Hypothesis testing for leakage assessment in side channel analysis

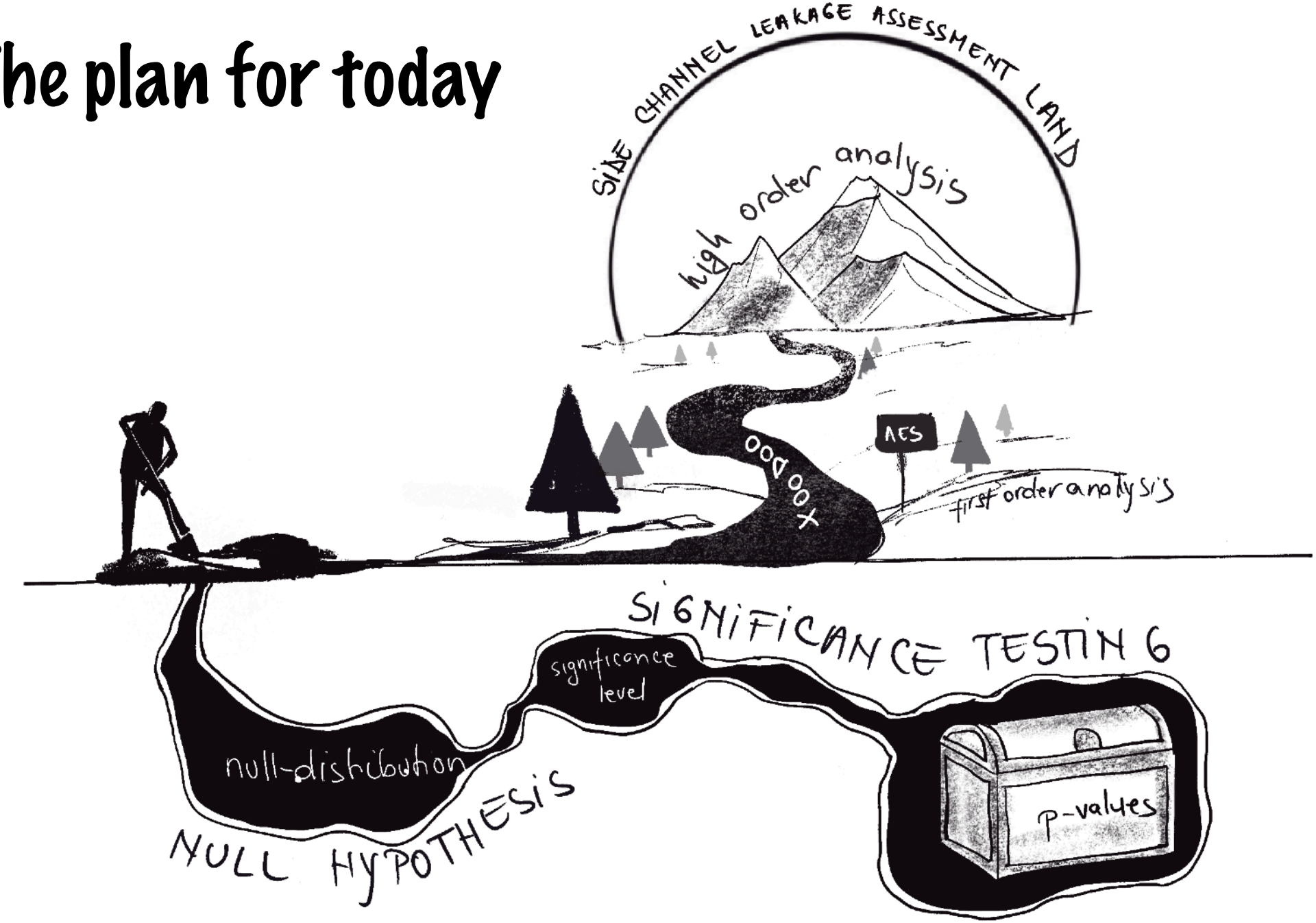
Making decisions is easy, making the right decision less so

