Faculty of Computer Science | Data Analysis | Moscow 2025

Lecture 3
Investigating Relationships (part 2)

Lecturer: Alisa Melikyan, amelikyan@hse.ru, PhD,
Associate Professor of the School of Software Engineering

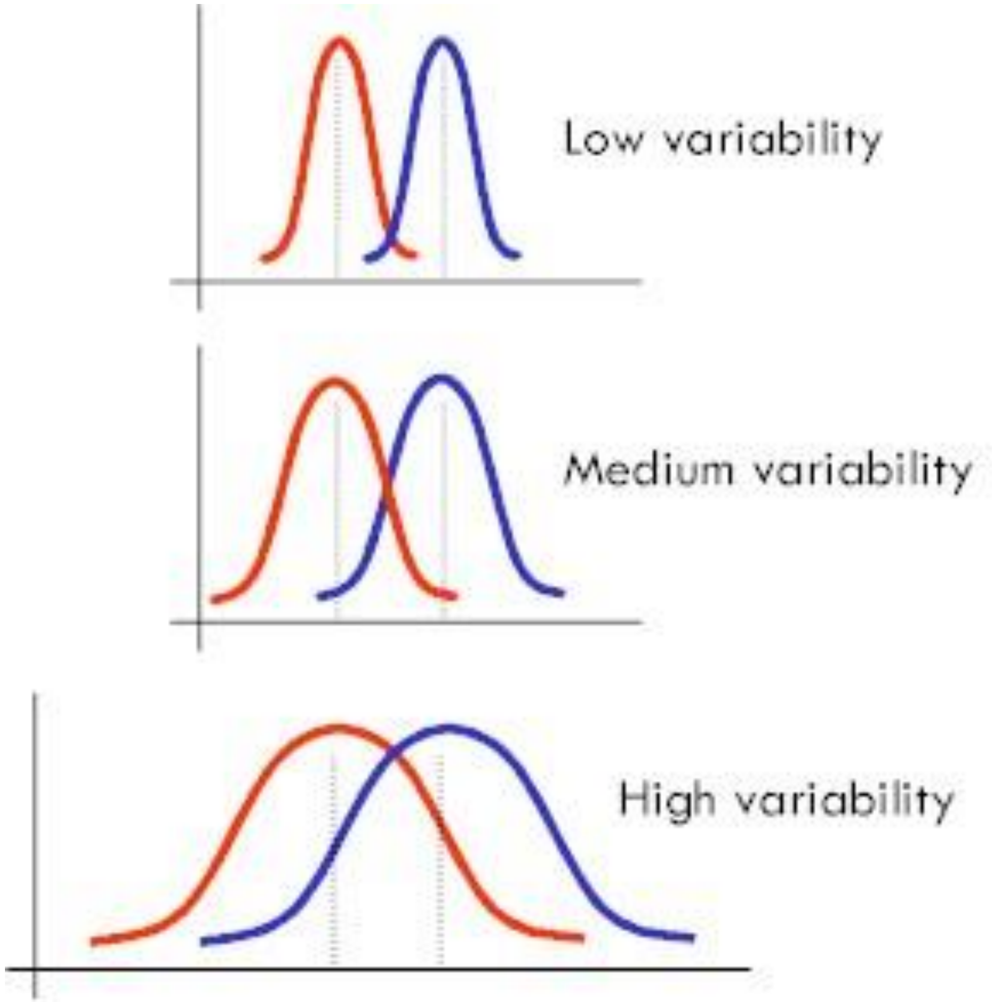
Student’s t-test

A t-test is a statistical test that checks if two means are reliably different from each other.

The values of the means may show a difference, but we can’t be sure if that is a **reliable** difference.

Descriptive statistics only describes the data but can’t be generalized beyond that. Inferential statistics, like t-test, allows to make inferences about the population beyond our data, generalize the findings to a whole population beyond the sample, that is tested.

Same means with different variabilities



t-test

$$t = \frac{\text{variance between groups}}{\text{variance within groups}}$$

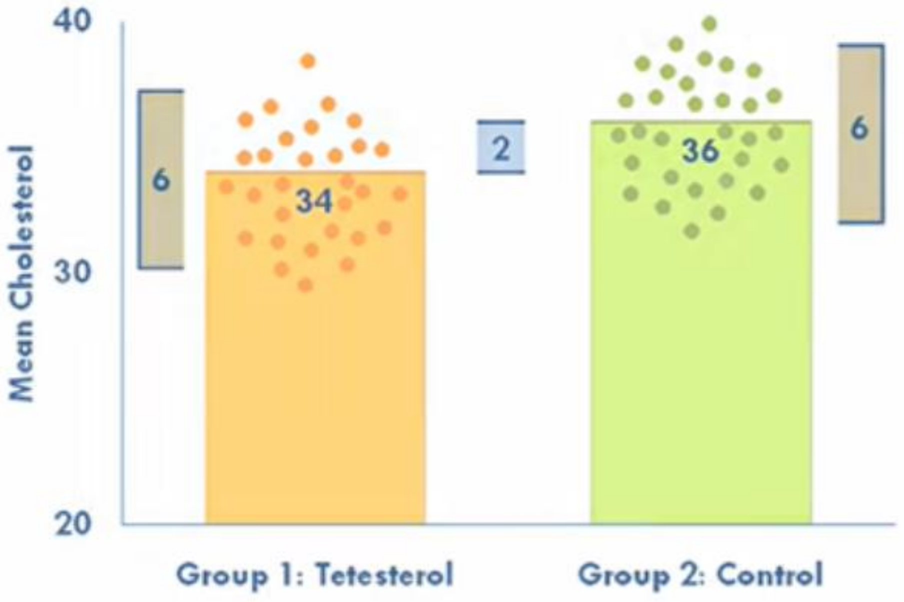
A big t-value = different groups

A small t-value = similar groups

$$\text{t-value} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t-test

$$t = \frac{2}{6}$$



t = 2/6 → the value is not big enough to be reliable → we can't say that there is a statistically significant difference between the means.

