the test with the chosen main KPI in order to validate that "there's no disaster." Meaning, plan the test (sample size, etc.) in order to identify an effect of 10% (if exists). This will enable the ability to run a smaller sample size (as margin of error increases, the minimum sample size decreases. See formula in 'sample size' section).
   - 3.3 Carefully consider whether the expected impact of the new change (e.g. feature development) justifies running an AB test.

2.3 RUNNING A TEST

2.3.1 Test groups sampling

TL;DR
See 1 under "the long version" section below.

The long version

1. Key sampling method guidelines:
   - 1.1 Sample users from the test's target population only.
   - 1.2 The sampling and allocation to test groups should occur at the "entrance" to the test zone (e.g. relevant page, search bar, etc.).
   - 1.3 Maintain 'user stickiness' - no migration of users between test groups.
   - 1.4 Validate consistency of each test group experience - no changes of experience during a test.

2. Users sampling mechanism.
   - 2.1 Random sampling:

     Randomly allocate users to test groups. This method is sometimes flawed in terms of biases towards specific test groups, especially when dealing with small populations. Even though the overall

random sampling is unbiased, it may result in an unbalanced samples per specific subpopulations i.e. allocating more users from specific countries to one group. This can be problematic since these users are worth much more than rest of the world.

2.2 Options for alternative sampling method:

2.2.1 Pick a seed for a hash function by which the users are allocated. Analyze the data from the week prior to the experiment. Look at potential problematic segments (e.g. US) and check that control and treatment are not statistically different. If they are, pick a new seed and try again. Remember that if there are 5 metrics and 3 segments, there's a huge chance (1-0.95^15) > 50% that one of them will be off. Performing this check helps verify that the allocation of users will be okay for all metrics and segments.

2.2.2 Extend 2.2.1 by trying 200 seeds and looking for the optimal one under some criteria (e.g. smallest p-value is maximized).

2.2.3 Employ stratified sampling (also useful for planning the test groups allocation to fit the desired subpopulations analysis).

2.2.4 Employ CUPED method: http://bit.ly/expCUPED

2.3.2 Consistency and test duration

*** also relevant for "3. Analyzing a test" section*

TL;DR

1. A test's duration usually lasts between 2 weeks and 2 months.
2. For criteria required to stop a test and end the lifecycle see 1.1, 1.2, 1.3, 2, 3, 4, 5 in the long version section below.

The long version

1. Run the test as needed for reaching minimum sample size (N).

   1.1 Run the test in weekly cycles - 2 weeks/3 weeks/4 weeks/...

   1.2 When reaching "disaster N", check for disaster, and stop/continue the test accordingly.

   1.3 Disclaimer - It's possible to stop a test in case p-value is significant for 3 days in a row, for the test's main KPI.

   1.3.1 Motivation - it's unlikely to get type 1 error (false positive) 3 days in a row, or more.

   1.3.2 Example - identifying 4% increase in main KPI with the above 3 days consistency, WO reaching the minimum sample size (N) that was set originally (e.g. for a 2% increase).

2. Wait at least 2 full weeks (due to seasonality reasons).
3. Wait until "disaster N".
4. Don't keep running a test for more than is needed (according to N and by weekly cycles - the stopping criteria defined).

   4.1 The false positive rate (type 1 error) is slightly inflated as time goes by.

When there is a suspected issue, trend, etc. longer experiments are encouraged.

   4.2 When there is a suspected issue, trend, etc. longer experiments are encouraged.

2.3.3 Interactions between different tests

TL;DR

1. Prevention of strong interactions should be implemented as part of any controlled experiments management tool.
2. Detecting them should be implemented as part of any controlled experiments analysis tool. Alerts should be sent to relevant stakeholders when clear and meaningful interactions are discovered.

The long version

As we increase the number of tests running in parallel, the risk of interactions between different treatments becomes a growing concern. A statistical interaction between two treatments X and Y exists if their combined effect is not the same as the sum of the two individual treatment effects.

