A/B testing

Course roadmap

1. Project valuation: valuation metrics, planning and rules
2. Model quality and decision making. Benefit curve
3. Estimating model risk discounts
4. A/B testing and financial result verification
   - What is A/B testing?
   - Five principles of A/B testing
   - Preparing an A/B test
   - Evaluation of A/B testing results
5. Unobservable model errors, metalearning

What is A/B testing

What is A/B testing

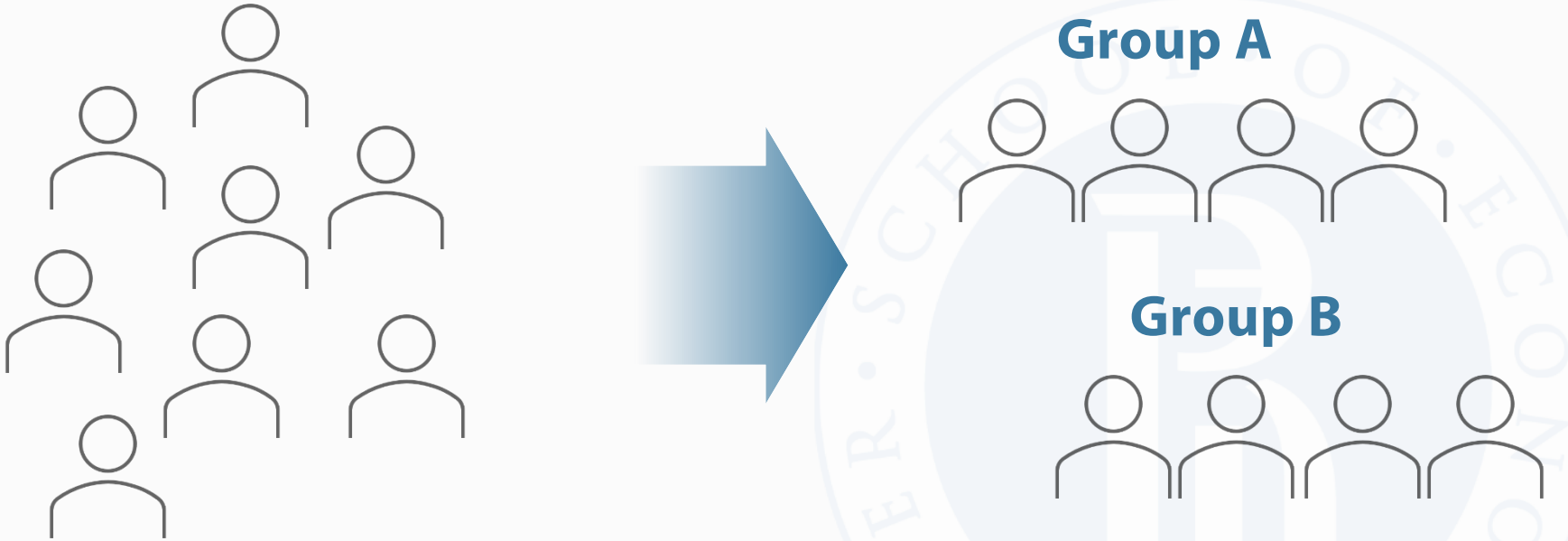
- A/B testing is a method of research to test hypothesis of our influence on a process / products. It allows to mitigate the risk of wrong assumptions that can worsen business process performance

What is A/B testing

- A/B testing is a method of research to test hypothesis of our influence on a process / products. It allows to mitigate the risk of wrong assumptions that can worsen business process performance
- This methodology allows one to clear the effect of the project / AI solution from macro changes in economy, supply and demand, seasonality and other exogenous (environmental) factors

What is A/B testing

- In order to conduct A/B testing we need to form two (or more) randomly split samples so that before our experiment both samples (A and B) have the same features and behavior



What is A/B testing

- To verify whether randomization is done accurately, preliminary we need to list the features of sample units that determine the result we are going to have an impact on

Group A

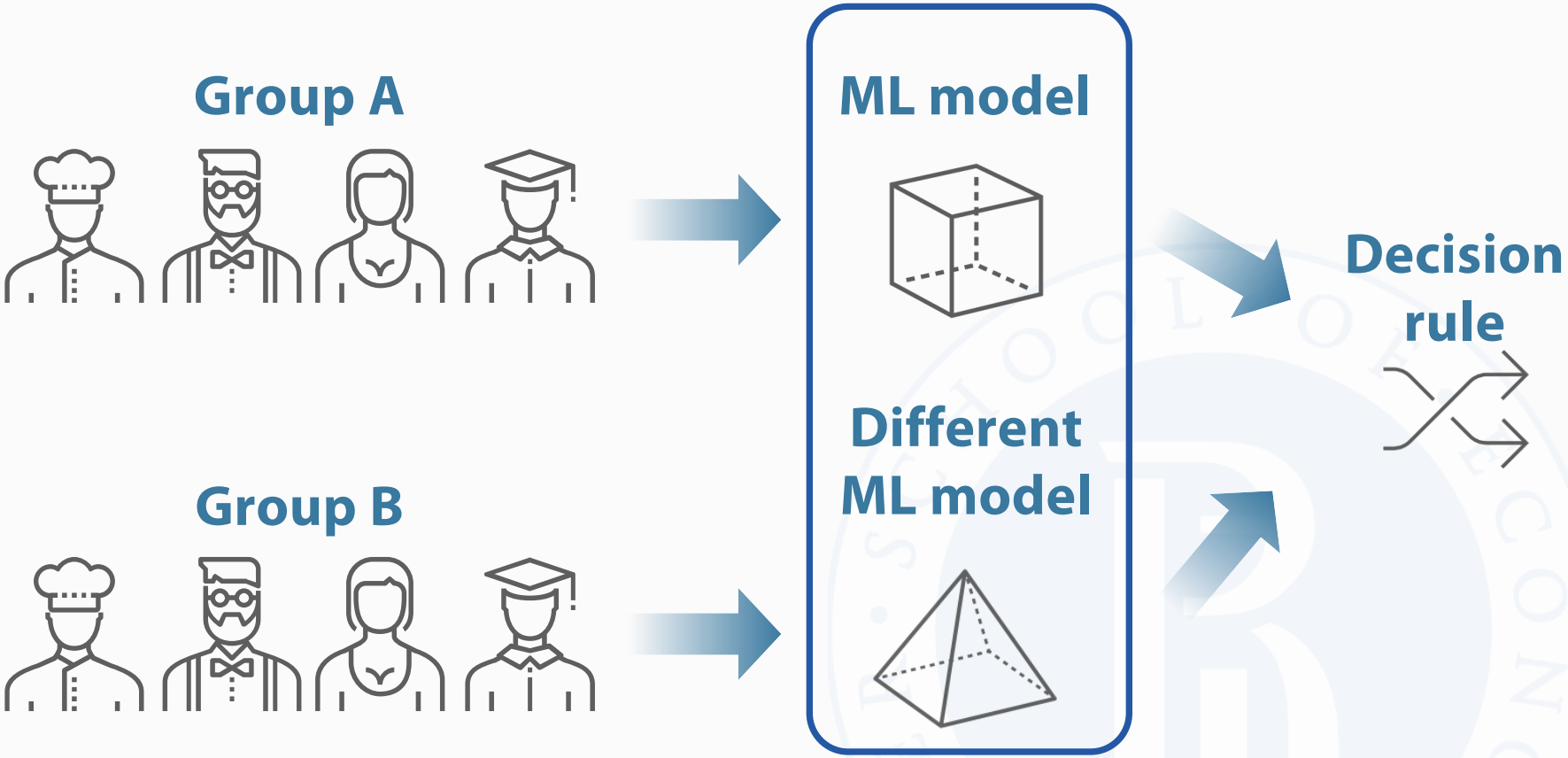
Group B

Features: sex, occupation

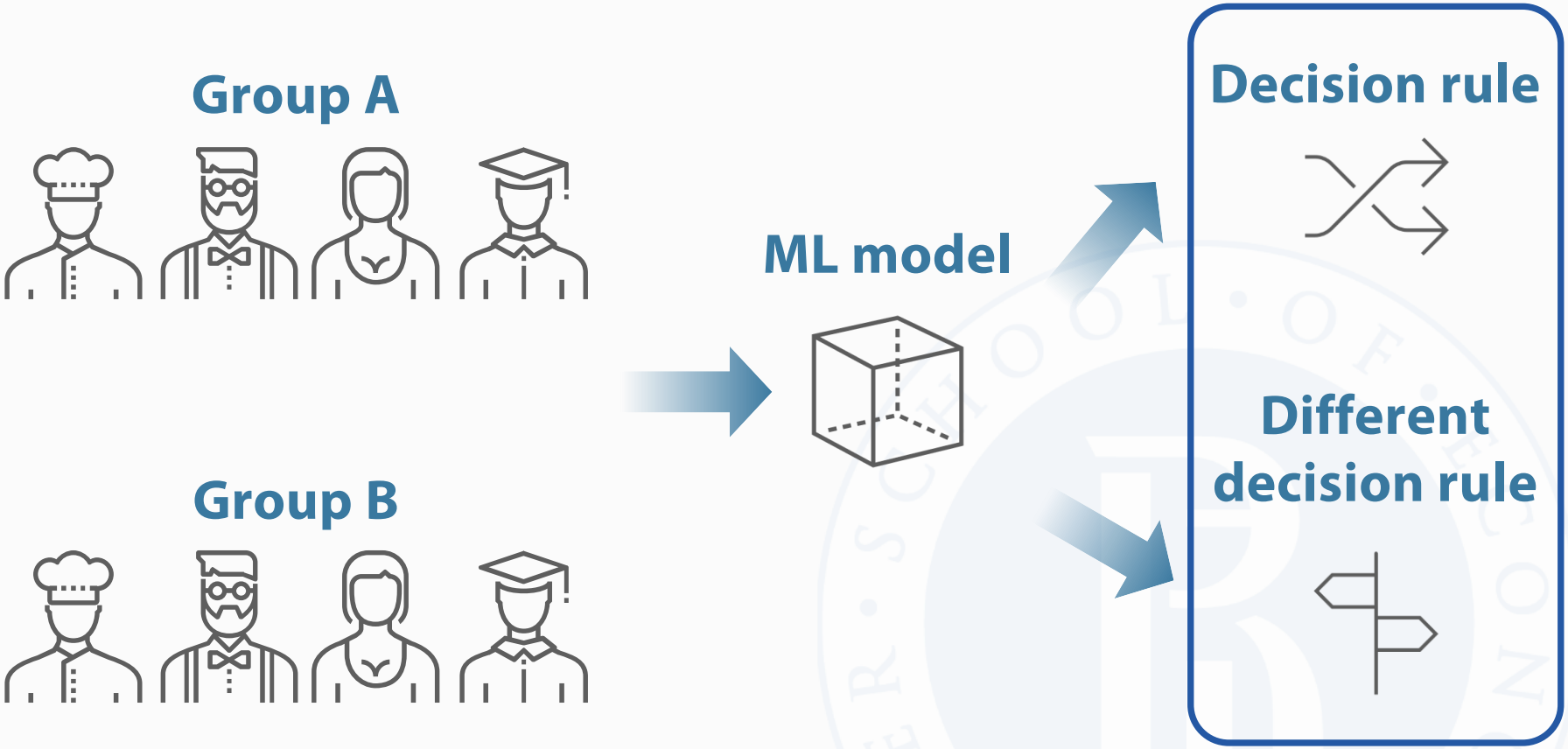
25% are women
25% are Master students

Both in groups A and B

A/B testing. Different ML model



A/B testing. Different ML model



Wrap-up

1. A/B testing is a concept of controlled experiment that allows one to identify the effect of proposed changes in a business process

Wrap-up

1. A/B testing is a concept of controlled experiment that allows one to identify the effect of proposed changes in a business process
2. A/B testing is working subject to “other things equal” principle

Wrap-up

1. A/B testing is a concept of controlled experiment that allows one to identify the effect of proposed changes in a business process
2. A/B testing is working subject to “other things equal” principle
3. When applying to ML models financial impact estimation, one can A/B test either new ML model, or a new set of decision rules, or both