t-test

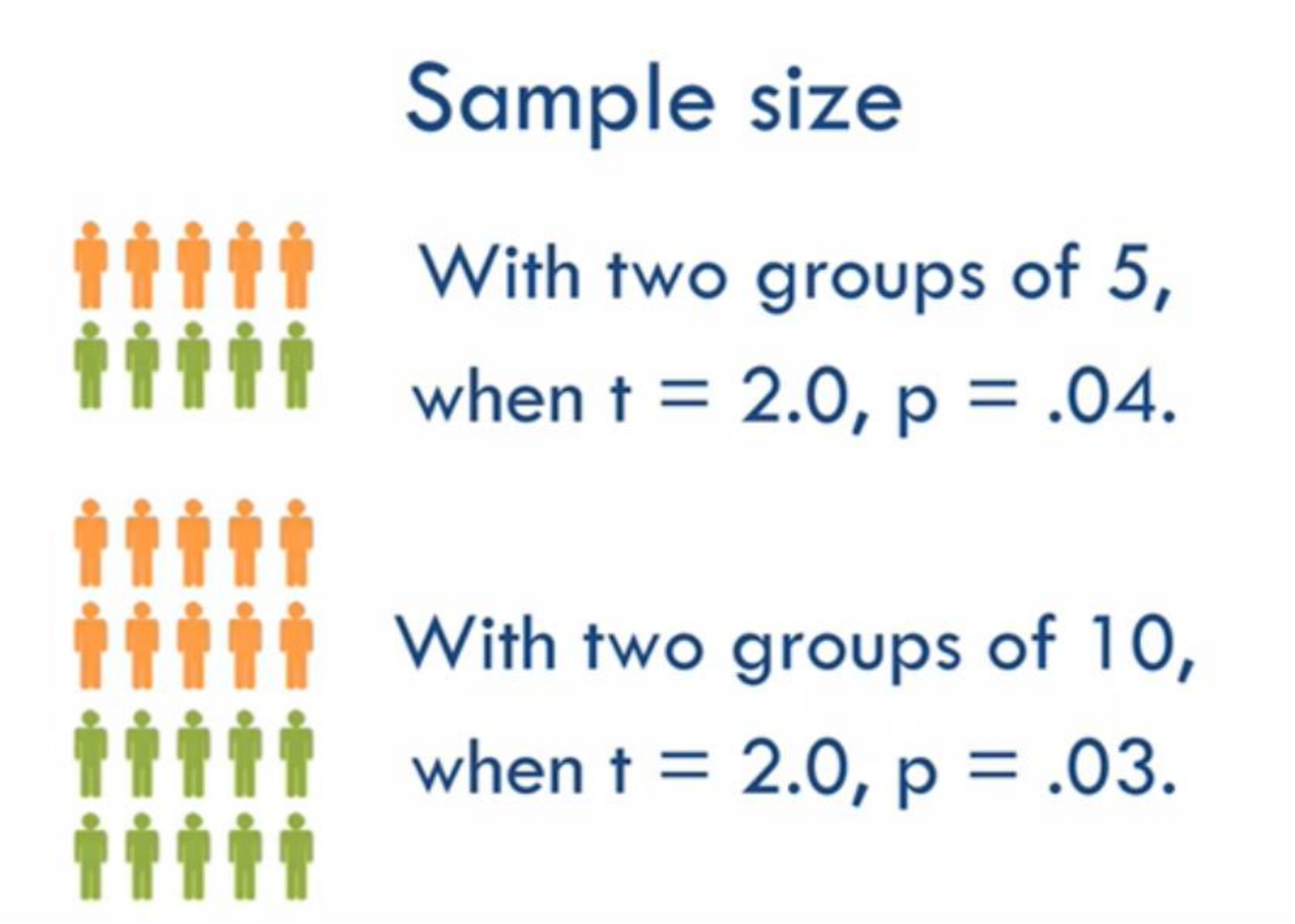
Each t-value has a corresponding p-value, which tells us the likelihood that there is a real difference. P-value is the probability that the pattern produced by our data could be produced by random data. It tells us whether the difference between our groups is real or if it's just a fluke. The p-value for each t-value depends on the sample size. Bigger samples make it easier to detect differences.

If p = .10, there is a 10% chance.

If p = .05, there is a 5% chance there is no real difference.

If p = .01, there is a 1% chance.

t-test


Sample size
With two groups of 5,
when t = 2.0, p = .04.
With two groups of 10,
when t = 2.0, p = .03.