Ileana Buhan, June 2023

@ileanabuhan

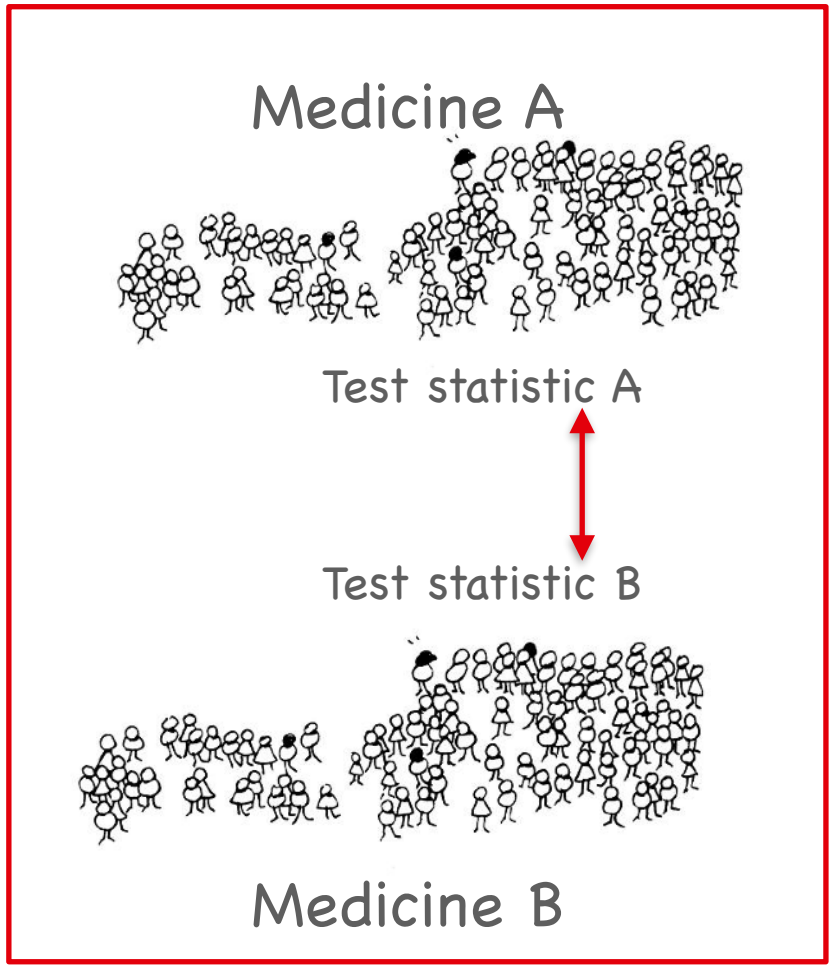
Radboud University

The plan for today

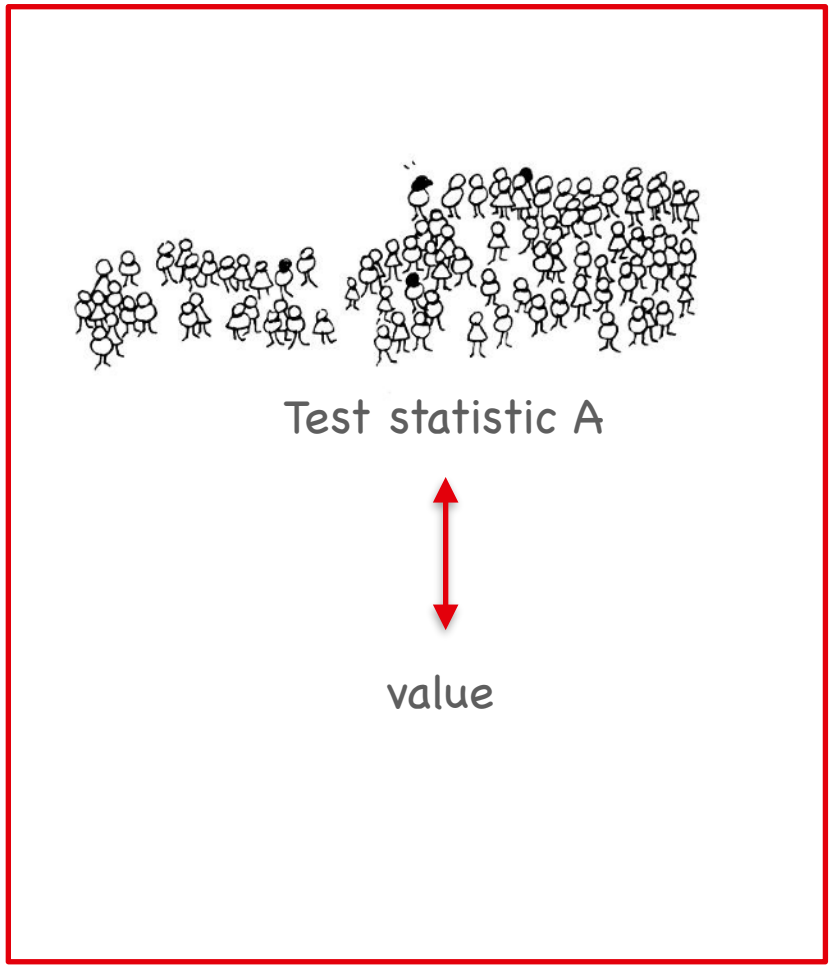


Null Hypothesis Significance Testing (NHST)

Two types of questions



Two-sample test



One-sample test

What is hypothesis?

A hypothesis* is tentative assumption made in order to draw out and test its logical or empirical consequences.

*Source for definition https://www.merriam-webster.com/dictionary/hypothesis

Food for brain

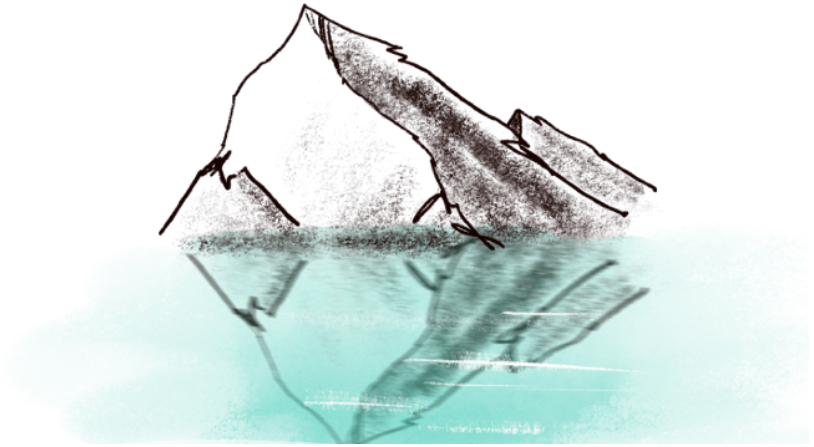
Which of the following statements are hypotheses?

1. The average height of all PhD students is 1.93m;
2. Gas in NL in expensive;
3. Green trolls have an average height if 54cm;
4. The ratio of left- to right handed people in Croatia is equal;
5. Eminem would make a better presindent for the USA than Justin Bieber;

Answer: 1 and 4.

Population vs sample data

The average concentration of salt for the water in the lake is 3%.



Population

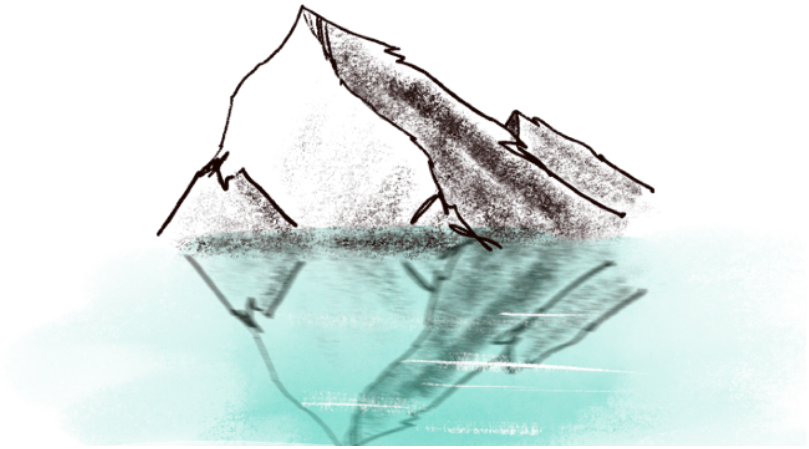
μ, σ

Sample data

$\bar{x}, s$

Population vs sample data

The average concentration of salt for the water in the lake is 3%.



Population

μ, σ

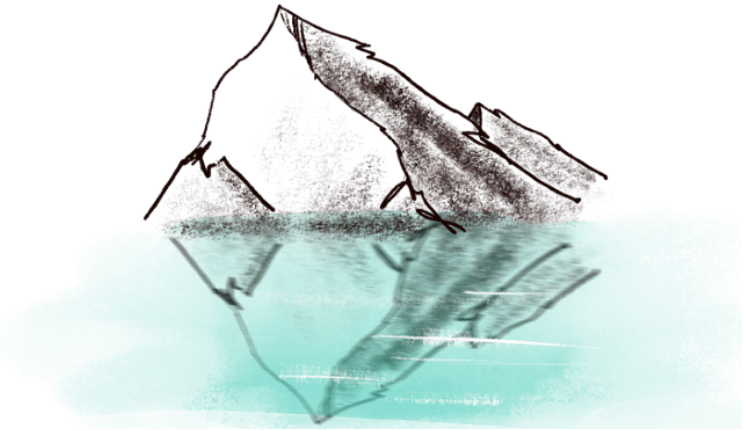
Is this question:

(A)Two-sample test
(B) One-sample test

Answer: the B

What is hypothesis testing?

A tool for making decisions about a population (lake) given some sample data (glass of water).