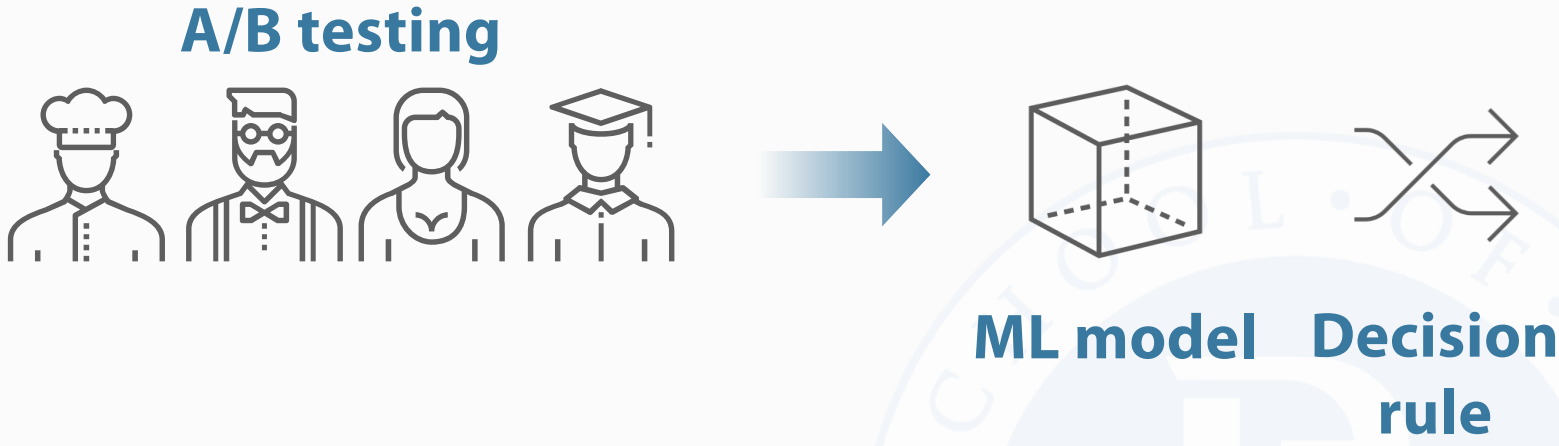
5 principles of A/B testing

A/B testing 5 principles

1. The process that is tested via A/B testing must replicate the process after scaling the solution

A/B testing 5 principles

1. The process that is tested via A/B testing must replicate the process after scaling the solution



A/B testing 5 principles

1. The process that is tested via A/B testing must replicate the process after scaling the solution

A/B testing

ML model **Decision rule**

Deployment

Different ML model **Different Decision rule**

A/B testing 5 principles

1. The process that is tested via A/B testing must replicate the process after scaling the solution

A/B testing

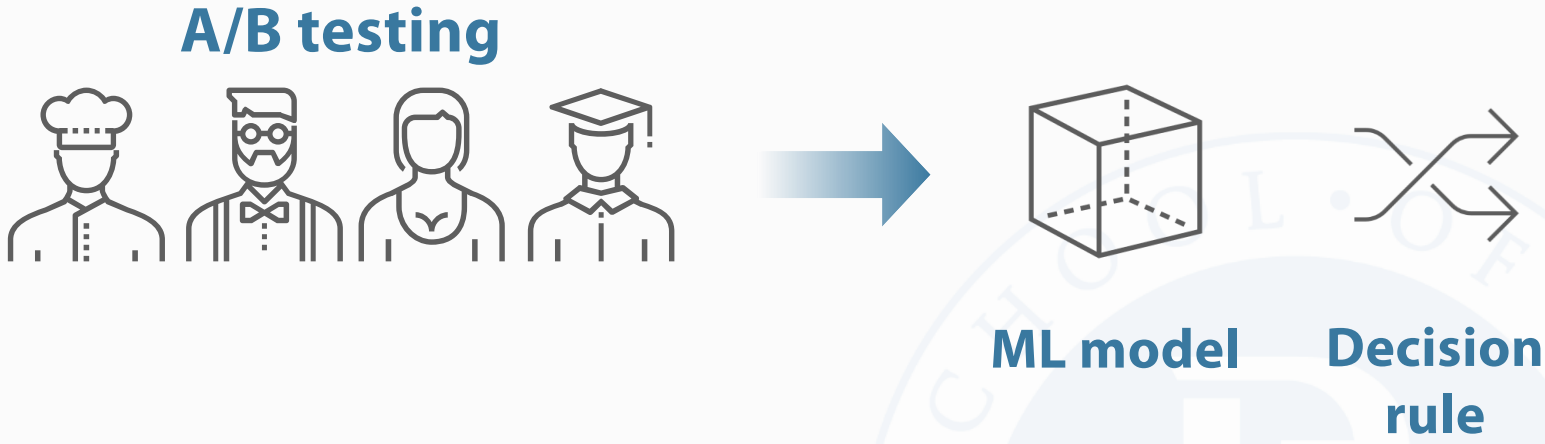
ML model **Decision rule**

Deployment

Precisely the same ML model **Same Decision rule**

A/B testing 5 principles

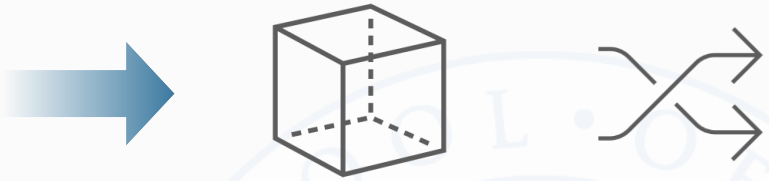
2. The structure and features of units within A/B test must be representative of structure and features after scaling our solution



A/B testing 5 principles

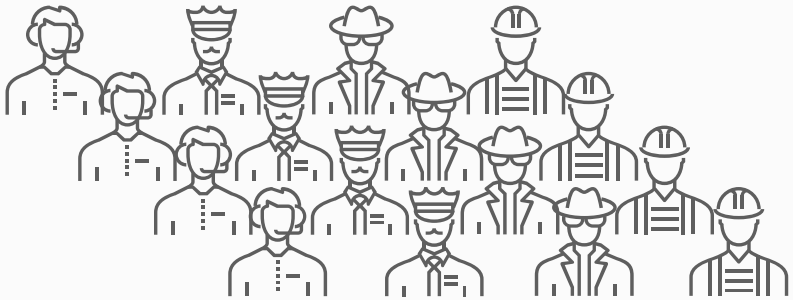
2. The structure and features of units within A/B test must be representative of structure and features after scaling our solution

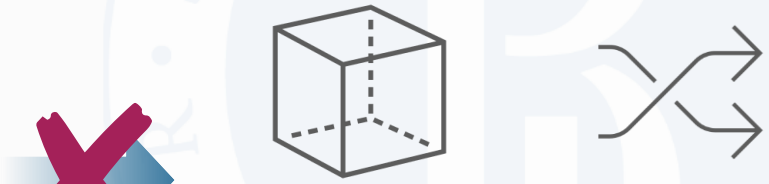
A/B testing



ML model **Decision rule**

Deployment





Same ML model **Same Decision rule**

Different client profile

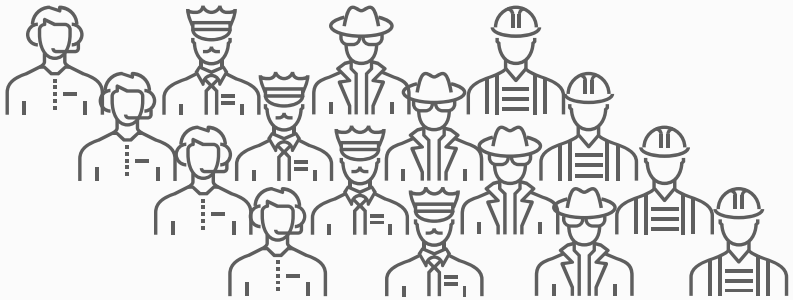
A/B testing 5 principles

2. The structure and features of units within A/B test must be representative of structure and features after scaling our solution

A/B testing

ML model **Decision rule**

Deployment



Same client profile

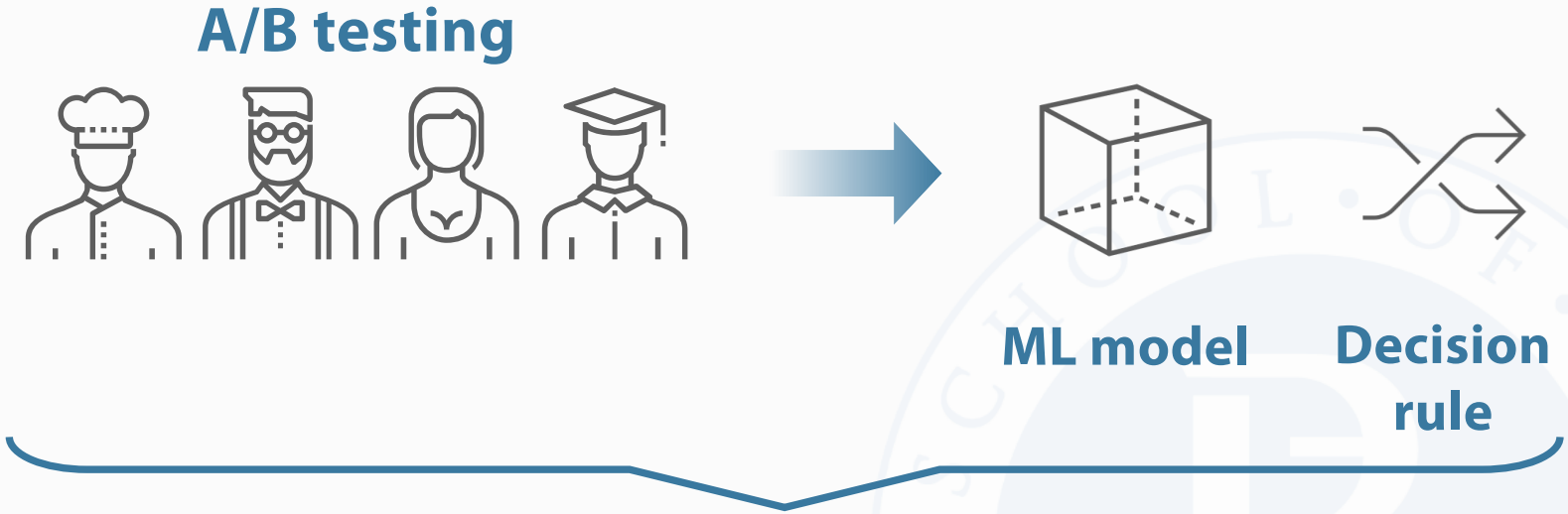
Same ML model **Same Decision rule**

A/B testing 5 principles

3. The criterion for decision making and all terms of piloting must be determined before starting an A/B test

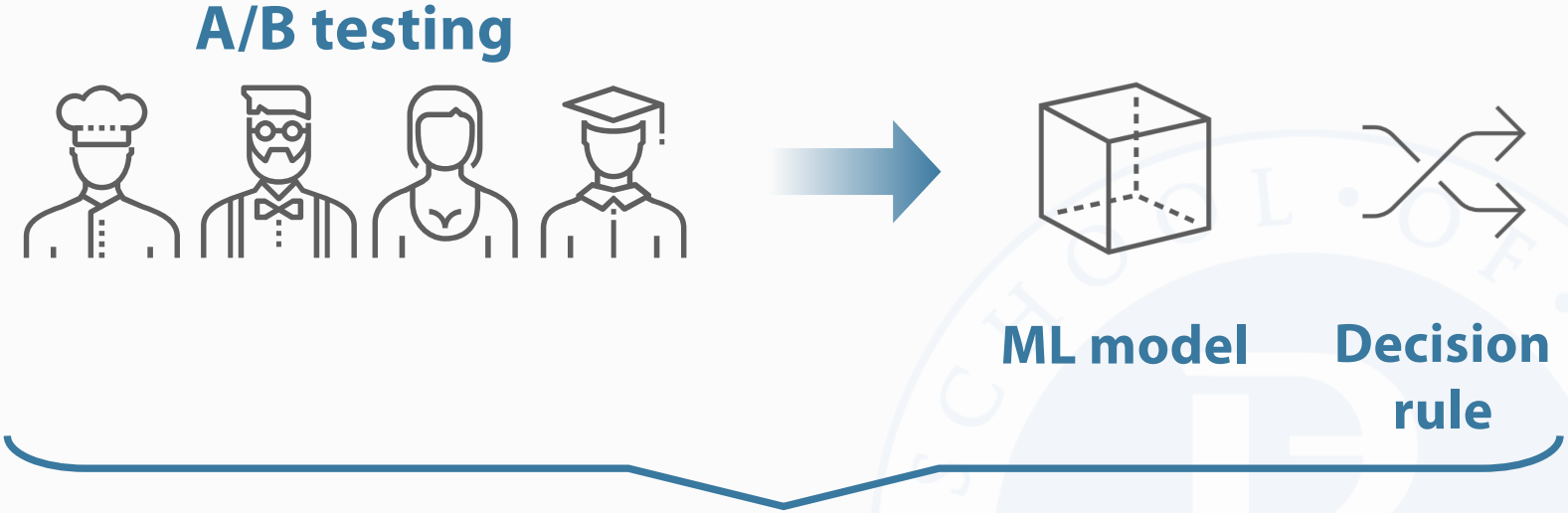
A/B testing 5 principles

4. The decision to scale the solution must be based only on the terms determined at the beginning of A/B test and its results