Types of t-test

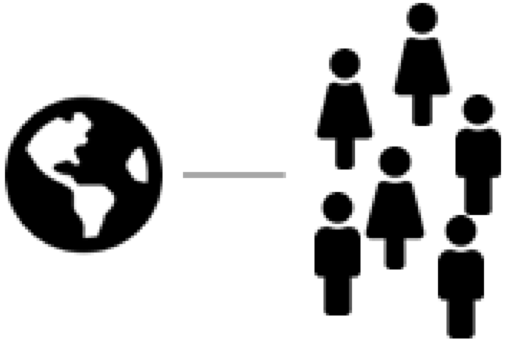
Independent-samples t-test tests the means of two different groups. Is also called between-samples or unpaired-samples t-test.

Paired-samples t-test tests the mean of one group twice. For example, students' knowledge in data analysis before and after taking the course. It's also called within-subjects, repeated-measures or dependent-samples t-test.

One-sample t-test compares the mean in one group with a hypothetical value or known population mean.

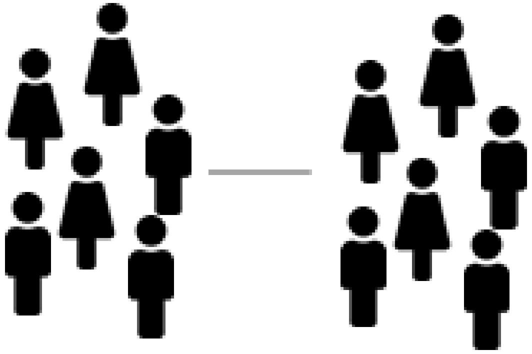
Types of t-test

One sample t-test



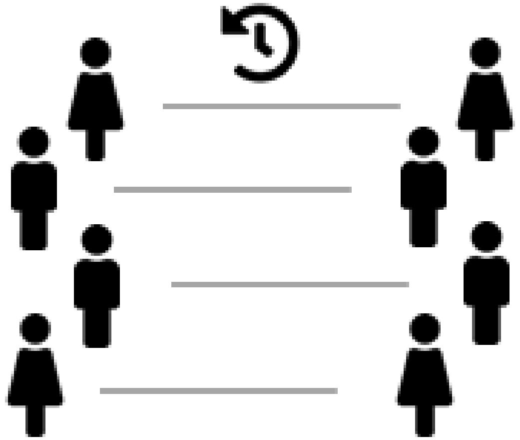
Is there a difference between a group and the population

Unpaired t-test



Is there a difference between two groups

Paired t-test



Is there a difference in a group between two points in time

Hypotheses

Independent-samples t-test

H1: the mean salaries of men and women are different

Paired-samples t-test

H1: the mean salaries of employees of this organization before and after restructuring are different

One-sample t-test

H1: the mean salary of employees of this organization is different from the average salary in the region

Limitations of t-test

1. The results of inferential statistics can only be applied to populations that resemble the sample that was tested.
2. The data should be approximately interval-level or higher.
3. The sample as well as the population should be roughly normal in distribution.
4. Each group should have about the same number of data points, but not less than 20-30.
5. Scores should not be influenced by each other.

Overcoming several limitations

Non-parametric tests do the same job as t-tests but can be applied to non-normal distributions and ordered-level data. But these tests are less powerful.

Independent Samples t-test

The World Bank assigns the world's economies to four income groups – low, lower-middle, upper-middle, and high-income countries.

H1: the **mean** values of the *percent of travel services our of service exports* for <u>high income</u> countries and <u>other</u> countries are different.

Dependent variable: Travel services (% of service exports)

Grouping variable: 1 – high income countries, 0 – other countries.

p-value is 0.0002 < 0.05 → the probability of making an error while rejecting H_0 and accepting H_1 is very low, we will accept H_1 and conclude that there is a statistically significant difference in mean values of the dependent variable in the two groups.

Paired Samples t-test

We have the values of the variable Travel services (% of commercial service exports) for the same countries for two years – 2023 and 2024.

H1: the **mean** values of the *percent of travel services our of commercial service exports* for countries in 2023 and in 2023 are different.

H0: mean values are not different

p-value is 0.43 > 0.05 → the probability of making an error while rejecting H_0 and accepting H_1 is too high, we will reject H_1 and conclude that there is no statistically significant difference in mean values of the variable for two time periods.

ANOVA

ANalysis Of Variance (ANOVA) tests whether more then two population means are equal or not, and therefore generalizes the t-test beyond two means.

H0: all means are equal $m_0 = m_1 = m_2 = \ldots = m_n$

H1: not all means are equal (at least one pair of means is different)

ANOVA

Dependent variable: Travel services (% of service exports)

Grouping variable: 1 – high income countries, 2 – middle income countries, 3 – low income countries

H1: the mean values of the *percent of travel services our of service exports* for <u>groups of countries based on income</u> are different

H0: mean values are not different

p-value is 0.0001 < 0.05 → the probability of making an error while rejecting H_0 and accepting H_1 is very low, we will accept H_1 and conclude that there is a statistically significant difference in mean values of the dependent variable for the three groups.