1. Prevention -
    1.1 A series of suggested approaches for preventing and detecting interactions between tests, can be found in "Online Controlled Experiments at Large Scale" under section 5.2, by Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, Nils Pohlmann.
    1.2 The above prevention approaches are recommended for implementation as part of any controlled experiments management tool.

2. Detection -
    2.1 Can be implemented as an ongoing detection algo, which will send alerts for tests with strong interaction.
    2.2 Tests that are flagged will likely need to be stopped.

2.3.4 Leakage

*** Also relevant for "3. Analyzing a test" section.*

TL;DR
See the long version section below.

The long version

This section is relevant for cases where there is a migration of users between test groups, wherein the controlled experiment management tool should prevent it from happening.

1. Rule of thumb for users leakage handling:
    1.1 Leakage up to 10% - filter out all the leaked users.
    1.2 Leakage above 10% - test is not valid and should reopen.

2.4 ANALYZING A TEST

2.4.1 Statistical methods

TL;DR
See 1.1, 1.2.1, 2.1 in "the long version" section below.

The long version

Use 1-tail tests with 0.95 significance level, for both proportional and continuous KPIs.

1. 2 group test (A and B)
    1.1 Proportional KPIs - Chi-squared test
    1.2 Continuous KPIs - T test for Independent (unpaired) samples
        1.2.1 Use Welch's t-test
        1.2.2 The above is a two-sample location test which is used to test the hypothesis that two populations have equal means. Welch's t-test is an adaptation of Student's t-test, and is more reliable when the two samples have unequal variances and unequal sample sizes.

2. More than 2 groups (A,B,C,D,...)

    2.1 Pairwise comparison between each treatment group and the control group (A vs. B, A vs. C, A vs. D, etc.).

2.4.2 Outliers

TL;DR
See 1, 2.1, 2.2 in "the long version" section below.

The long version

1. Remove all "non-legitimate" data (bots, frauds, etc.).
2. Employ percentiles approach for capping -

    2.1 Calculate the 99.5% value out of all non-zero values of the relevant KPI, across all groups combined (A,B,C,D,...).

    2.2 Perform capping - change all relevant KPI values above the 99.5% value to be that exact value.

    2.3 This is being done in order to keep the top performing participants in the test, while lowering the KPI's variance.

2.4.3 Confidence intervals

TL;DR

1. Formulas for confidence intervals (CIs) of the difference between test groups are supplied in 1.1, 2.1 in "the long version" section below.
2. In order to transform the CIs to relative difference (in %), divide the CIs bounds by the control group mean value.

The long version

1. Continuous KPIs

    1.1 CI formula for test groups mean difference:

When $\sigma_1 \neq \sigma_2$ are unknown, the appropriate two-sided confidence interval for $\mu_1 - \mu_2$ is

$$\overline{X}_1 - \overline{X}_2 \pm t_{1-\frac{\alpha}{2},v}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

Where

$$v = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_1-1)}}$$

    1.2 In order to translate this to a relative difference between groups (in %) just divide the CI bounds by the control group mean.

2. Proportional KPIs

2.1 CI formula for test groups difference:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \cdot s.e.(\hat{p}_1 - \hat{p}_2)$$

where $s.e.(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

2.2 As in continuous KPIs, divide the CI bounds by the control group P to get the relative difference (in %).

2.3 The fact that this is based on the normal approximation and not chi distribution is not an issue.

2.3.1 Chi distribution is based on the normal distribution, and the P values obtained with each method should be equivalent.

2.4.4 Subpopulations analysis

TL;DR
See "the long version" section below :) ...

The long version

Refer to the process of comparing treatment and control groups per subpopulations in order to identify significant and drastic differences from the overall findings of the test. For example, compare A and B groups for U.S. users only.

1. Perform ANOVA (analysis of variance) to detect significant interaction of the relevant variable (e.g. country) and the test groups, with relation to the tested KPI.

1.1 In case there is significant interaction, perform subpopulation tests per each specific level of the relevant KPI (e.g. USA, Canada, etc.).