A/B testing 5 principles

4. The decision to scale the solution must be based only on the terms determined at the beginning of A/B test and its results



Measurable results of A/B test

(determined terms of success)

A/B testing 5 principles

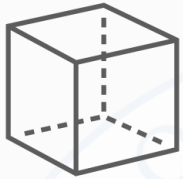
5. The difference between A and B groups in terms of examined target must be statistically significant

A/B testing 5 principles

5. The difference between A and B groups in terms of examined target must be statistically significant

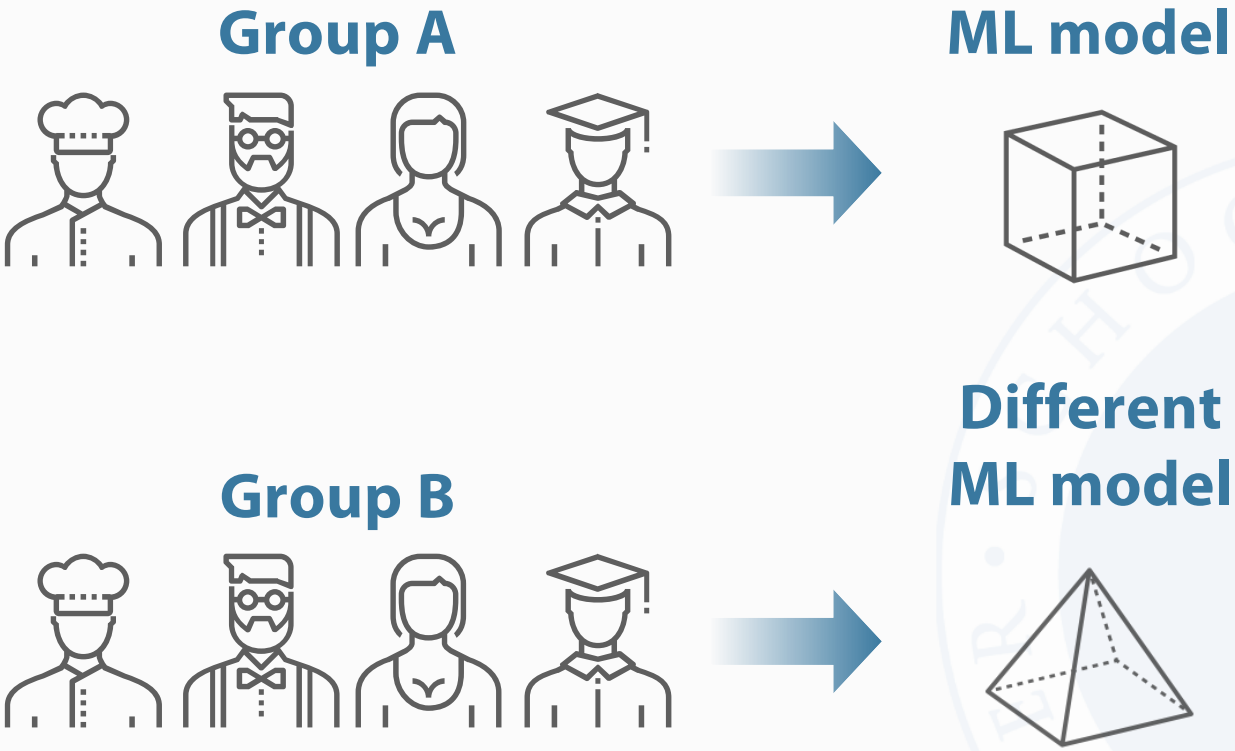
Group A

ML model



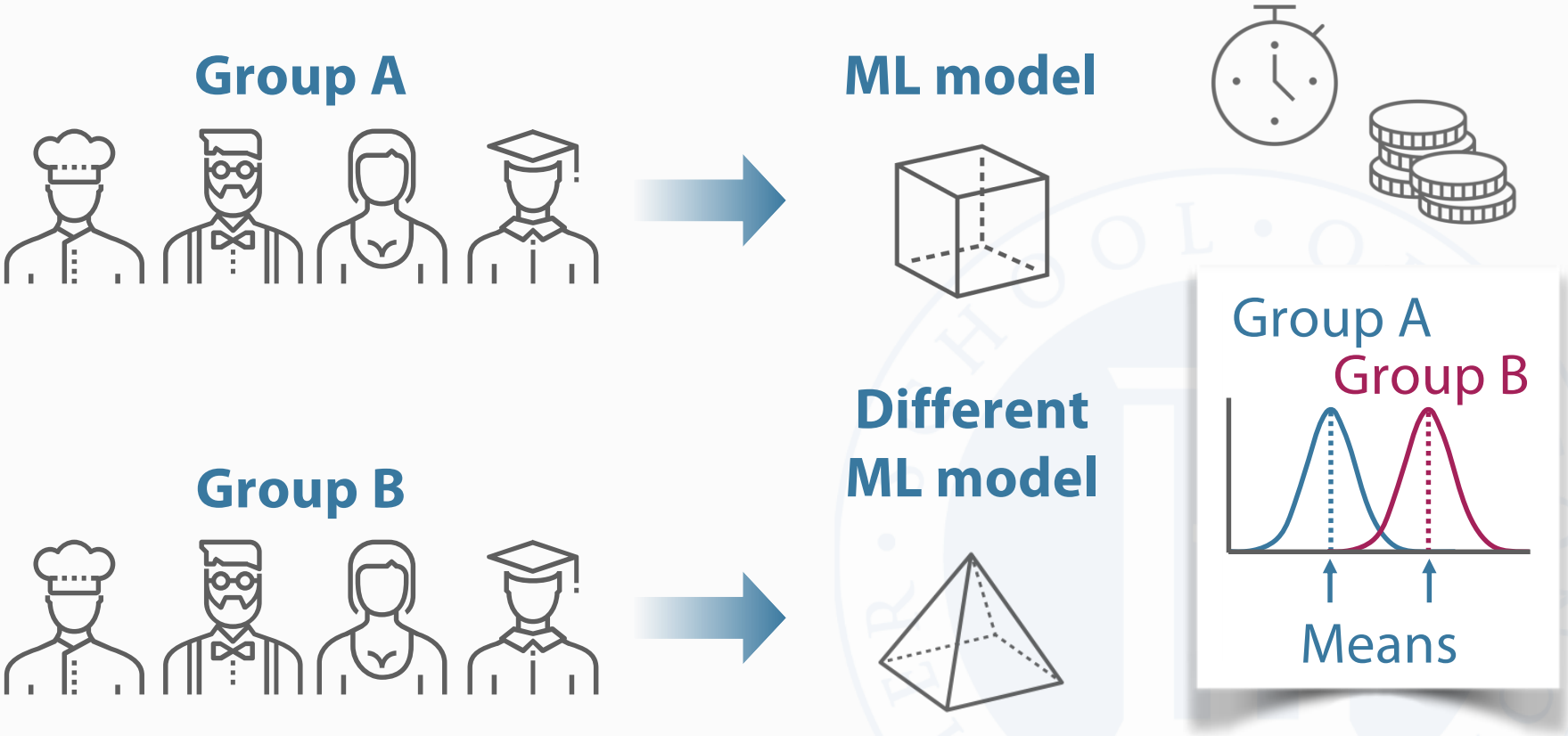
A/B testing 5 principles

5. The difference between A and B groups in terms of examined target must be statistically significant



A/B testing 5 principles

5. The difference between A and B groups in terms of examined target must be statistically significant



Wrap-up

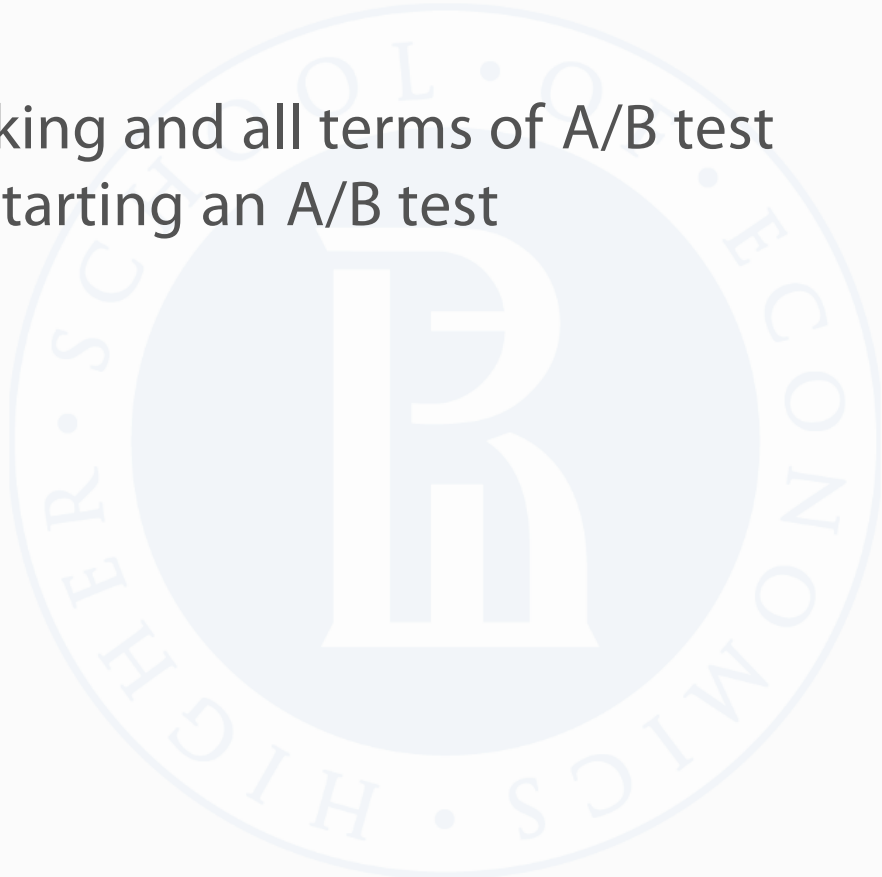
1. The process that is tested via A/B testing must replicate the process after scaling the solution

Wrap-up

1. The process that is tested via A/B testing must replicate the process after scaling the solution
2. The structure and features of units within an A/B test must be representative of structure and features after scaling our solution

Wrap-up

1. The process that is tested via A/B testing must replicate the process after scaling the solution
2. The structure and features of units within an A/B test must be representative of structure and features after scaling our solution
3. The criterion for decision making and all terms of A/B test must be determined before starting an A/B test



Wrap-up

1. The process that is tested via A/B testing must replicate the process after scaling the solution
2. The structure and features of units within an A/B test must be representative of structure and features after scaling our solution
3. The criterion for decision making and all terms of A/B test must be determined before starting an A/B test
4. The decision to scale the solution must be based only on the terms determined at the beginning of A/B test and its results

Wrap-up

1. The process that is tested via A/B testing must replicate the process after scaling the solution
2. The structure and features of units within an A/B test must be representative of structure and features after scaling our solution
3. The criterion for decision making and all terms of A/B test must be determined before starting an A/B test
4. The decision to scale the solution must be based only on the terms determined at the beginning of A/B test and its results
5. The difference between A and B groups in terms of examined target must be statistically significant

Preparing an A/B test

A/B testing preparation. Features checking

Example:

We are going to improve the response rate in communication campaigns / direct sales for legal entities.

The following steps are mandatory before A/B testing:

A/B testing preparation. Features checking

Example:

We are going to improve the response rate in communication campaigns / direct sales for legal entities.

The following steps are mandatory before A/B testing:

Step 1. Determine the object of treatment (what we tend to have an impact on) — it might be sales managers / clients / outlets

A/B testing preparation. Features checking

Example:

We are going to improve the response rate in communication campaigns / direct sales for legal entities.

The following steps are mandatory before A/B testing:

Step 1. Determine the object of treatment (what we tend to have an impact on) — it might be sales managers / clients / outlets

Step 2. Determine the indicator of improvement (e.g. average NPV of sales, response rate, time-to-market)

A/B testing preparation. Features checking

Example:

We are going to improve the response rate in communication campaigns / direct sales for legal entities.

The following steps are mandatory before A/B testing:

Step 3. Collect and analyze current sales statistics (trends, sales per manager / outlet, regional specificity)

A/B testing preparation. Features checking

Example:

We are going to improve the response rate in communication campaigns / direct sales for legal entities.