ANOVA

The significance of ANOVA is tested based on the value of F statistic.

$$SST = SSW + SSB$$

Number of groups

$$F = \frac{MSB}{MSW} = \frac{SSB/(c-1)}{SSW/(n-c)}$$

Number of cases

ANOVA: SST

Group 1: 1, 3, 5

Group 2: 5, 7, 9

Group 3: 4, 5, 6

Sample mean:

(1+3+5+5+7+9+4+5+6)/9 = 5

$$SST = \sum(X-\overline{X})^2$$

Total sum of squares (SST): 42

$(1-5)^2 + (3-5)^2 + (5-5)^2 + (5-5)^2 + (7-5)^2 + (9-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 =$
$16 + 4 + 0 + 0 + 4 + 16 + 1 + 0 + 1 = 42$

ANOVA: SSW

Group 1: 1, 3, 5 mean 1 = (1+3+5)/3 = 3
Group 2: 5, 7, 9 mean 2 = (5+7+9)/3 = 7
Group 3: 4, 5, 6 mean 3 = (4+5+6)/3 = 5

Sum of squares within groups (SSW):

$SSW1 = (1-3)^2 + (3-3)^2 + (5-3)^2 = 4 + 0 + 4 = 8$
$SSW2 = (5-7)^2 + (7-7)^2 + (9-7)^2 = 4 + 0 + 4 = 8$
$SSW3 = (4-5)^2 + (5-5)^2 + (6-5)^2 = 1 + 0 + 1 = 2$