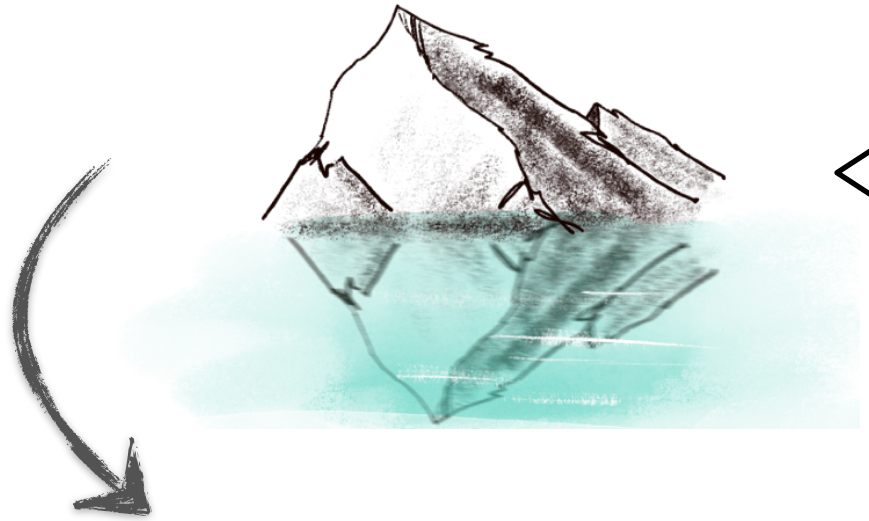
Hypothesis test

A hypothesis test evaluates two mutually exclusive statements about a _population_ to determine which statement is best supported by the _sample data_.

Test statistic

A test statistic is a number calculated from sample data that is used to evaluate hov compatible the experimental results are with the hypothesis test.

population parameter



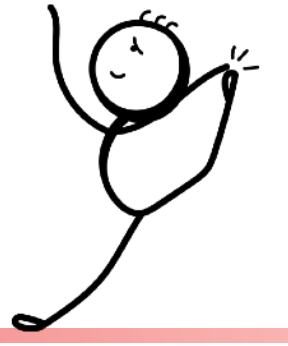
Hypothesis test

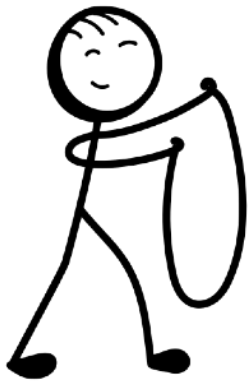
test statistic

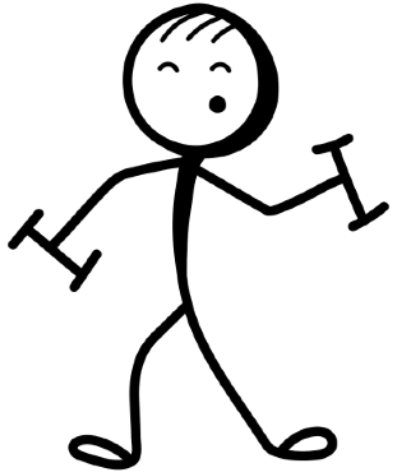
The average concentration of salt for the water in the lake is 3%.

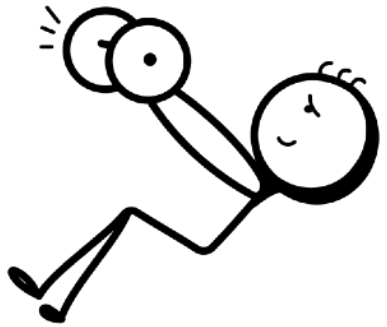
The average concentration of salt for the water in our glass is 2.7%.

Exercise 1 (1,2,3)



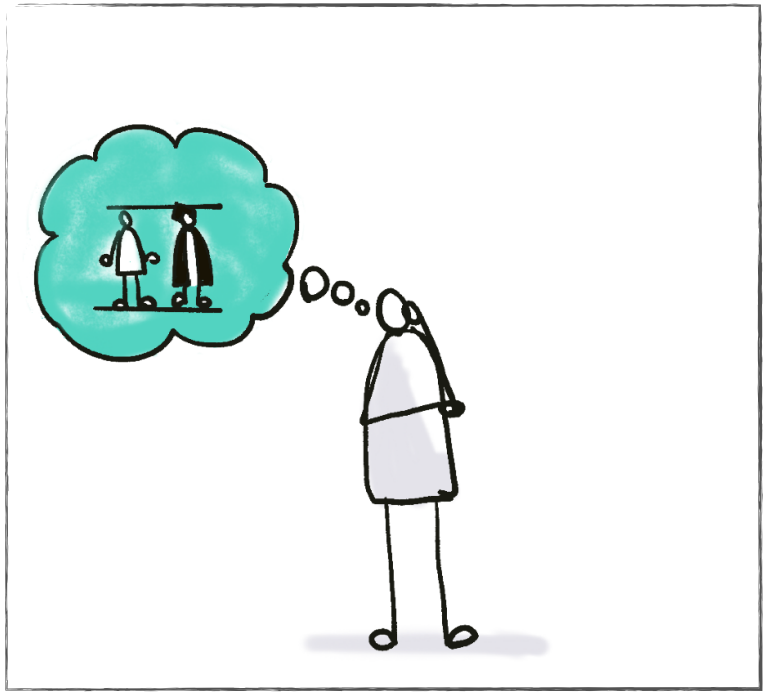




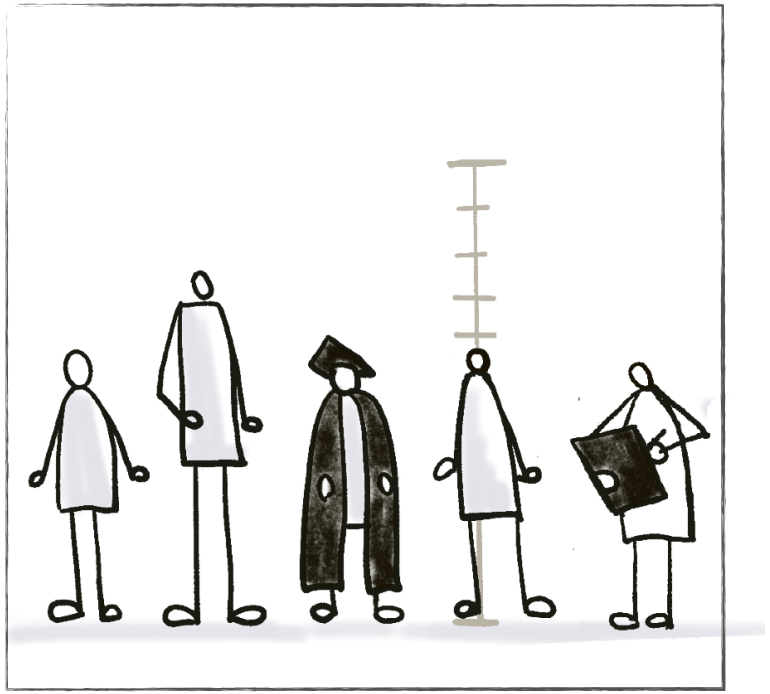


NHST in three steps

1. Select the H_0, α;



2. Collect data

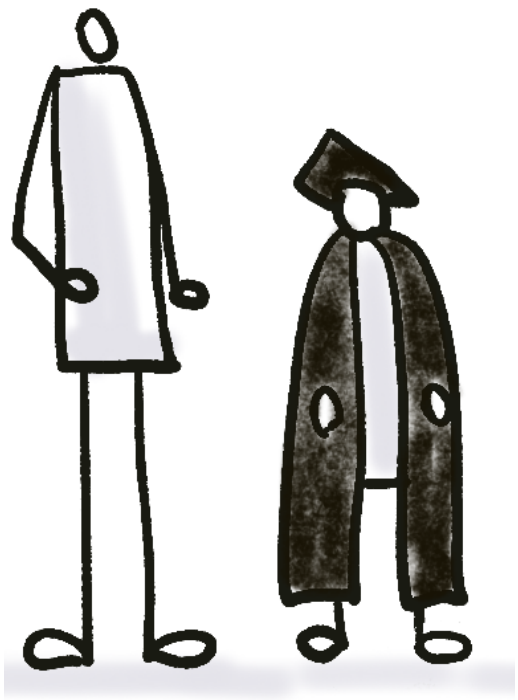


3. Test

Hypothesis testing quantifies how unusual the data is, assuming the null hypothesis to be true.