The following steps are mandatory before A/B testing:

Step 3. Collect and analyze current sales statistics (trends, sales per manager / outlet, regional specificity)

Step 4. Explore clients' data — determine customer groups that differ significantly in terms of demand (sales structure) and sales parameters in order to correctly build a control group

A/B testing preparation. Features checking

Example:

We are going to improve the response rate in communication campaigns / direct sales for legal entities.

The differences in demand and sales parameters (average bill, interest / commission rate, other terms) may be explained by

- company size,
- business field,
- being part of a corporate group,
- regional specificity.

A/B testing preparation. Randomization

Randomization implies splitting the objects of experiment into 2 groups that have the same behavior and distribution of sales and client features (e.g. age, sex, average income) before our influence.

Group A (control group) — sample with units that we do not influence

Group B (treatment group) — sample with units we will treat

A/B testing preparation. Randomization

In some cases we cannot achieve true randomization

A/B testing preparation. Randomization

In some cases we cannot achieve true randomization

For instance, legislation does not allow one to run different price policies (pricing discrimination) in one region and we can only run an A/B test so that samples A and B are 2 different regions with regional specificity.

A/B testing preparation. Randomization

In the cases when we cannot create truly randomized samples the following methods are applicable to solve it:

- Contrasting subsamples that are comparable in A and B (and scaling effect afterwards if structure of subsamples in A and B is constant and subsamples have the same distribution of features)

A/B testing preparation. Randomization

In the cases when we cannot create truly randomized samples the following methods are applicable to solve it:

- Contrasting subsamples that are comparable in A and B (and scaling effect afterwards if structure of subsamples in A and B is constant and subsamples have the same distribution of features)
- Creating synthetic control group — constructing synthetic control units as a weighted combination of comparison units that approximates the features of the sample that is treated (method is described in Abadie and Gardeazabal, 2003 and Abadie, Diamond, and Hainmueller (2010, 2011, 2014))

A/B testing preparation. Randomization

In the cases when we cannot create truly randomized samples the following methods are applicable to solve it:

- Propensity score matching is a method used when there was no controlled experiment, or its conditions were violated

A/B testing preparation. Randomization

In the cases when we cannot create truly randomized samples the following methods are applicable to solve it:

- Propensity score matching is a method used when there was no controlled experiment, or its conditions were violated

We have to additionally select clients from historical data to divide them into artificial A and B Groups to guarantee “other things equal” condition.

A/B testing preparation. Randomization

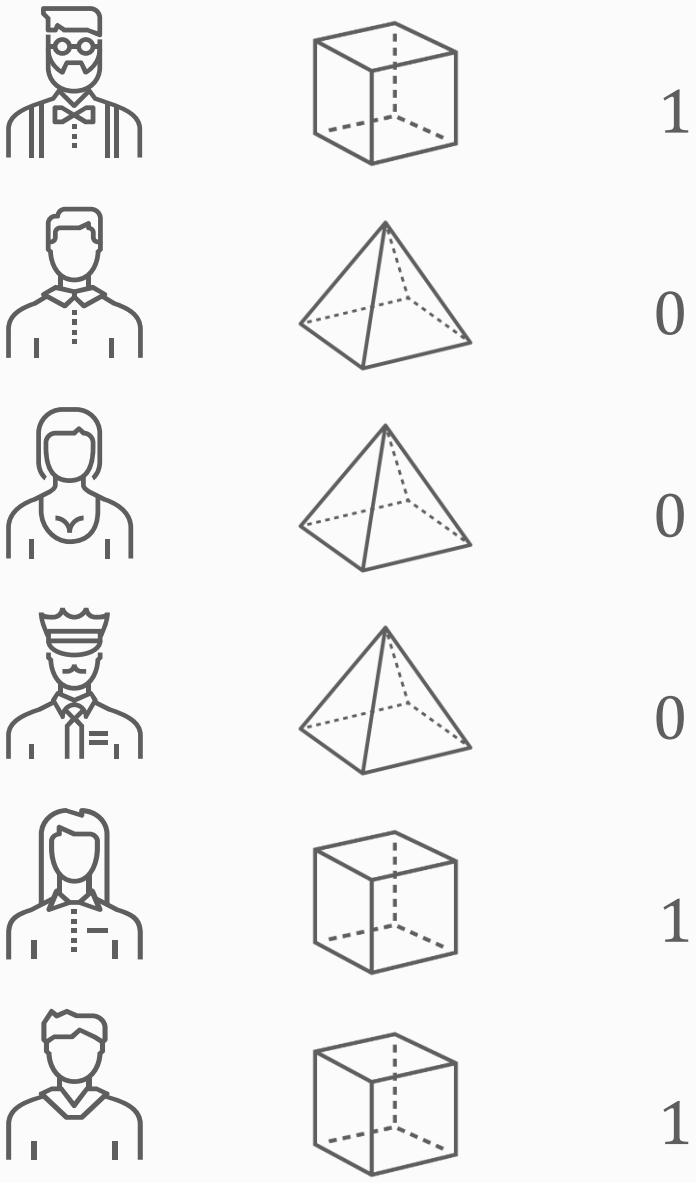
In the cases when we cannot create truly randomized samples the following methods are applicable to solve it:

- Propensity score matching is a method used when there was no controlled experiment, or its conditions were violated

We have to additionally select clients from historical data to divide them into artificial A and B Groups to guarantee "other things equal" condition.

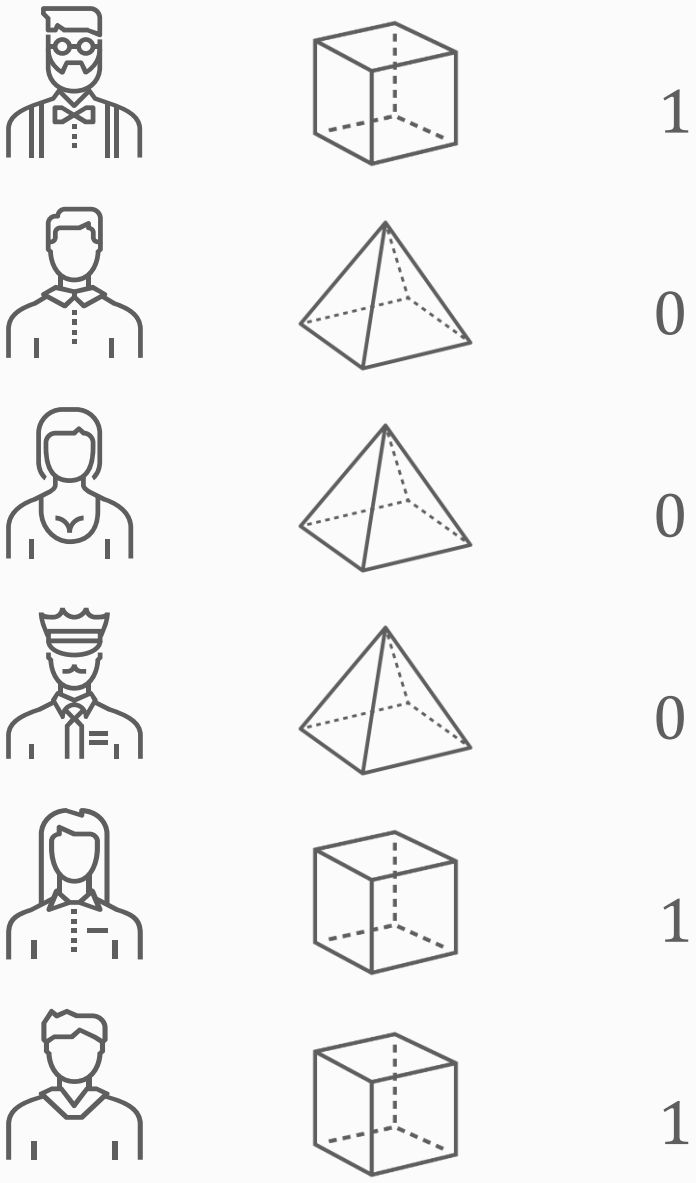
How does it work?

A/B testing preparation. Propensity score match



Suppose, there was no controlled experiment

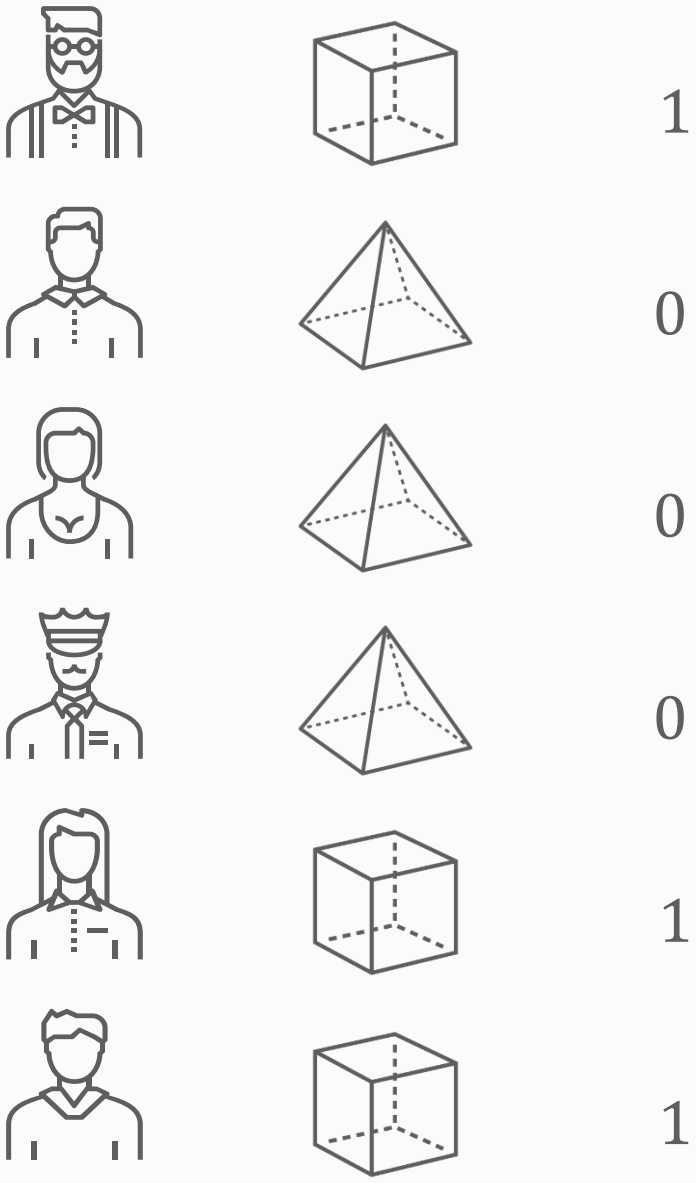
A/B testing preparation. Propensity score match



We can build an auxiliary classifier:

Client's $X \rightarrow Prob$ of being treated by cube or pyramid ML model

A/B testing preparation. Propensity score match

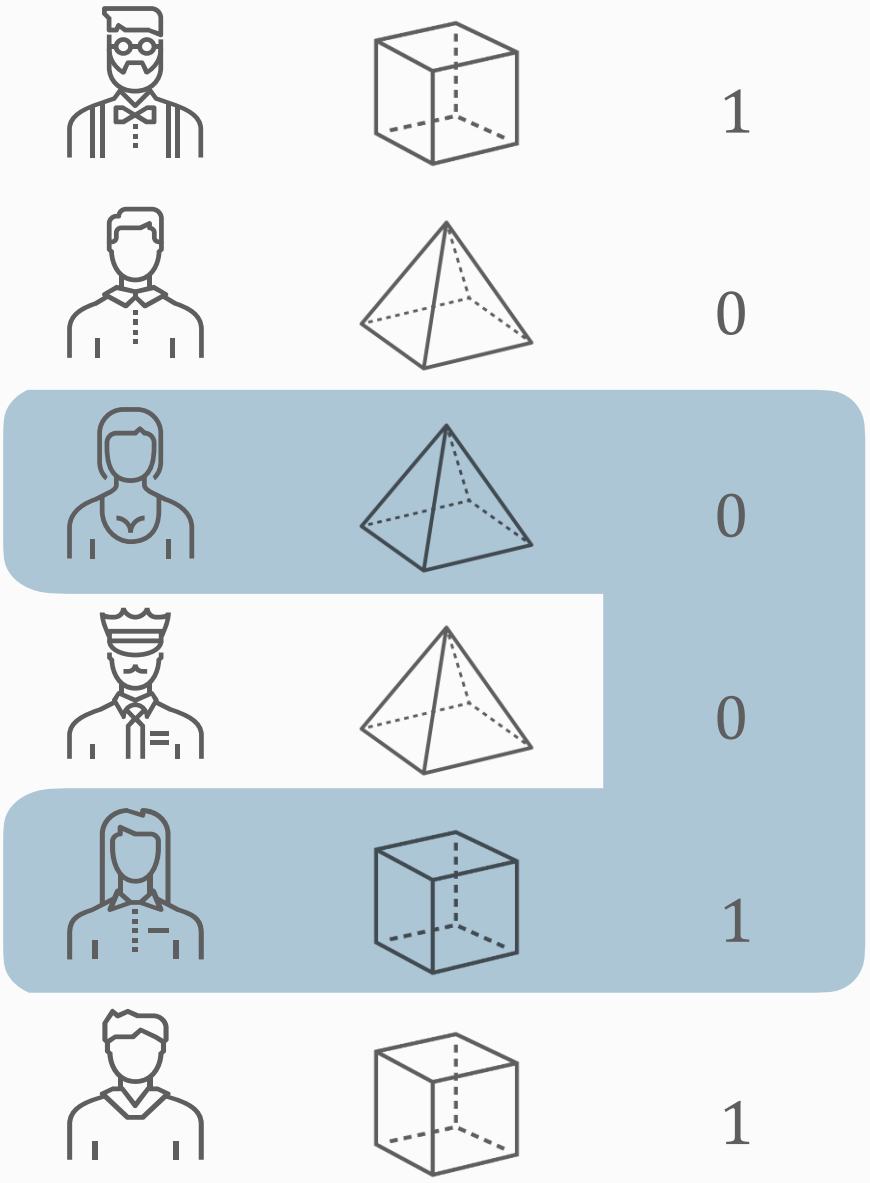


We can build an auxiliary classifier:

Client's $X \rightarrow Prob$ of being treated by cube or pyramid ML model

Clients with closest Probs (but different treatments) can be "matched" as pairs

A/B testing preparation. Propensity score match

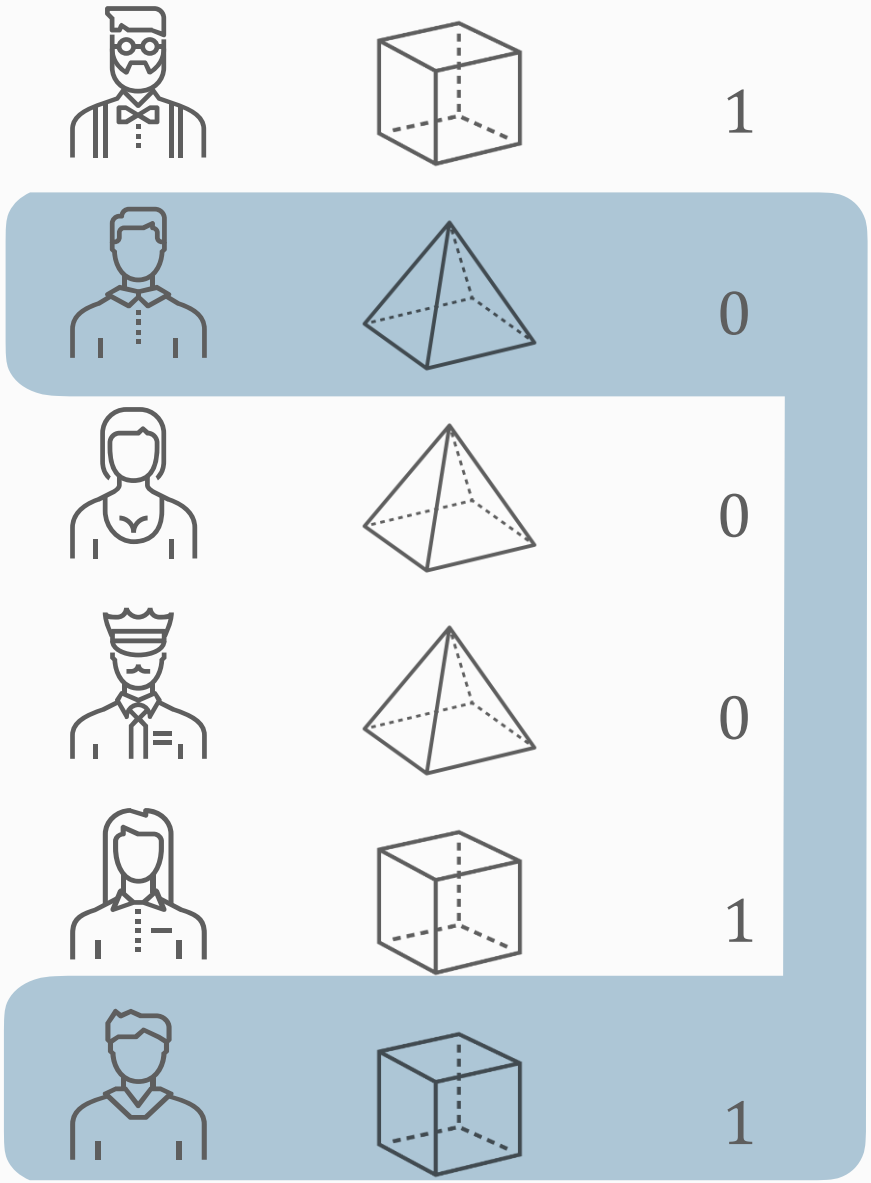


We can build an auxiliary classifier:

Client's $X \rightarrow Prob$ of being treated by cube or pyramid ML model

Clients with closest Probs (but different treatments) can be "matched" as pairs

A/B testing preparation. Propensity score match

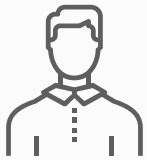
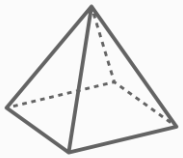


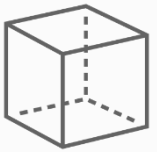
We can build an auxiliary classifier:

Client's $X \rightarrow Prob$ of being treated by cube or pyramid ML model

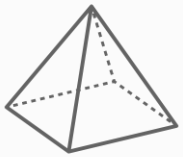
Clients with closest Probs (but different treatments) can be "matched" as pairs

A/B testing preparation. Propensity score match

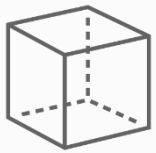



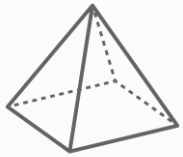


Pair 1

Pair 2

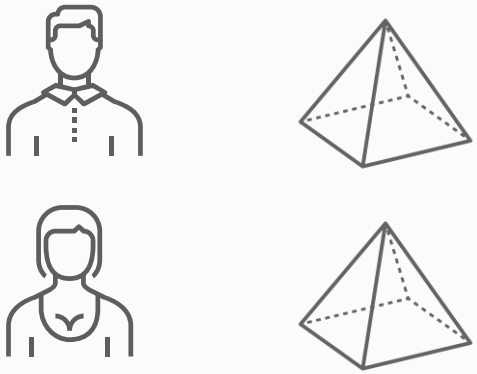





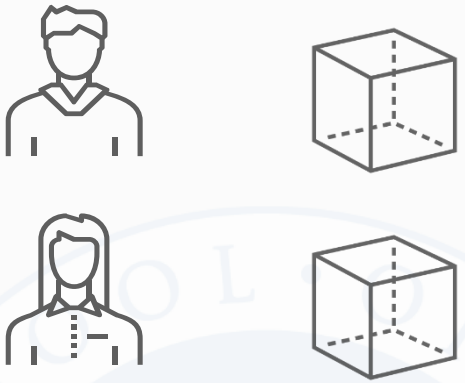
Unpaired (discarded)

A/B testing preparation. Propensity score match

“Group A”

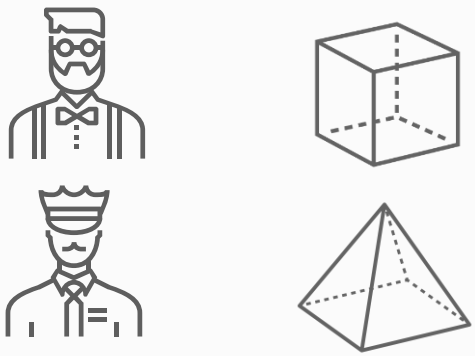


“Group B”



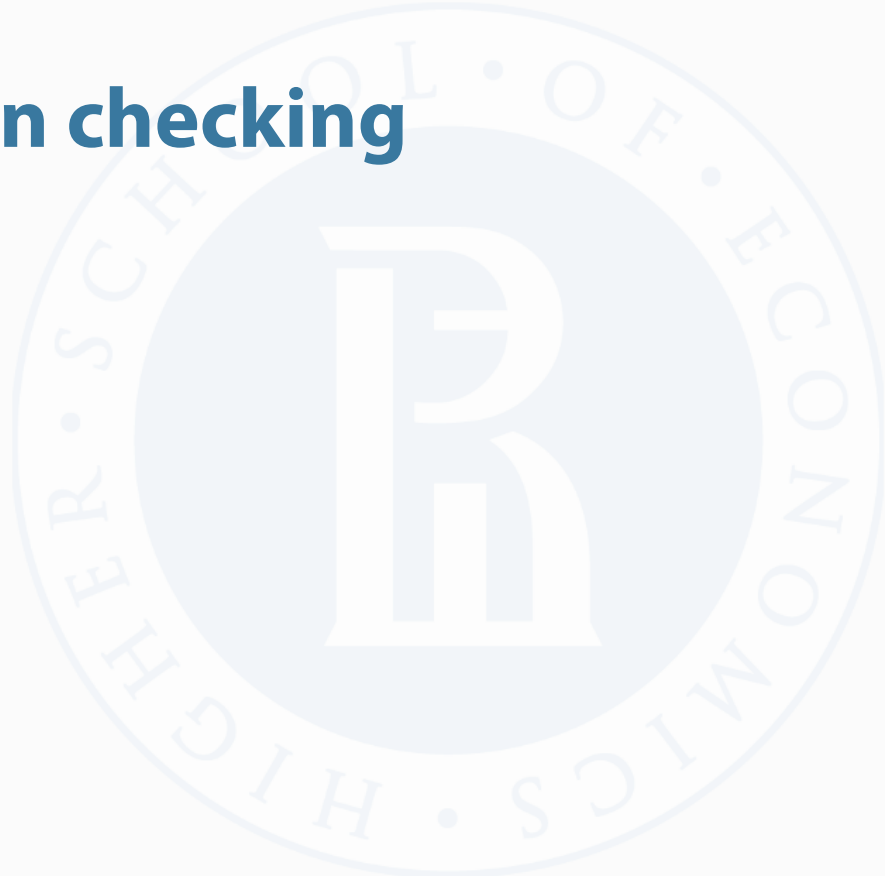
Pair 1

Pair 2



Unpaired (discarded)

Randomization checking



Randomization checking

After randomization and determination of unit properties that are crucial for the result we are going to improve, we need to verify that these properties are homogeneous in control and treatment groups.

Randomization checking

After randomization and determination of unit properties that are crucial for the result we are going to improve, we need to verify that these properties are homogeneous in control and treatment groups.

To verify the homogeneity, we can compare the following measures for 2 randomized groups:

- Sample means
- Sample medians
- Sample variances

Null hypothesis (H_0) for testing: the means / medians / variances of A and B samples are the same

Randomization checking. Sample means

Sample mean — an arithmetic average of target values in a sample

$$\bar{Y} = \frac{1}{N} \cdot \sum_{i=1}^{N} y_i$$

y_i — value of unit property

N — size of sample

Randomization checking. Sample means

Sample mean — an arithmetic average of target values in a sample

$$\bar{Y} = \frac{1}{N} \cdot \sum_{i=1}^{N} y_i$$

y_i — value of unit property

N — size of sample

$\bar{Y}$ is used as an estimator of $\boldsymbol{\mu}$, which is true mean of general population

$H_0: \boldsymbol{\mu}_A = \boldsymbol{\mu}_B$

A, B — indicator of A sample and B sample

Randomization checking. Sample variances

Sample variance — a measure of how far in terms of target variable units are spread out from the average value

$$s^2 = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (y_i - \bar{Y})$$

y_i — value of unit property

N — size of sample

$\bar{Y}$ — sample mean

Randomization checking. Sample variances

Sample variance — a measure of how far in terms of target variable units are spread out from the average value

$$s^2 = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (y_i - \bar{Y})$$

y_i — value of unit property

N — size of sample

$\bar{Y}$ — sample mean

s^2 is used as unbiased estimator of $\boldsymbol{\sigma}^2$: true variance of general population

$H_0: \boldsymbol{\sigma}^2{}_A = \boldsymbol{\sigma}^2{}_B$

A, B — indicator of A sample and B sample

Randomization checking. Sample medians

Sample median — a value that separates the sample so that one half of sample contains values that are less than the estimated median and the other one half with values that are greater than the estimated median.