SSW = SSW1+SSW2+SSW3 = 8 + 8 + 2 = 18

ANOVA: SSB

Group 1: 1, 3, 5
Group 2: 5, 7, 9
Group 3: 4, 5, 6

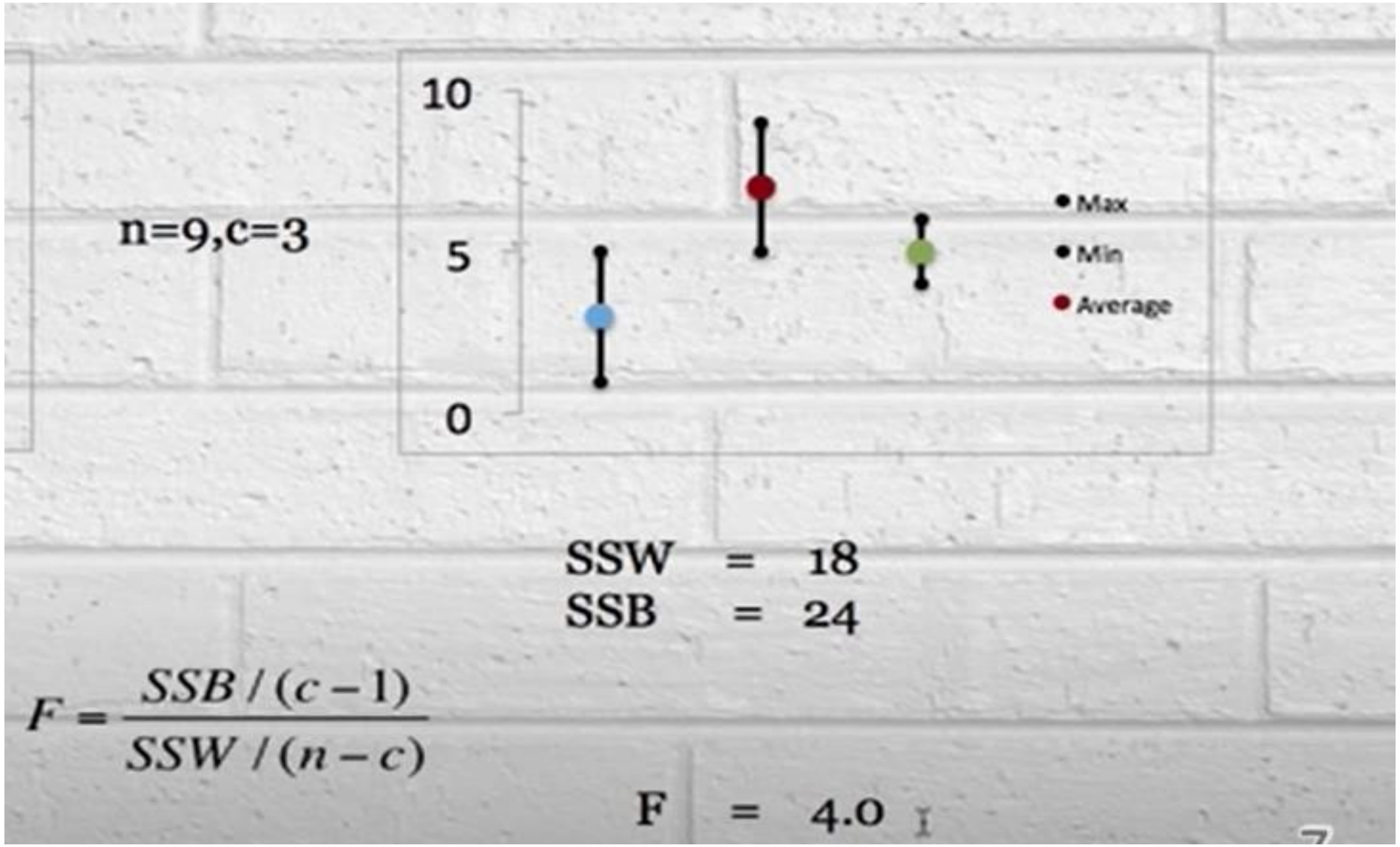
Sample mean:
(1+3+5+5+7+9+4+5+6)/9 = 5

Mean 1 = (1+3+5)/3 = 3
Mean 2 = (5+7+9)/3 = 7
Mean 3 = (4+5+6)/3 = 5

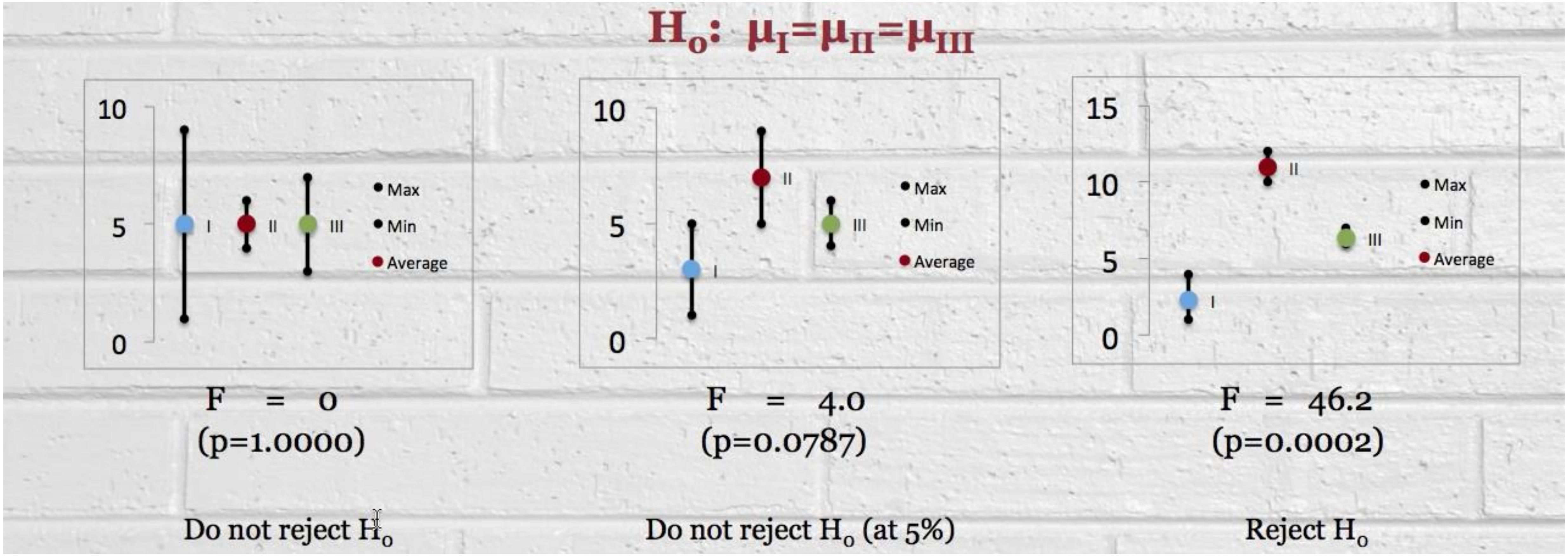
SSB1 = $3*(3-5)^2$ = 12
SSB2 = $3*(7-5)^2$ = 12
SSB3 = $3*(5-5)^2$ = 0

SSB = SSB1+SSB2+SSB3 = 12 + 12 + 0 = 24

ANOVA



ANOVA



Source: https://www.youtube.com/watch?v=9cnSWads6oo

Post-hoc test

If ANOVA confirms that the means are not equal, we can do pairwise post-hoc testing to determine whether there is a difference between the means of all possible pairs (all pairwise comparisons).

Mann-Whitney U Test

Is a non-parametric test, analogue of Independent-samples t-test. Can be performed for ordinal or non-normally distributed interval or ratio scale data.

H_1: The two populations are not equal

Mann-Whitney U Test

		Total Sample (Ordered Smallest to Largest)		Ranks	
Placebo	New Drug	Placebo	New Drug	Placebo	New Drug
7	3		1		1
5	6		2		2
6	4		3		3
4	2	4	4	4.5	4.5
12	1	5		6	
		6	6	7.5	7.5
		7		9	
		12		10	

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/bs704_nonparametric4.html

Wilcoxon signed-rank test

Is a non-parametric test, analogue of Paired-samples t-test.

H_1: ranks of the group before and after treatment are not equal https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric6.html#headingtaglink_1

Wilcoxon signed-rank test

Calculator: https://www.socscistatistics.com/tests/signedranks/default2.aspx

Patient	Before Treatment	After 1 Week of Treatment	Difference (Before-After)
1	85	75	10
2	70	50	20
3	40	50	-10
4	65	40	25
5	80	20	60
6	75	65	10
7	55	40	15
8	20	25	-5

Ordered Absolute Values of Differences	Ranks	Signed Ranks
-5	1	-1
10	3	3
-10	3	-3
10	3	3
15	5	5
20	6	6
25	7	7
60	8	8

W+ = 32 and W- = 4. The sums of positive and negative ranks are calculated and the smallest sum is selected. As a result, W=4.

Kruskal-Wallis test

Is a non-parametric test, analogue of ANOVA.

Null hypothesis: the samples (groups) are from identical populations.

Alternative hypothesis: at least one of the samples (groups) comes from a different population than the others.

Kruskal-Wallis test

First, we rank all the values regardless of the group to which they belong.