Lets test!

Are PhD students taller than faculty staff?



Is this question:

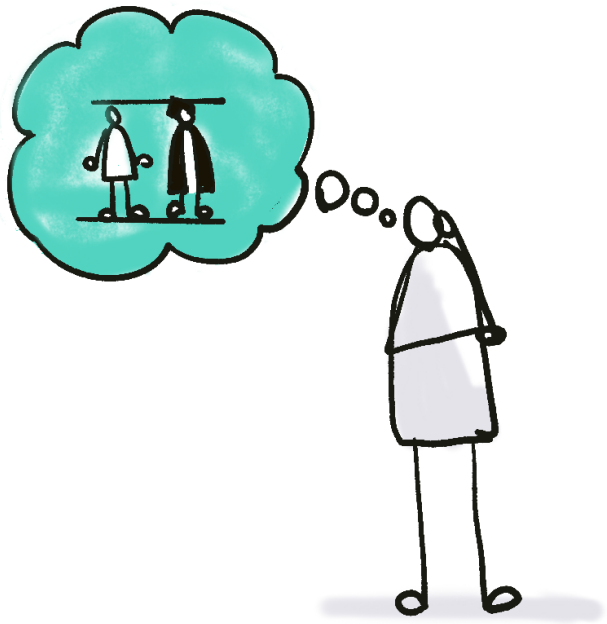
(A)Two-sample test
(B) One-sample test

Answer: (A)

We need two mutually exclusive statements for testing:

1. the null-hypothesis H_0
2. the alternative hypothesis H_a

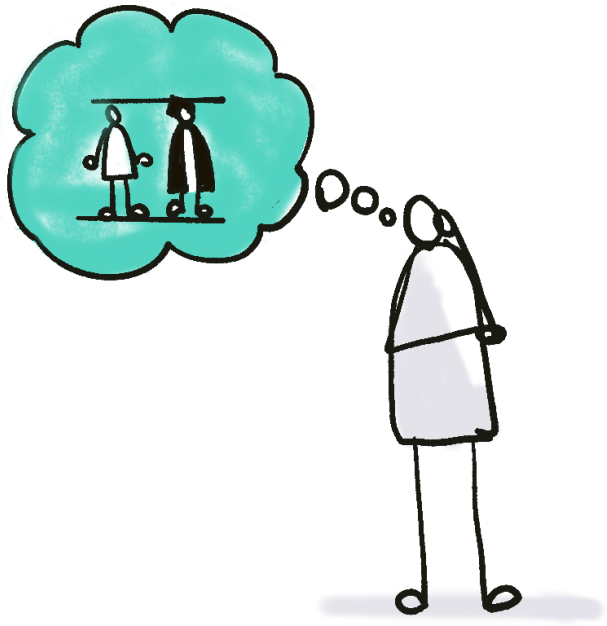
1. Select the null



The null hypothesis (H_0) is a specific statment about a population parameter generated by the researcher for the purpose of an argument.

A good null-hypothesis is interesting to reject and must be specific.

Which statement is a good choice for H_0?



1. The polio vaccine has no effect on the probability of developing paralytic polio;
2. Adding free gifts does not increase sales;
3. The power consumption of this device does not depend of the processed data;
4. The ratio of left -, right - handed people is equal in the population;

Answer: all.

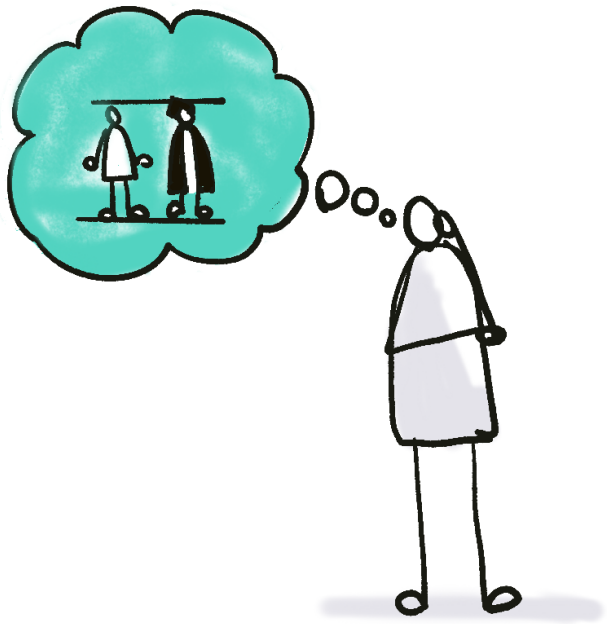
The null vs the alternative

The two hypothesis are NOT equal. The only hypothesis tested with the data is H_0.

A good null-hypothesis is interesting to reject and must be specific.

The alternative-hypothesis is not-specific and contains all other values.

Which statement is more suited as H_0 or H_a :

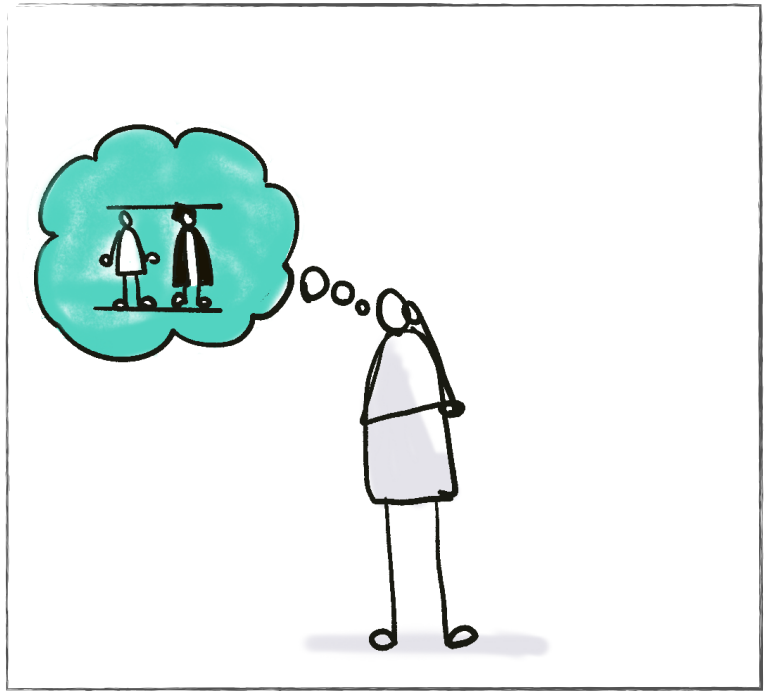


1. The number of hours preschool children spend watching TV affects how they behave at daycare.
2. Social skills influence the number of friends a person has.
3. Smoking influences the risk of allergies.
4. Growth rate of trees are unaffected by increases in carbondioxide levels in the atmosphere.