Randomization checking. Sample medians

Sample median — a value that separates the sample so that one half of sample contains values that are less than the estimated median and the other one half with values that are greater than the estimated median.

To calculate sample medium:

1. Sort the values in ascending order
2. For odd sample — a median is a value with $\left(\frac{N-1}{2}+1\right)$ ordinal number
3. For even sample — a median is a sum of values with ordinal numbers $\left(\frac{N}{2}\right)$ and $\left(\frac{N}{2}+1\right)$ that is divided by 2

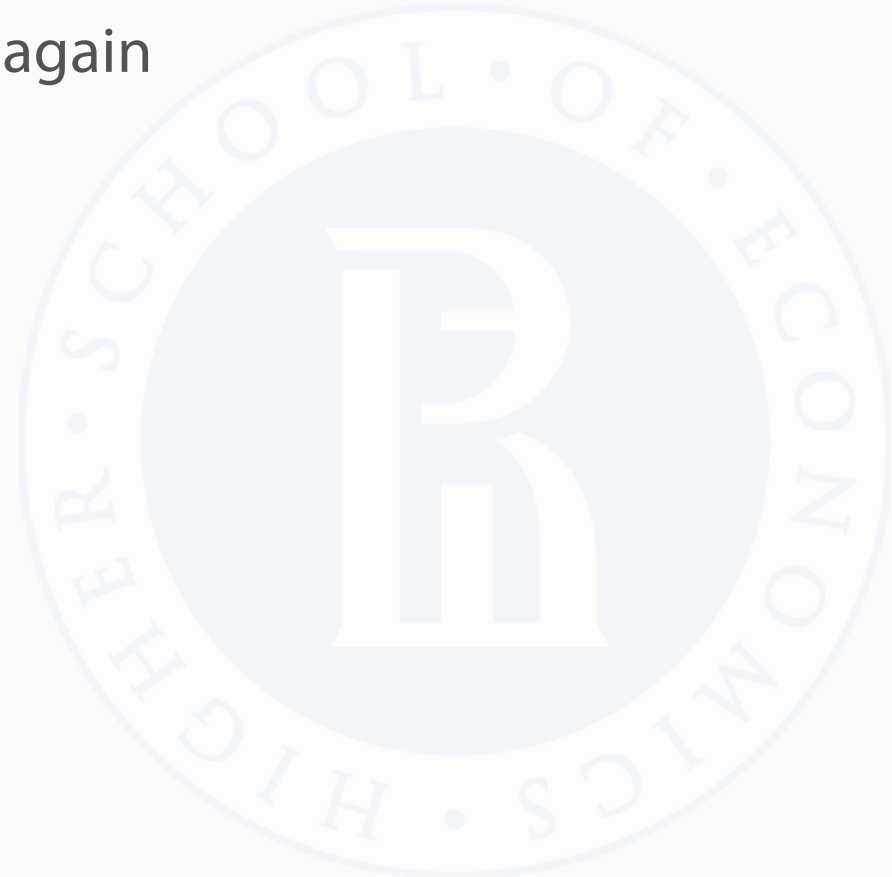
$$H_0: Median_A = Median_B$$

Randomization checking

If we observe significant inequality of means, medians or variances it indicates that

- the randomization was conducted incorrectly

→ we have to randomize units again



Randomization checking

If we observe significant inequality of means, medians or variances it indicates that

- the randomization was conducted incorrectly

→ we have to randomize units again

- or there are factors that explain this difference and do not allow to randomize units without additional adjustment

→ we have to analyze the causes and change the terms of a pilot respectively (e.g. by creating subsamples in groups A and B, by creating synthetic control groups)

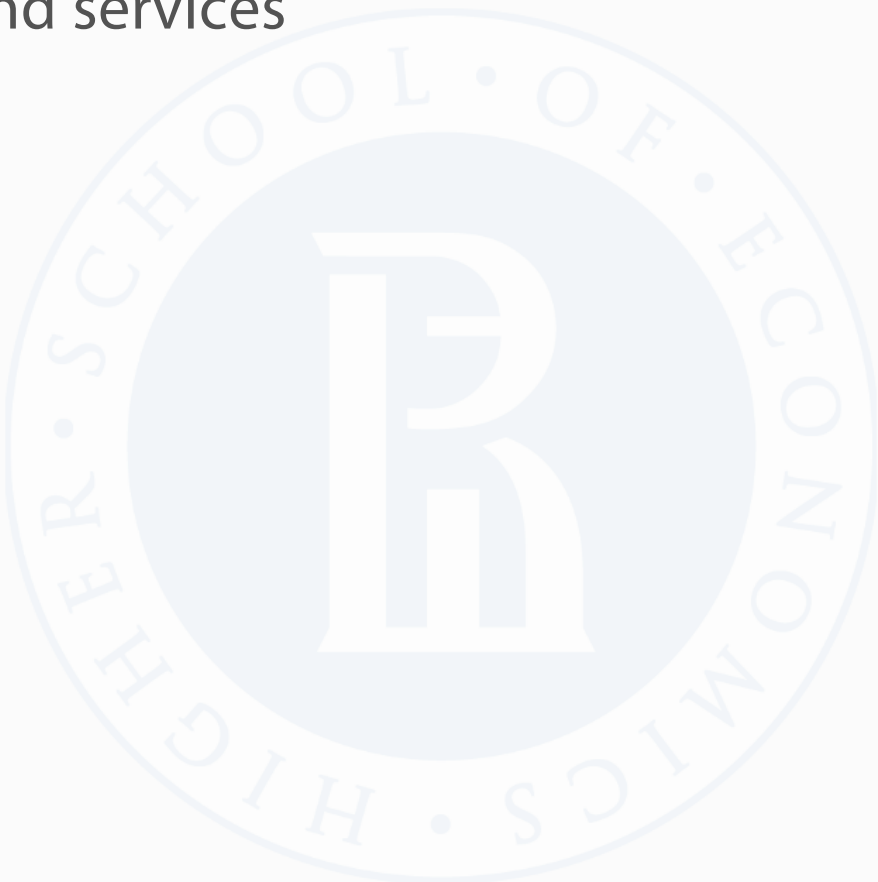
A/B testing preparation. Negative effects

Before running a pilot the following risks must be assessed to design A/B testing appropriately:

A/B testing preparation. Negative effects

Before running a pilot the following risks must be assessed to design A/B testing appropriately:

- Cannibalization — a situation when our initiatives to boost sales of one product or service worsen the sales of our other competitive products and services



A/B testing preparation. Negative effects

Before running a pilot the following risks must be assessed to design A/B testing appropriately:

- Cannibalization — a situation when our initiatives to boost sales of one product or service worsen the sales of our other competitive products and services
- Cream skimming — a situation when our initiatives to boost sales have a maximum short-term effect at the beginning (reassessment of customer base) and then degrade

A/B testing preparation. Negative effects

Before running a pilot the following risks must be assessed to design A/B testing appropriately:

Information asymmetry effects:

- Moral hazard — a concept that client's behavior may change because of transferring the cost of risk on the other party (e.g. detecting clients with extra-risk and selling them insurance makes customers more reckless and risk-taking)

A/B testing preparation. Negative effects

Before running a pilot the following risks must be assessed to design A/B testing appropriately:

Information asymmetry effects:

- Moral hazard — a concept that client's behavior may change because of transferring the cost of risk on the other party (e.g. detecting clients with extra-risk and selling them insurance makes customers more reckless and risk-taking)
- Adverse selection — a concept that our initiatives to select clients who have higher risks and offer them higher price may worsen company's profits as the portion of more risky clients will increase and average margin will decrease

Wrap-up

1. Before running an A/B test we should list all relevant clients', business process features that can have influence on target event

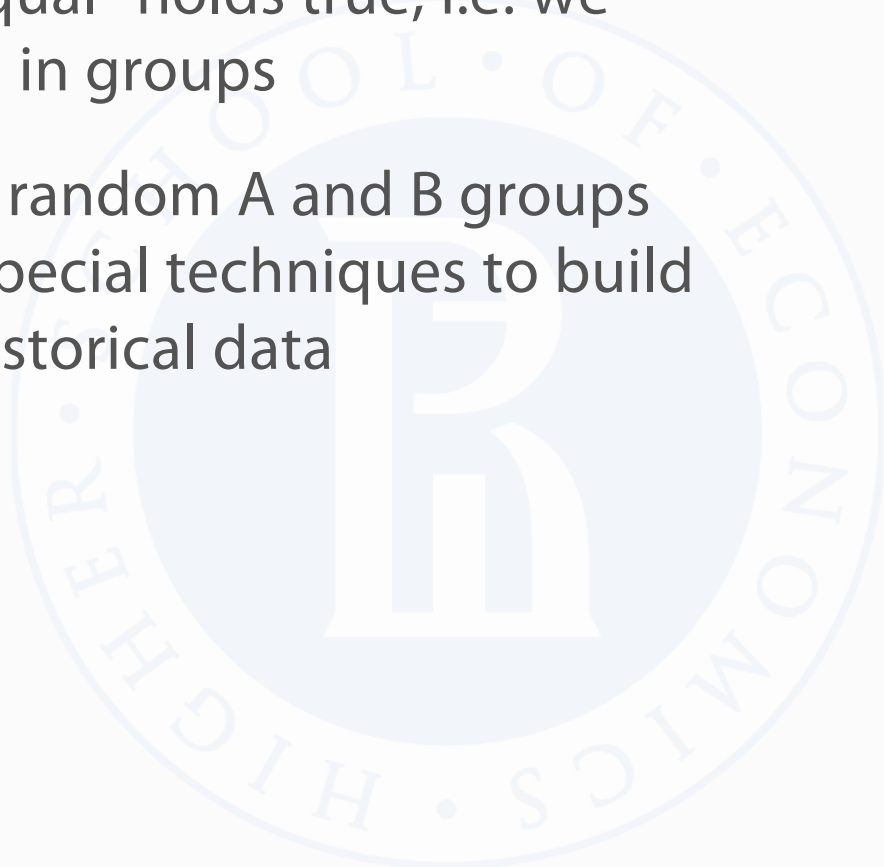
Wrap-up

1. Before running an A/B test we should list all relevant clients', business process features that can have influence on target event

2. Next, we ensure that in terms of the listed features the principle "other things equal" holds true, i.e. we have to check randomization in groups

Wrap-up

1. Before running an A/B test we should list all relevant clients', business process features that can have influence on target event
2. Next, we ensure that in terms of the listed features the principle "other things equal" holds true, i.e. we have to check randomization in groups
3. In case when designing truly random A and B groups is not available, we can use special techniques to build synthetic groups based on historical data



Wrap-up

1. Before running an A/B test we should list all relevant clients', business process features that can have influence on target event
2. Next, we ensure that in terms of the listed features the principle "other things equal" holds true, i.e. we have to check randomization in groups
3. In case when designing truly random A and B groups is not available, we can use special techniques to build synthetic groups based on historical data
4. When designing an A/B test we should always consider potential negative effects and biases that arise from cannibalization, adverse selection effects etc.

Evaluation of A/B testing results

Object of evaluation

We analyze difference between control group and treatment group in terms of treatment effect:

Object of evaluation

We analyze difference between control group and treatment group in terms of treatment effect:

Average treatment effect (ATE) — the average difference of the sample parameter

Object of evaluation

We analyze difference between control group and treatment group in terms of treatment effect:

Average treatment effect (ATE) — the average difference of the sample parameter

$$ATE = E(Y_{\mathrm{B}} - Y_{\mathrm{A}}) = E(Y_{\mathrm{B}}) - E(Y_{\mathrm{A}}) \approx \bar{Y}_B - \bar{Y}_A$$

Y_{B} — target value of units that are treated

Y_A — target value of units that are not treated

As Y can serve any measurable (sales, profits, time-to-market, response rate, default event etc.)

Statistical significance testing

Statistical significance testing is a technique to retain or reject the null hypothesis.

Statistical significance testing

Statistical significance testing is a technique to retain or reject the null hypothesis.