Group 1	Group 2	Group 3
27	20	34
2	8	31
4	14	3
18	36	23
7	21	30
9	22	6

Original Score	Rank
2	1
3	2
4	3
6	4
7	5
8	6
9	7
14	8
18	9
20	10
21	11
22	12
23	13
27	14
30	15
31	16
34	17
36	18

Kruskal-Wallis test

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric7.html#headingtaglink_3

Then we replace values with ranks and in each group the sum of the ranks (T) is calculated

	Group 1	Group 2	Group 3
	14	10	17
	1	6	16
	3	8	2
	9	18	13
	5	11	15
	7	12	4
T	39	65	67
n	6	6	6

$$H = \frac{12}{N(N+1)}\left(\sum \frac{T_i^2}{n}\right) - 3(N+1)$$

$$H = \frac{12}{18(18+1)}\left(\frac{39^2}{6} + \frac{65^2}{6} + \frac{67^2}{6}\right) - 3(18+1) = 2.854$$

Source: https://www.youtube.com/watch?v=BkyGuNuaZYw

Comparing tests

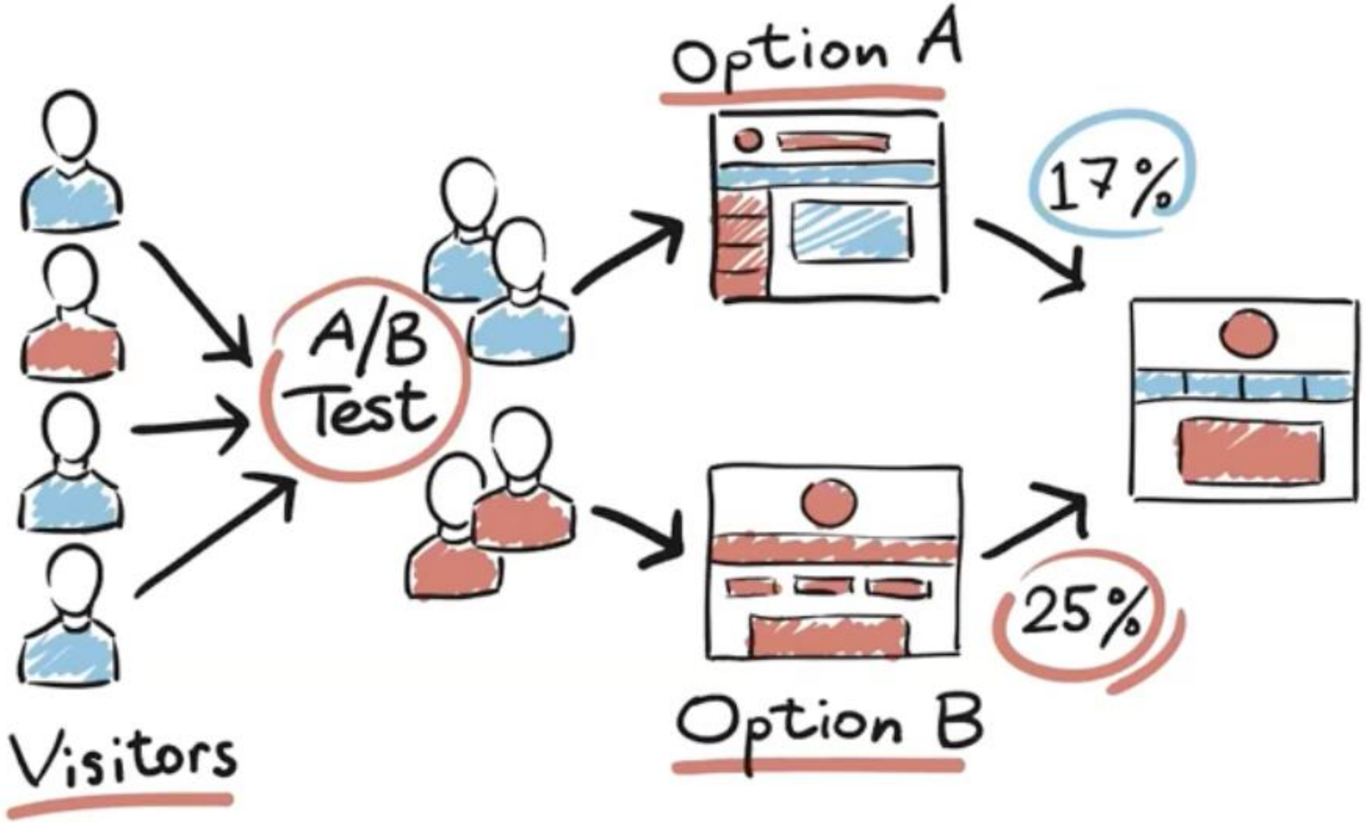
Type of samples	Parametric tests	Non-parametric tests
2 independent samples	t-test for two independent samples	Mann-Whitney U test
2 related samples	t-test for two paired samples	Wilcoxon signed-rank test
> 2 independent samples	ANalysis Of VAriance (ANOVA)	Kruskal-Wallis test