1.2 Flag subpopulations whose results are opposite of the general population's results.

2. Analyze A/B test results per each of these subpopulations:

2.1 New users and existing users

2.1.1 Check funnel KPIs for new users (for example, rates of: visit→registration, registration→conversion).

2.2 Specific to marketplaces, can be sellers and non-sellers

2.4.5 Differences in lifetime (prediction)

TL;DR
See "the long version" section below :) ...

The long version
Check the trend in each group in order to flag the possibility that in the future the relation between group A and group B will change.

2.4.6 Adjusted p-value

TL;DR
See 2 in "the long version" section below.

The long version
This section relates to the "multiple comparisons problem".

1. The more inferences made, the more likely erroneous inferences are to occur. Several statistical techniques have been developed to prevent this from happening, allowing significance levels for single and multiple comparisons to be directly compared. These techniques generally require a stricter significance threshold for individual comparisons, so as to compensate for the number of inferences being made. 2 common techniques are:

    1.1 Bonferroni correction is a method that compensates for the increased likelihood of incorrectly rejecting a null hypothesis as the number of groups in the test increases. That's being done by testing each individual hypothesis at a significance level of $\frac{\alpha}{M}$, where α is the desired overall significance level and *M* is the number of hypotheses (treatments, subpopulations analyses etc.).

    1.2 False Discovery Rate (FDR) is a method of conceptualizing the rate of type 1 errors in null hypothesis testing when conducting multiple comparisons. FDR-controlling procedures are designed to control the expected proportion of "discoveries" (rejected null hypotheses) that are false (incorrect rejections).This provides less stringent control of Type 1 errors as compared to familywise error rate (FWER) controlling procedures (such as the Bonferroni correction).

2. Using a global mistake (α)

    2.1 A global mistake (α) will force us to have bigger Ns

    2.2 Note that using a global mistake is less relevant when an A/B/C/D test to be a "3 A/B tests", and we're not interested in controlling the overall false discovery rate (type 1 error). This is typically more suitable when dealing with a much higher magnitude of comparisons.

2.4.7 Normalizing KPIs

TL;DR
See 1.1, 2.1, 3.1 in "the long version" section below.

The long version

1. Continuous KPIs

    1.1 Normalize KPI by users (user_id).

    1.2 For example, when the test main KPI is revenue, we should actually analyze the revenue per user in each of the test groups.

2. Proportional KPIs

    2.1 Make sure to calculate the proportion metric using the right denominator. Find the potential that a proportion metric had.

    2.2 For example, when calculating conversion rate, divide the number of conversions by the number of potential to be converted in the test - group members that entered the test as RNCs (registered-not converted) and guests.

3. Important - Relation/proportion of #allocation→#user_id throughout all test groups

    3.1 Keep track of the above relation to ensure no significant difference between test groups.

    3.2 Theoretically #allocations between test groups should be equal (as the controlled experiments management tool should take care of this), but the #users usually vary.

    3.3 If there is a significant difference, it means the normalization that was previously done by users (user_id) is actually invalid and has skewed the results. In this case the data needs to be re-analysed and normalized by allocations.

2.4.8 Reporting results

TL;DR
See “the long version” section below...

The long version

Key issues to address in the test analysis report:

1. Report on results per the main KPI and the secondary KPIs (if were chosen) only.
2. Report on non-significant results, while emphasizing that they’re not significant.
3. When reporting results, do so while elaborating on the confidence intervals and the mean.
4. When reporting on the test results, make sure to only report on traffic involved in the test, rather than all traffic on the site.

    4.1 For example - 3% revenue increase in a test designed to check a new formula in search mechanism. 50% of the traffic goes through the search box. Meaning that the overall expected affect after 100% roll out of the new formula is %3 * 50% = 1.5%.

5. When reporting on subpopulations, do so while elaborating on the subpopulation size out of test target population and out of all the population of users.
6. Automatically scan test subpopulations and report for “disaster”/”great success”.