Significance testing is used before an A/B test (to check randomization) as well as after A/B test (to verify the effect)

Statistical significance testing

Statistical significance idea:

1. Calculate some function (s.c. statistics) of observed data

Statistical significance testing

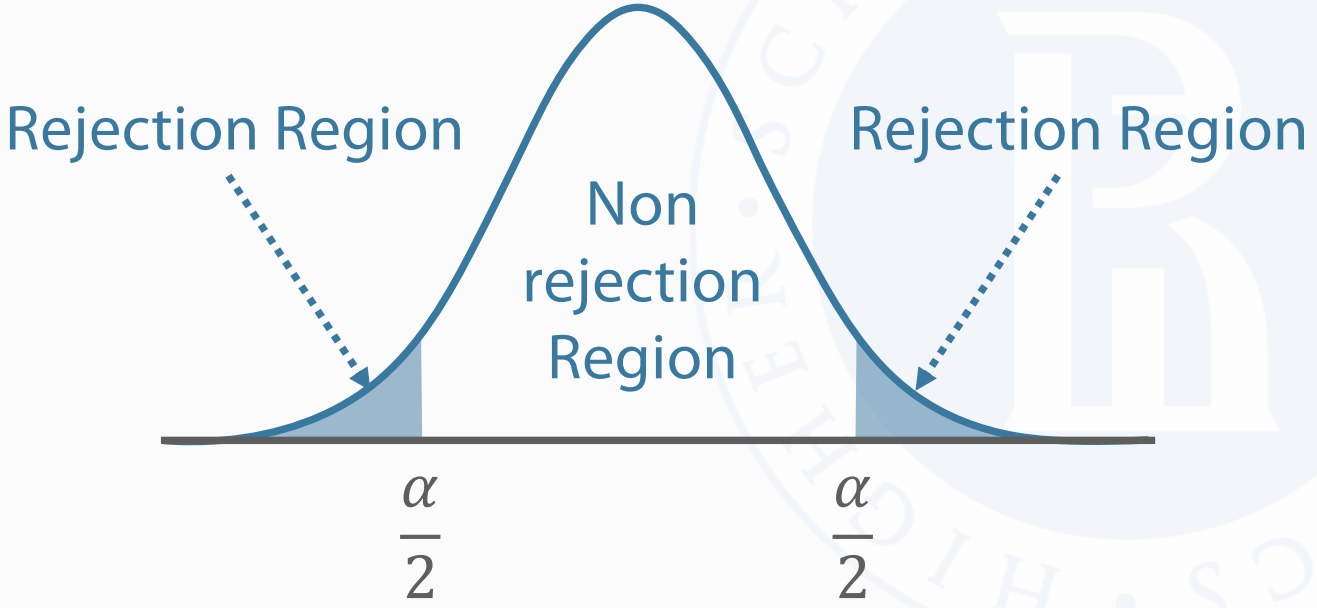
Statistical significance idea:

1. Calculate some function (s.c. statistics) of observed data
2. If null hypothesis is correct, then statistics' distribution is known

Statistical significance testing

Statistical significance idea:

1. Calculate some function (s.c. statistics) of observed data
2. If null hypothesis is correct, then statistics' distribution is known
3. Assess whether statistics value is probable given its distribution

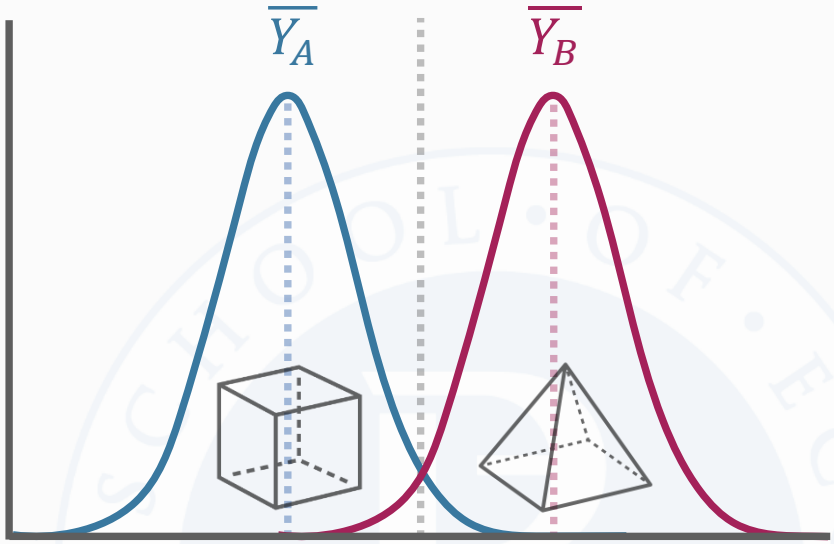


Statistical significance testing. Student's t-test

So, $H_0: \mu_A = \mu_B$,
s.t. the variances in Y within group A and B are similar

Statistical significance testing. Student's t-test

So, $H_0: \mu_A = \mu_B$,
s.t. the variances in Y within group A and B are similar

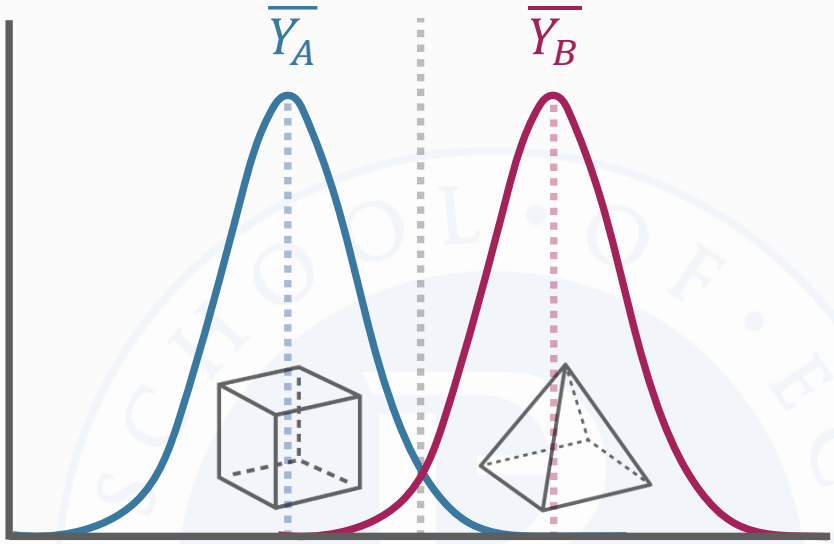


Statistical significance testing. Student's t-test

So, $H_0: \mu_A = \mu_B$,
s.t. the variances in Y within group A and B are similar

$$t = \frac{\bar{Y}_B - \bar{Y}_A}{\boldsymbol{\sigma}\sqrt{\frac{1}{N_A} + \frac{1}{N_B}}}$$

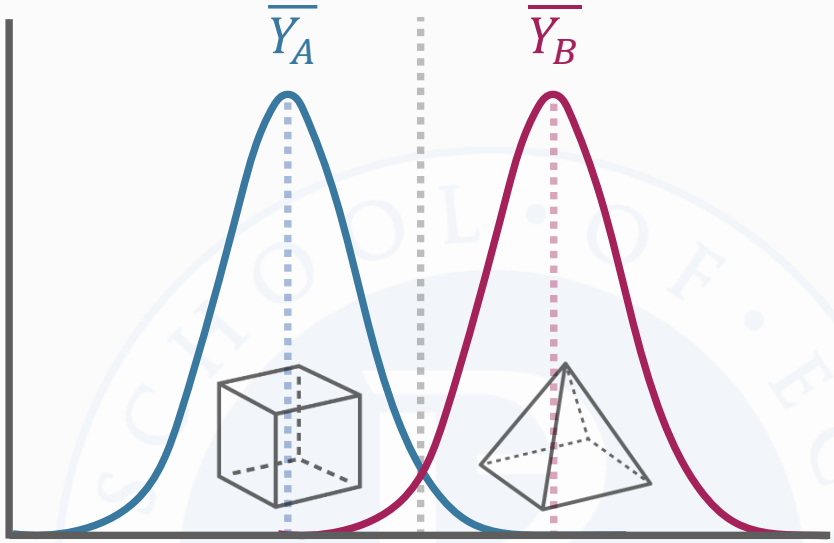
$$t \sim t_{N_A+N_B-2}$$



Statistical significance testing. Student's t-test

So, $H_0: \mu_A = \mu_B$,
s.t. the variances in Y within group A and B are similar

$$t = \frac{\bar{Y}_B - \bar{Y}_A}{\boldsymbol{\sigma}\sqrt{\frac{1}{N_A} + \frac{1}{N_B}}}$$



$$t \sim t_{N_A+N_B-2}$$

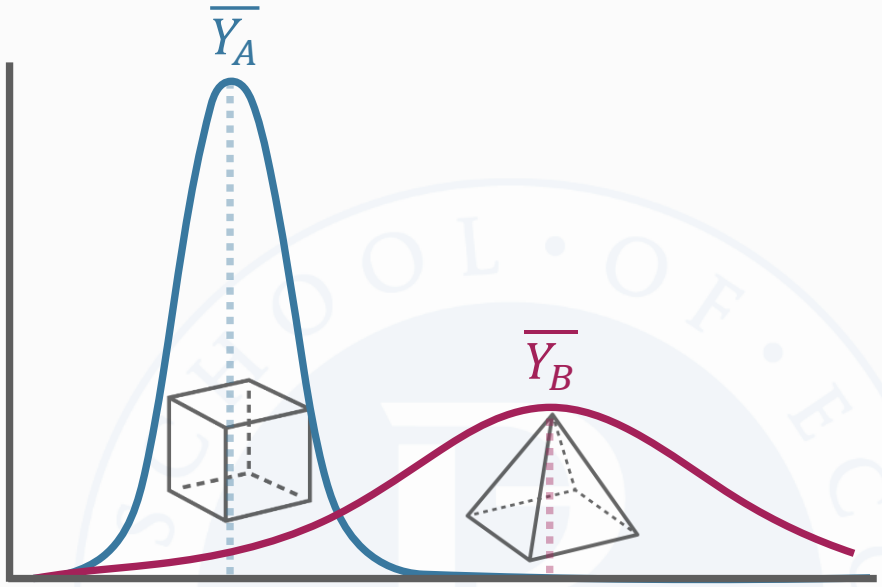
where

$$\boldsymbol{\sigma} = \sqrt{\frac{(N_A - 1)s_A^2 + (N_B - 1)s_B^2}{N_A + N_B - 2}}$$

, $s^2_{A/B}$ are unbiased sample variance estimators

Statistical significance testing. Welch's t-test

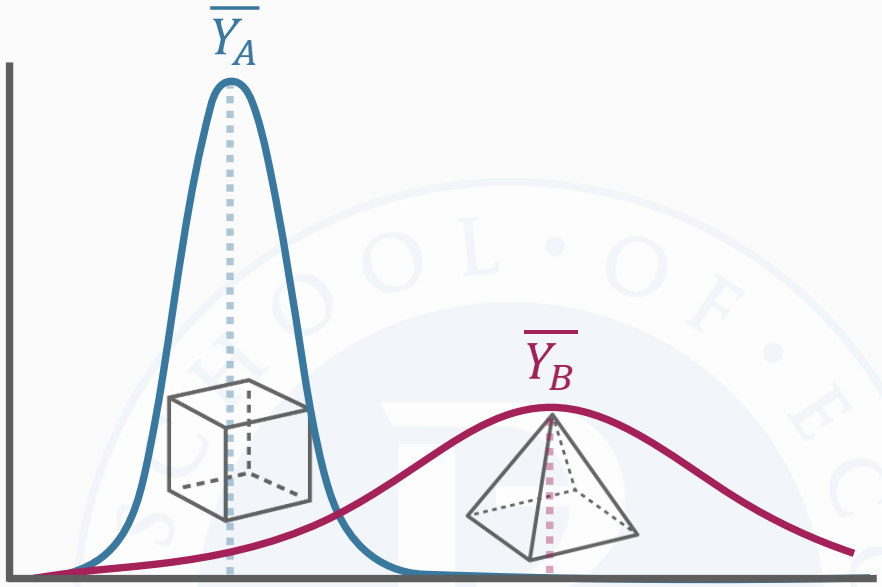
So, $H_0: \mu_A = \mu_B$,
s.t. the variances in Y within group A and B are different



Statistical significance testing. Welch's t-test

So, $H_0: \mu_A = \mu_B$,
s.t. the variances in Y within group A and B are different

$$t = \frac{\bar{Y}_B - \bar{Y}_A}{\sqrt{\frac{s_A^2}{N_A} + \frac{s_B^2}{N_B}}}$$

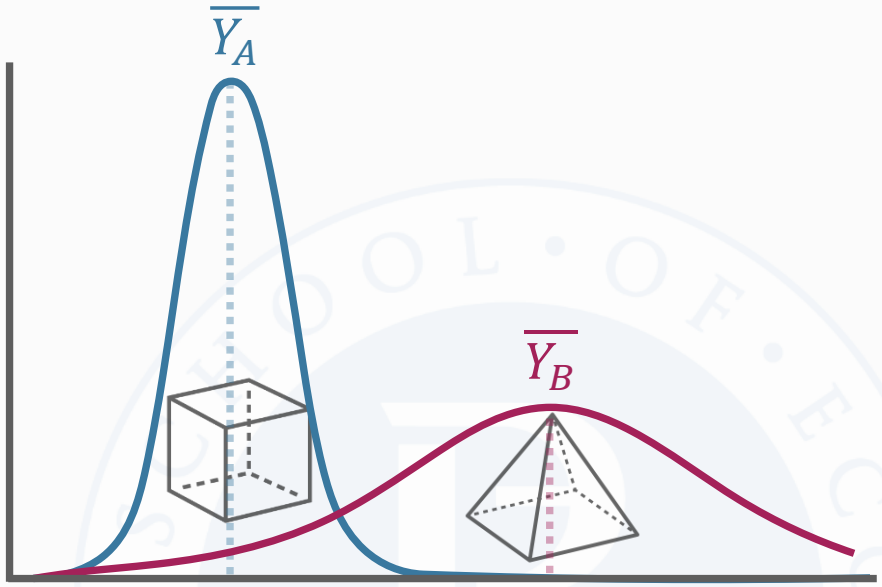


Statistical significance testing. Welch's t-test

So, $H_0: \mu_A = \mu_B$,
s.t. the variances in Y within group A and B are different

$$t = \frac{\bar{Y}_B - \bar{Y}_A}{\sqrt{\frac{s_A^2}{N_A} + \frac{s_B^2}{N_B}}}$$

$$t \sim t_v$$



where

$$v \approx \frac{(s_A^2/N_A + s_B^2/N_B)^2}{\frac{s_A^4}{N_A^2(N_A - 1)} + \frac{s_B^4}{N_B^2(N_B - 1)}}$$

Variance equality test. F-test

Suppose, we have two samples.