A/B testing

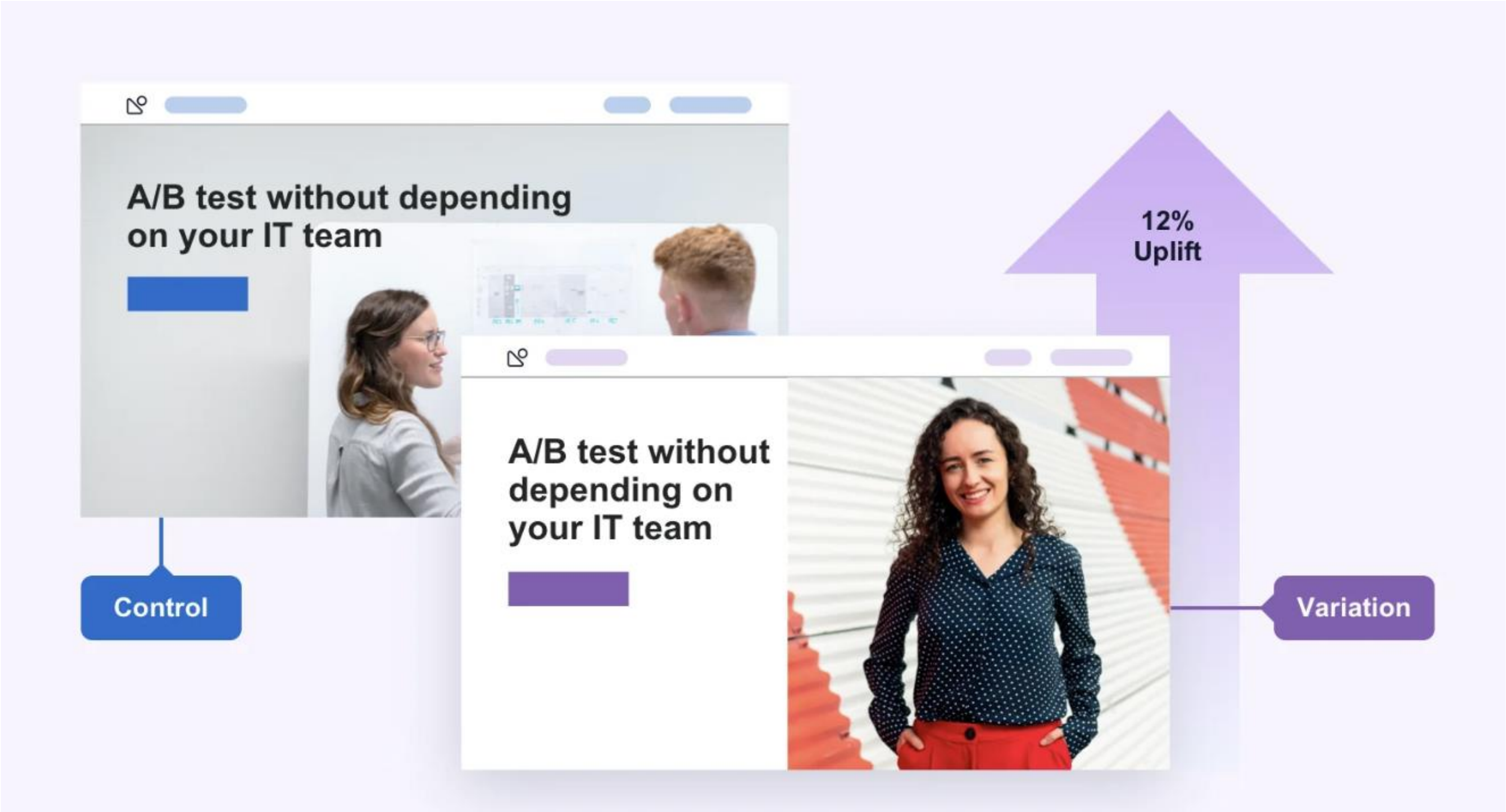
Compares two versions of an app or webpage to identify the better performer. It compares options to learn what customers prefer. You can test website/app layouts, email subject lines, product designs, colors, etc. It’s a randomized experimentation process where in two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to define which version leaves the maximum impact and drives business metrics.

In A/B testing, A refers to ‘control’ or the original testing variable. Whereas B refers to ‘variation’ or a new version of the original testing variable. The version that moves business metric(s) in the positive direction is known as the ‘winner.’

A/B testing



A/B testing


A/B test without depending on your IT team
A/B test without depending on your IT team
12% Uplift
Control
Variation