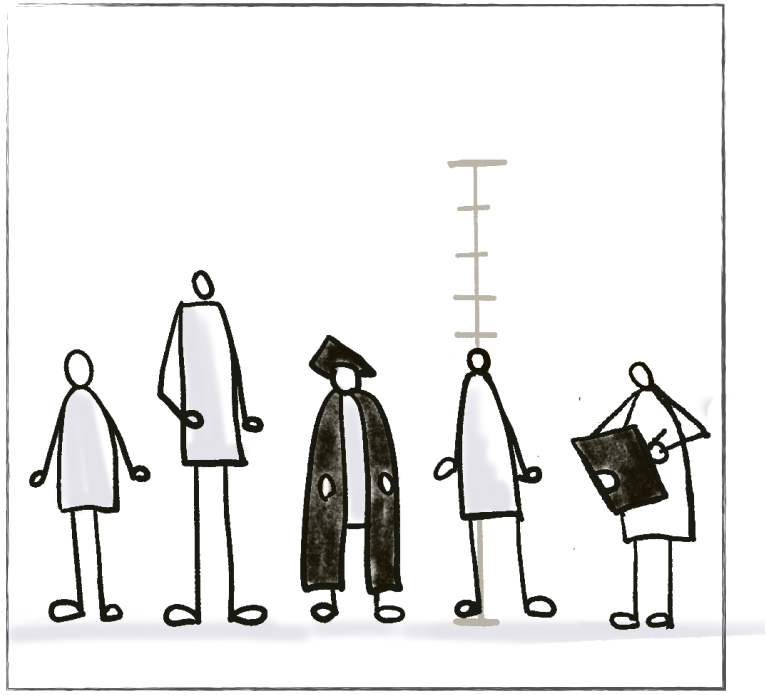
Answer: 1, 2, 3 H_a and 4, H_0.

NHST in three steps

1. Select the H_0, α;

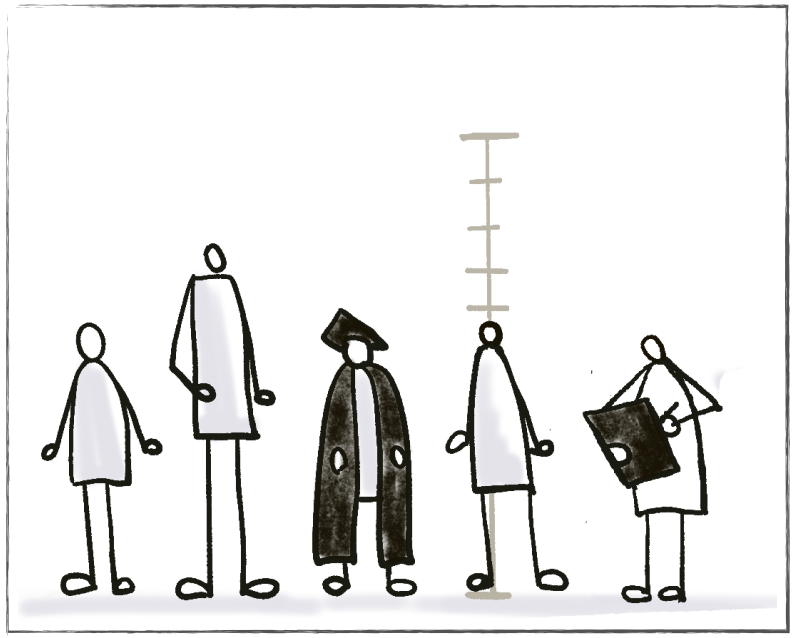


2. Sample data



3. Test

2. Sample data



The sample data must be representative for population.

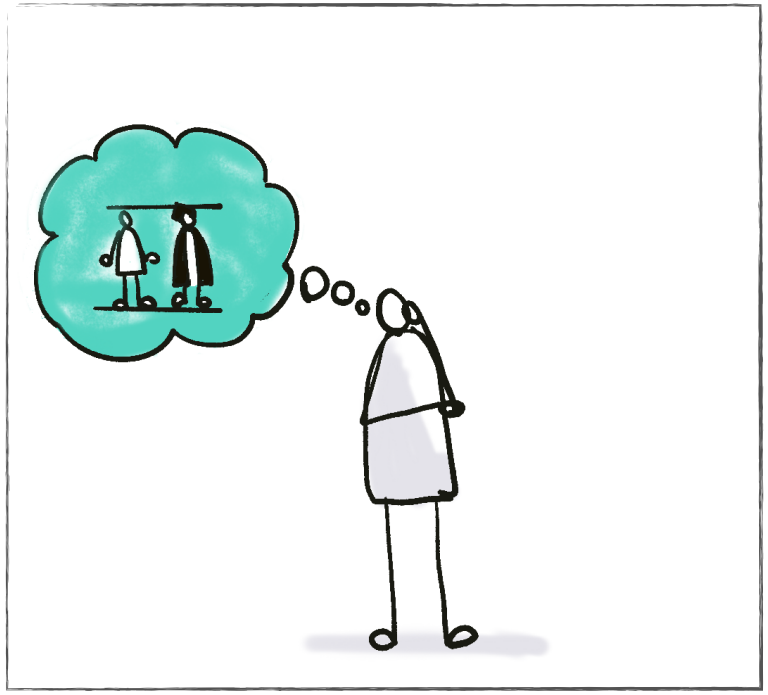
Known techniques:

- Random sampling
- Counting off
- Convenience sampling
- ....

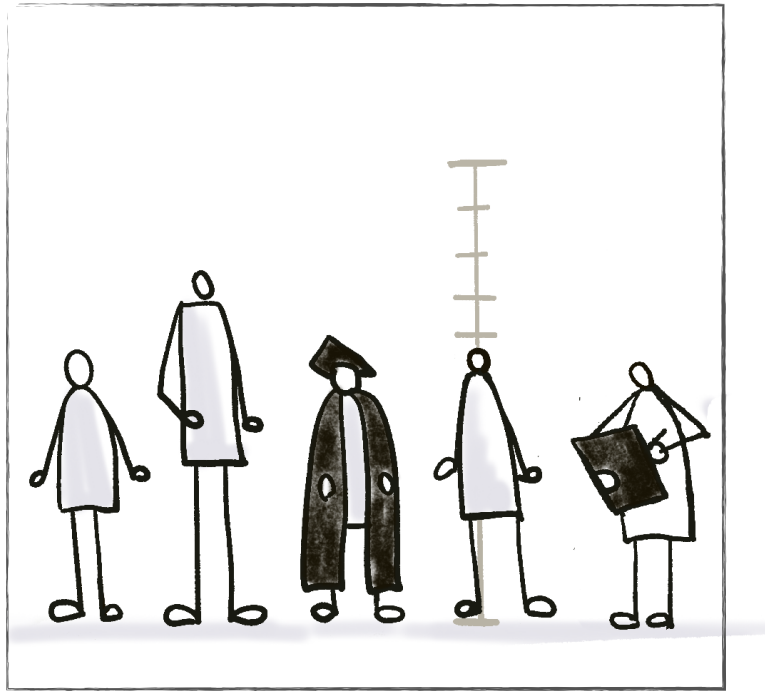
We will revisit the topic when discussing side-channel traces.