$H_0: \sigma_A^2 = \sigma_B^2$

Variance equality test. F-test

Suppose, we have two samples.

$H_0: \sigma_A^2 = \sigma_B^2$

1. We estimate unbiased sample variances

$s_A^2 = \frac{\sum_{i=1}^{N_A}(y_{A,i} - \bar{Y}_A)^2}{N_A - 1}$ and $s_B^2 = \frac{\sum_{i=1}^{N_B}(y_{B,i} - \bar{Y}_B)^2}{N_B - 1}$

Variance equality test. F-test

Suppose, we have two samples.

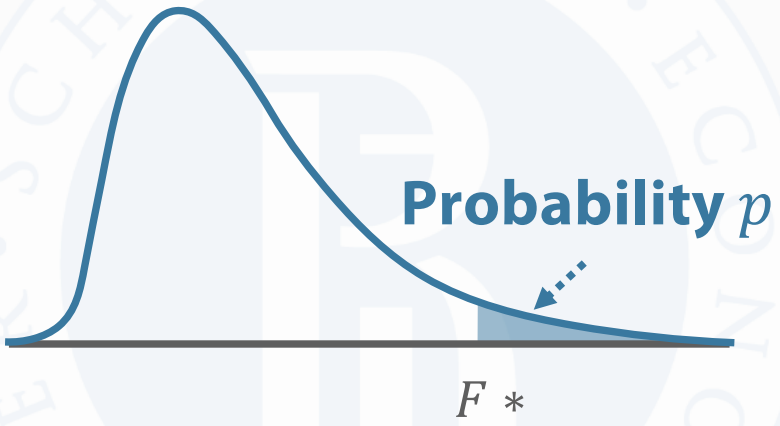
$H_0: \sigma_A^2 = \sigma_B^2$

1. We estimate unbiased sample variances

   $s_A^2 = \frac{\sum_{i=1}^{N_A}(y_{A,i} - \bar{Y}_A)^2}{N_A - 1}$ and $s_B^2 = \frac{\sum_{i=1}^{N_B}(y_{B,i} - \bar{Y}_B)^2}{N_B - 1}$

2. We calculate main statistics

   $$\mathrm{F} = \frac{s_A^2}{s_B^2} \sim F_{N_A-1, N_B-1}$$



Variance equality test. F-test

Suppose, we have two samples.

$H_0: \sigma_A^2 = \sigma_B^2$

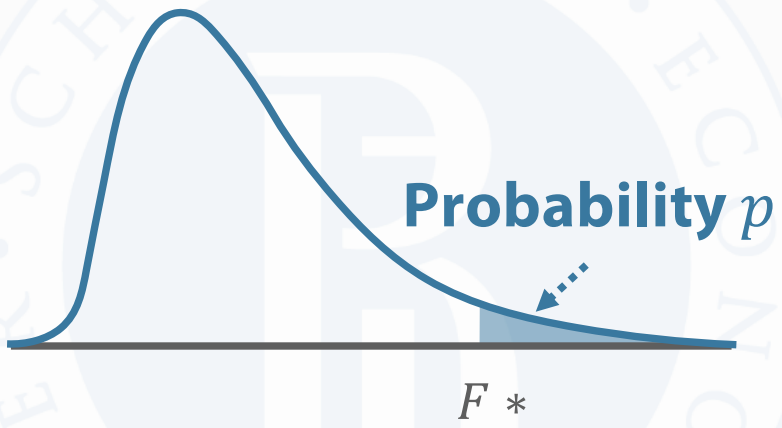
1. We estimate unbiased sample variances

   $s_A^2 = \frac{\sum_{i=1}^{N_A}(y_{A,i} - \bar{Y}_A)^2}{N_A - 1}$ and $s_B^2 = \frac{\sum_{i=1}^{N_B}(y_{B,i} - \bar{Y}_B)^2}{N_B - 1}$

2. We calculate main statistics

   $$\mathrm{F} = \frac{s_A^2}{s_B^2} \sim F_{N_A-1, N_B-1}$$

3. We check out the table with critical values for given confidence level α



Variance equality test. Bartlett's test

Suppose, we have multiple samples.

Variance equality test. Bartlett's test

Suppose, we have multiple samples.

$H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \ldots$

$H_A: \exists i, j: \sigma_i^2 \neq \sigma_j^2$ (at least two are different)

Variance equality test. Bartlett's test

Suppose, we have multiple samples.

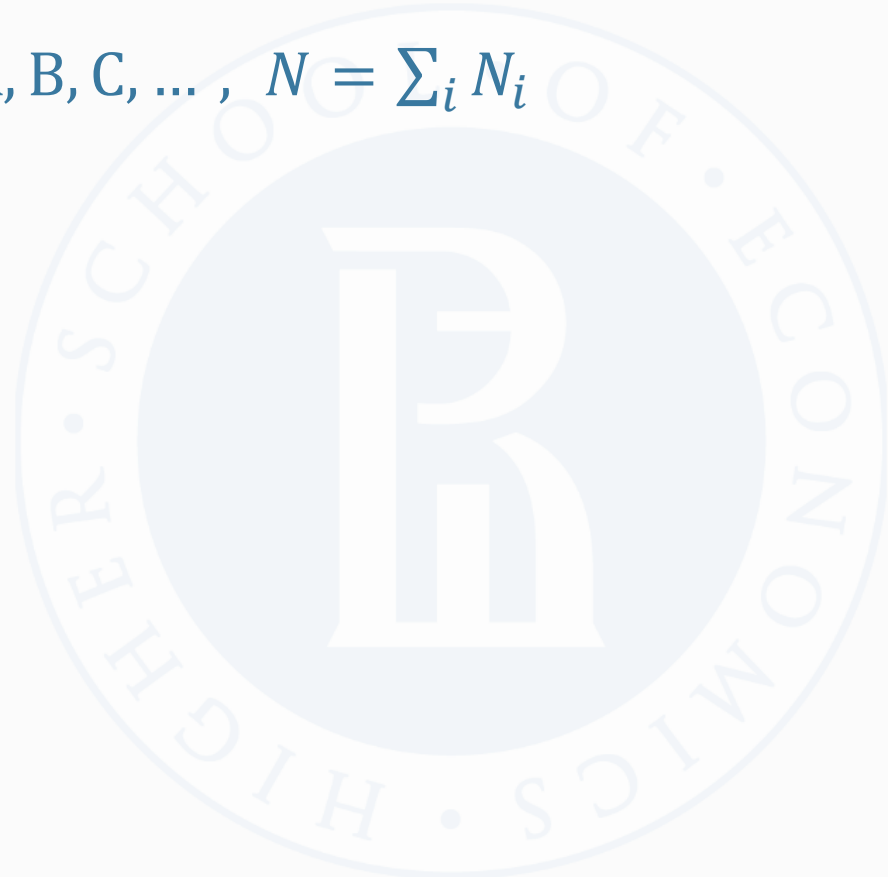
$H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \ldots$

$H_A: \exists i, j: \sigma_i^2 \neq \sigma_j^2$ (at least two are different)

1. We estimate s.c. *pooled* sample variance:

   $s_p^2 = \frac{\sum_i (N_i - 1) s_i^2}{N - k}$, where $i = \mathrm{A}, \mathrm{B}, \mathrm{C}, \ldots$, $N = \sum_i N_i$

   and k is number of samples



Variance equality test. Bartlett's test

Suppose, we have multiple samples.

$H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \ldots$

$H_A: \exists i, j: \sigma_i^2 \neq \sigma_j^2$ (at least two are different)

1. We estimate s.c. *pooled* sample variance:

   $s_p^2 = \frac{\sum_i (N_i - 1) s_i^2}{N - k}$, where $i = \mathrm{A}, \mathrm{B}, \mathrm{C}, \ldots$, $N = \sum_i N_i$ and k is number of samples

2. We calculate main statistics:

$$\chi^2 = \frac{(N-k)\ln\left(s_p^2\right) - \sum_i (N_i - 1)\ln(s_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_i \left(\frac{1}{N_i - 1}\right) - \frac{1}{N-k}\right)} \sim \chi_{k-1}^2$$

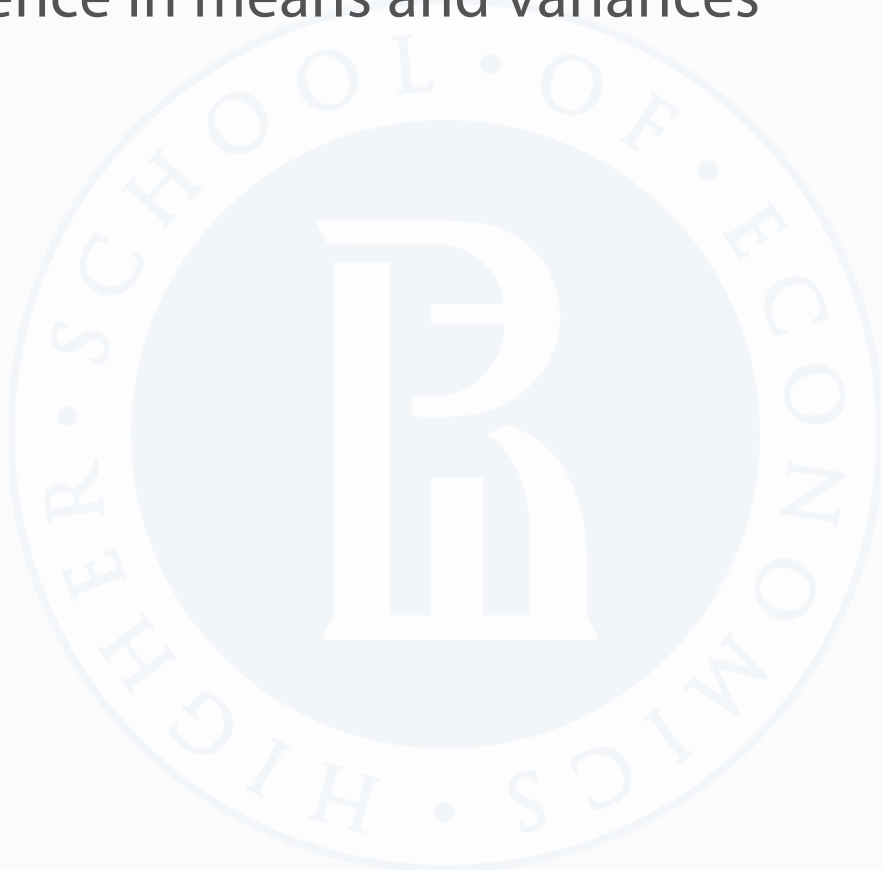
➤ *Bartlett, M. S. (1937). "Properties of sufficiency and statistical tests"*

Wrap-up

1. Evaluating A/B test results boils down to estimating whether there is a statistically significant difference in distributions of target event within A and B groups

Wrap-up

1. Evaluating A/B test results boils down to estimating whether there is a statistically significant difference in distributions of target event within A and B groups
2. There are several specially designed statistical tests that allows one to consider difference in means and variances of testing groups



Wrap-up

1. Evaluating A/B test results boils down to estimating whether there is a statistically significant difference in distributions of target event within A and B groups
2. There are several specially designed statistical tests that allows one to consider difference in means and variances of testing groups
3. Different tests have to be applied in relevant conditions, such as equality or inequality of variances within testing groups and total number of testing groups