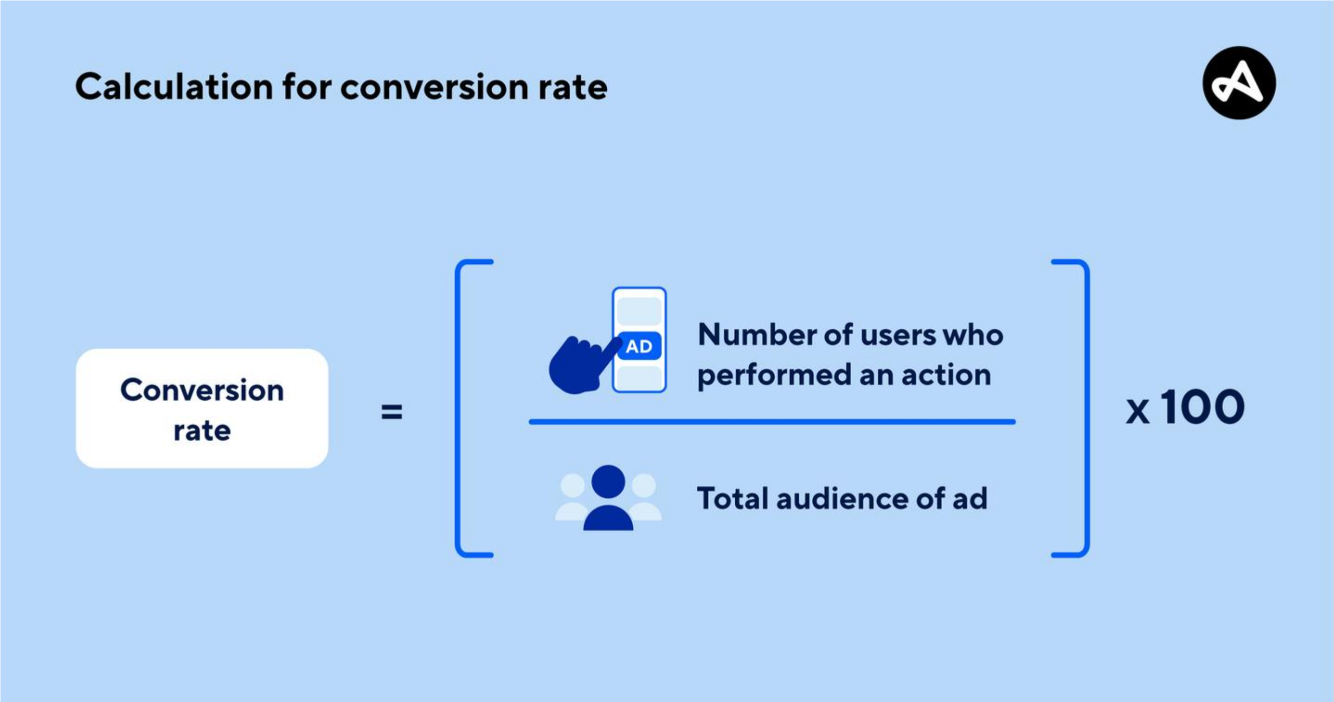
Conversion rate

Conversion rate records the percentage of users who have completed a desired action. Conversion rates are calculated by taking the total number of users who ‘convert’ (for example, by clicking on an advertisement), dividing it by the overall size of the audience and converting that figure into a percentage.

Conversion rates are an effective way of comparing and contrasting the performance of multiple advertising channels. Conversion rates are particularly important when running mobile user acquisition because they can measure the success of each campaign. A good conversion rate means a strong return on investment.

Conversion rate

For example, an advertiser runs a campaign with an audience of 20,000 people. Out of that group, 800 people clicked on the ad (ie. converted). To calculate this example, divide 800 by 20,000 to get 0.04, or a 4% conversion rate.



Bounch rate

In contrast to conversion rate, bounce rate is the percentage of visitors who enter a landing page but leave without taking additional actions — they bounce right out of there. These exits are known as single-page sessions.

$$\text{Bounce Rate} = \left(\frac{\text{One-page Visits}}{\text{Total Visits}} \right) \times 100$$

Click-through rate

Measures the percentage of clicks on a specific link, as compared to the number of times the link was shown (AKA the number of impressions).

$$\text{Click-Through Rate} = \left(\frac{\text{Clicks}}{\text{Impressions}} \right) \times 100$$

A/A testing

A/A tests enable to test two identical versions of an element. The traffic to the website is divided into two, with each group exposed to the same variation. A/A testing can help to define whether the conversion rates in each group are similar and confirm that the solution is working correctly. It is also a good way to identify bugs and outliers that impact results thereby raising the level of trust in the experiments.

A/A testing

A/A Testing Example

A/B testing algorithm

1. Understanding business problem & data
2. Detect and resolve problems in the data (Missing Value, Outliers, Unexpected Value)
3. Look at summary stats and plots
4. Apply hypothesis testing and check assumptions

Check Normality & Homogeneity

- Apply tests (Shapiro, Levene Test, T-Test, Mann Whitney U Test, etc.)

5. Evaluate the results
6. Make inferences
7. Recommend business decision to your customer/director/CEO etc.

Possible metrics

1. Revenue;
2. Conversion rate, bounch rate, click-through rate;
3. Profit margins (percentage of each dollar of revenue that the business keeps as profit after covering all costs);
4. Customer lifetime value (prediction of the total revenue a customer will generate throughout their relationship with the business);
5. User retention rate (percentage of users who keep engaging with the product, service, or content over a given period);
6. Customer satisfaction score;
7. Cart abandonment rate (percentage of shoppers who add items to their online shopping cart but leave without completing the purchase);
8. Click-through rate (how often people click on the links or ads after seeing them;
9. Average session duration;
10. Scroll depth.

Using proportions test

To define if the conversion rate for page A is significantly higher than page B, we do a difference of proportions test.

The test statistic for testing the difference in two population proportions, that is, for testing the null hypothesis $H_0 : p_1 - p_2 = 0$ is:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where:

$$\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}$$

the proportion of "successes" in the two samples combined.

Non- Smokers	Smokers
$n_1 = 605$	$n_2 = 195$
$y_1 = 351$ said "yes"	$y_2 = 41$ said "yes"
$\hat{p}_1 = \frac{351}{605} = 0.58$	$\hat{p}_2 = \frac{41}{195} = 0.21$

A/B testing examples from kaggle.com

- https://www.kaggle.com/code/ekrembayar/a-b-testing-step-by-step-hypothesis-testing
- https://www.kaggle.com/code/bahaulug/a-b-testing
- https://www.kaggle.com/code/babyoda/a-b-testing-in-practice

Thank you for your attention!