NHST in three steps

1. Select the null



2. Sample data



3. Test

3. Test the significance

null-distribution

How do we describe the universe where the null-hypothesis is true?

In a universe where the null-hypothesis is true, how surprized are we by the obversed data?

How do we measure surprize?

p-values

Significance level α

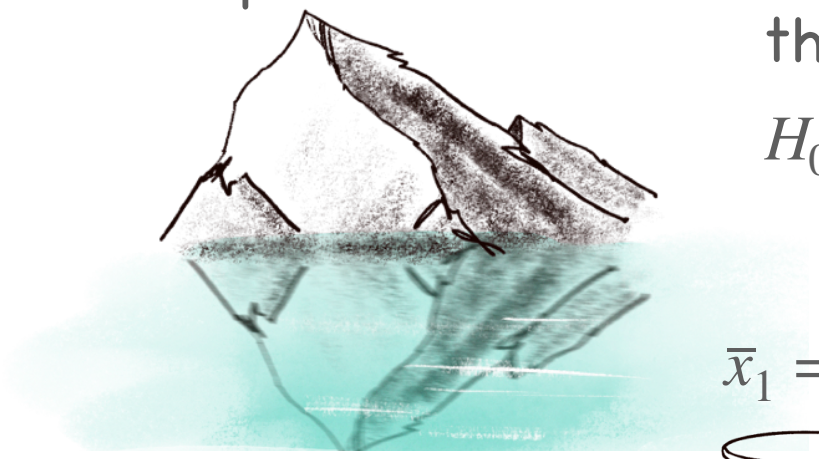
The null-distribution

The null-distribution is the sampling distribution of the outcomes for a test statistic under the assumption that the null-hypothesiss is true.

The null-distribution

The null-distribution is the sampling distribution of the outcomes for a test statistic under the assumption that the null hypothesis is true.

Example 1:*



The average concentration of salt for the water in the lake is 3%.

$H_0 : \mu_{salt} = 0.3$

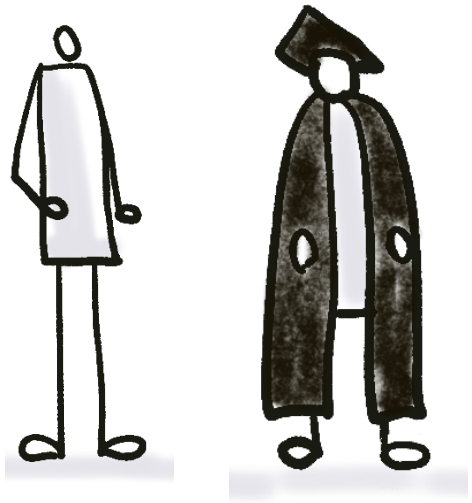
$\bar{x}_1 = 0.23$ $\bar{x}_2 = 0.2$ $\bar{x}_3 = 0.19$ $\bar{x}_4 = 0.35$ $\bar{x}_5 = 0.35$

*one-sample t-test

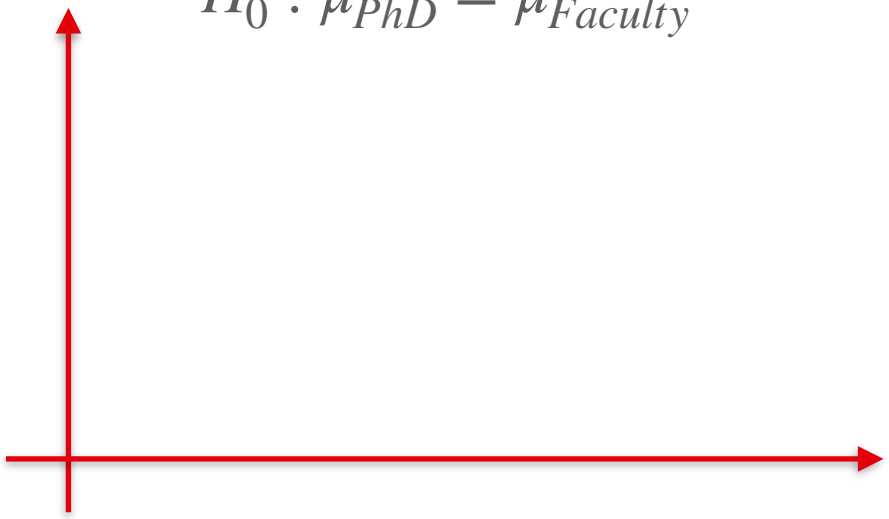
The null-distribution

The null-distribution is the sampling distribution of the outcomes for a test statistic under the assumption that the null hypothesis is true.

Example 2*



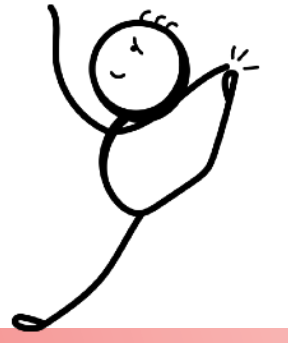
$$H_0 : \mu_{PhD} = \mu_{Faculty}$$

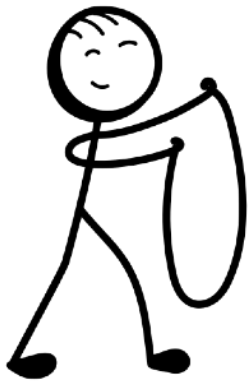


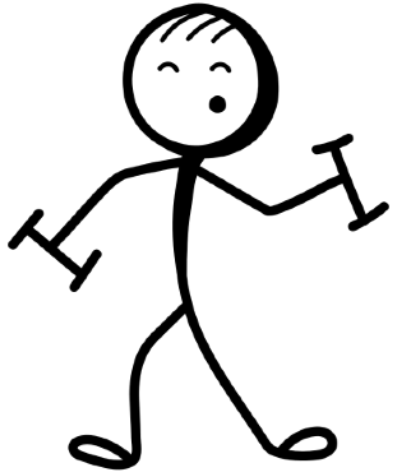
$\overline{x}$

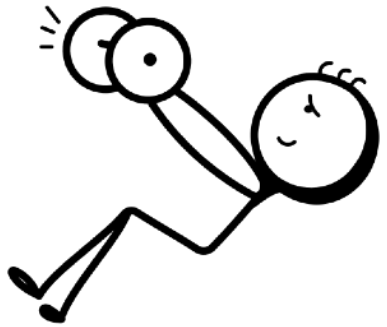
*two-sample t-test

Exercise 2







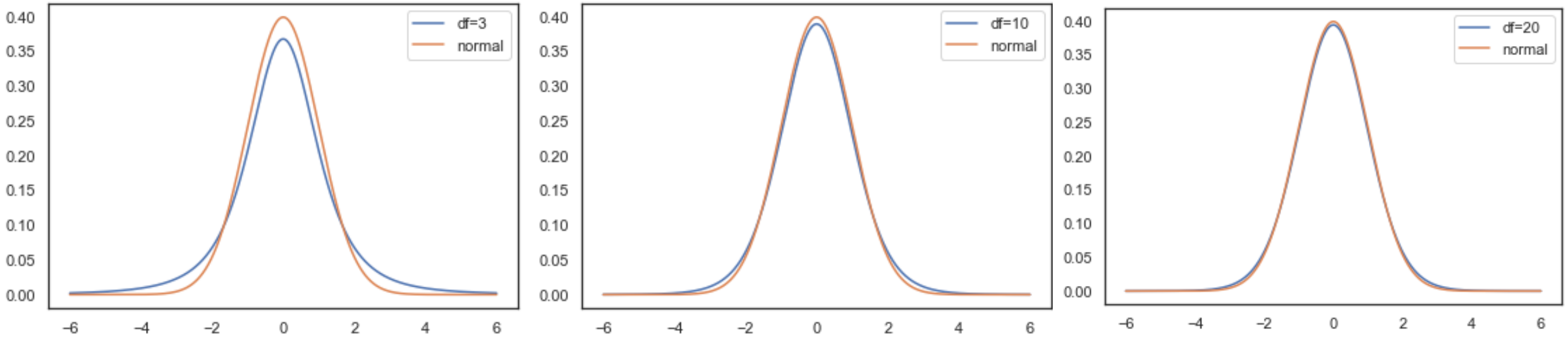


T-score

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s}{n-1}}}$$

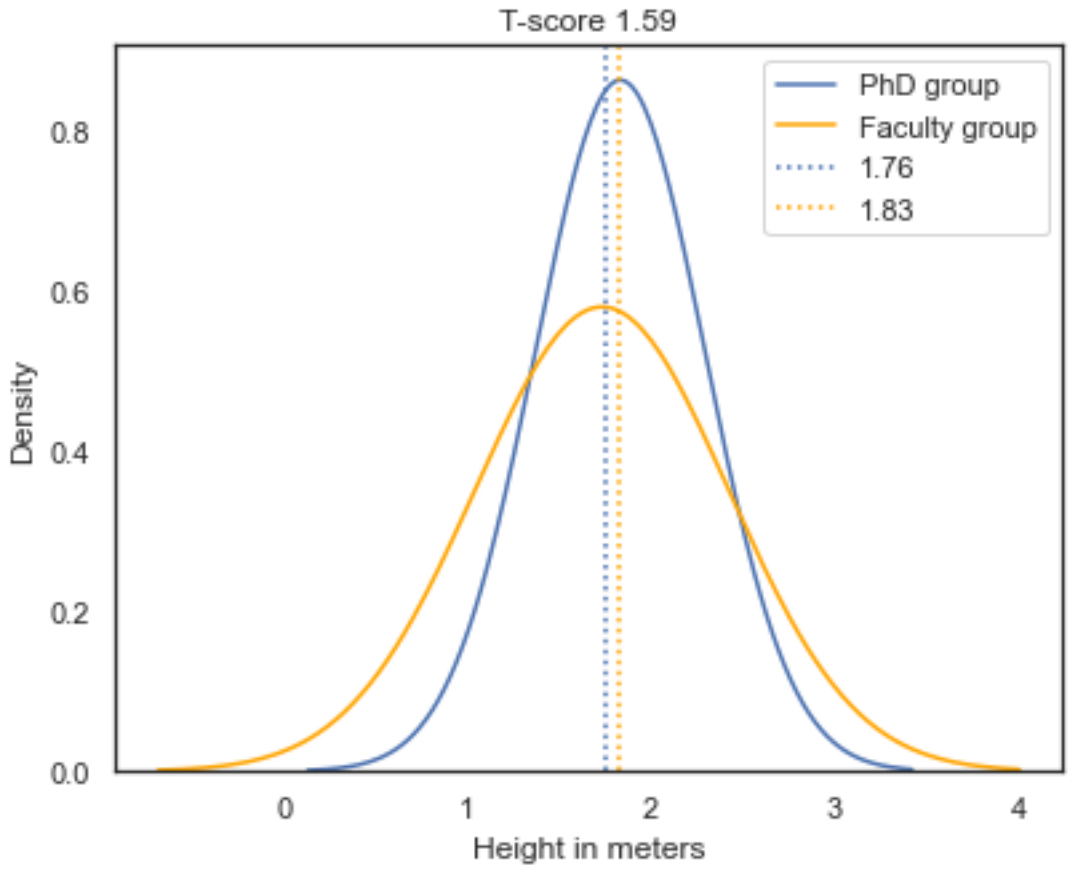
2.2 Normal vs t-distribution

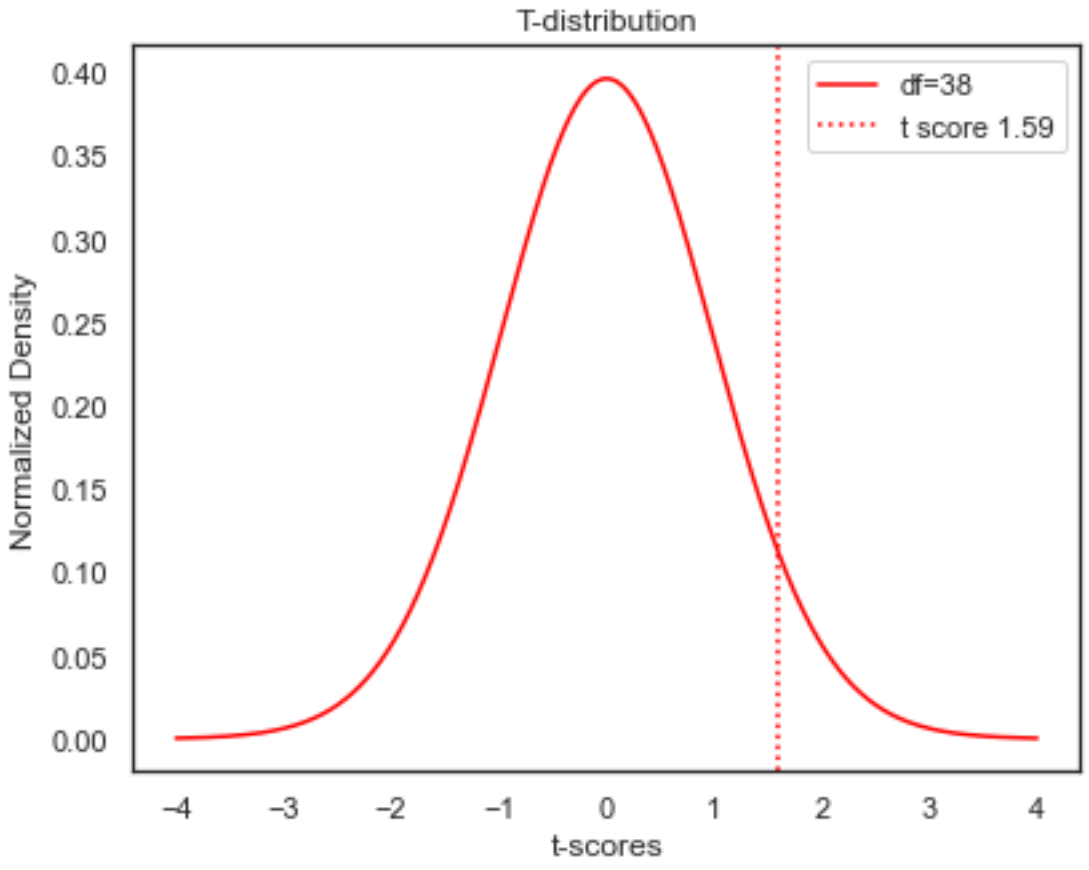
Point a)



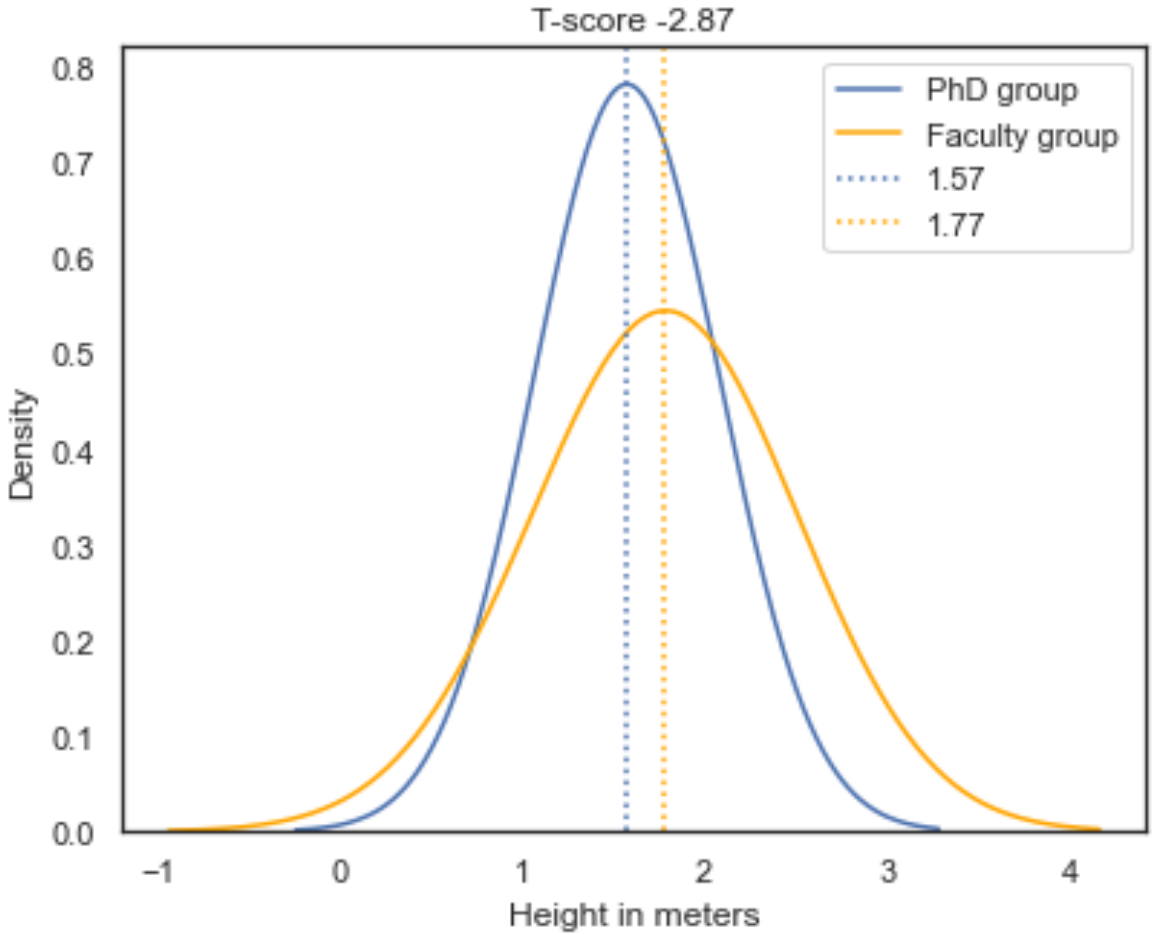
The t-distribution is heavier in the tail, but as the size sample increases (~30) it gets closer to a normal distribution.

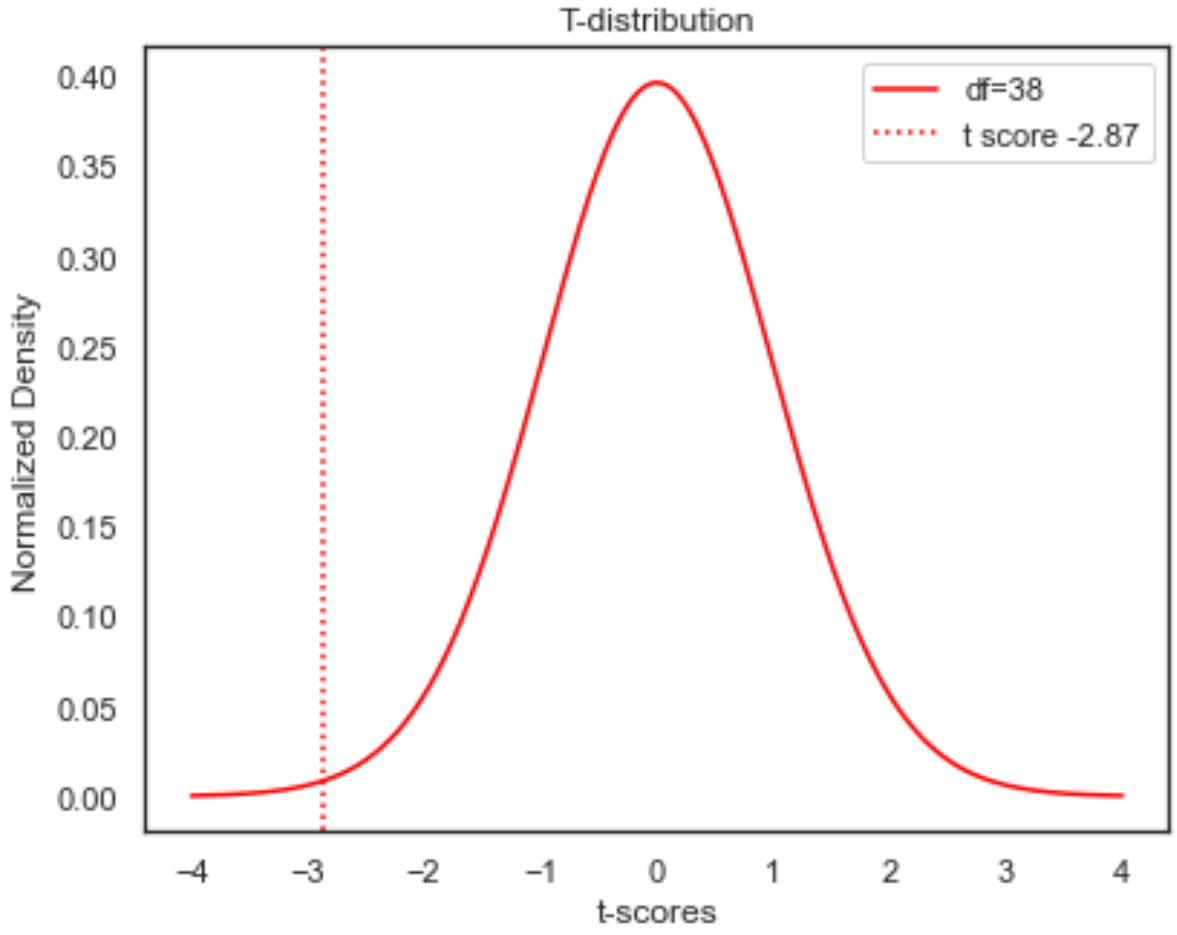
2.3 Two-sample mean test statistic



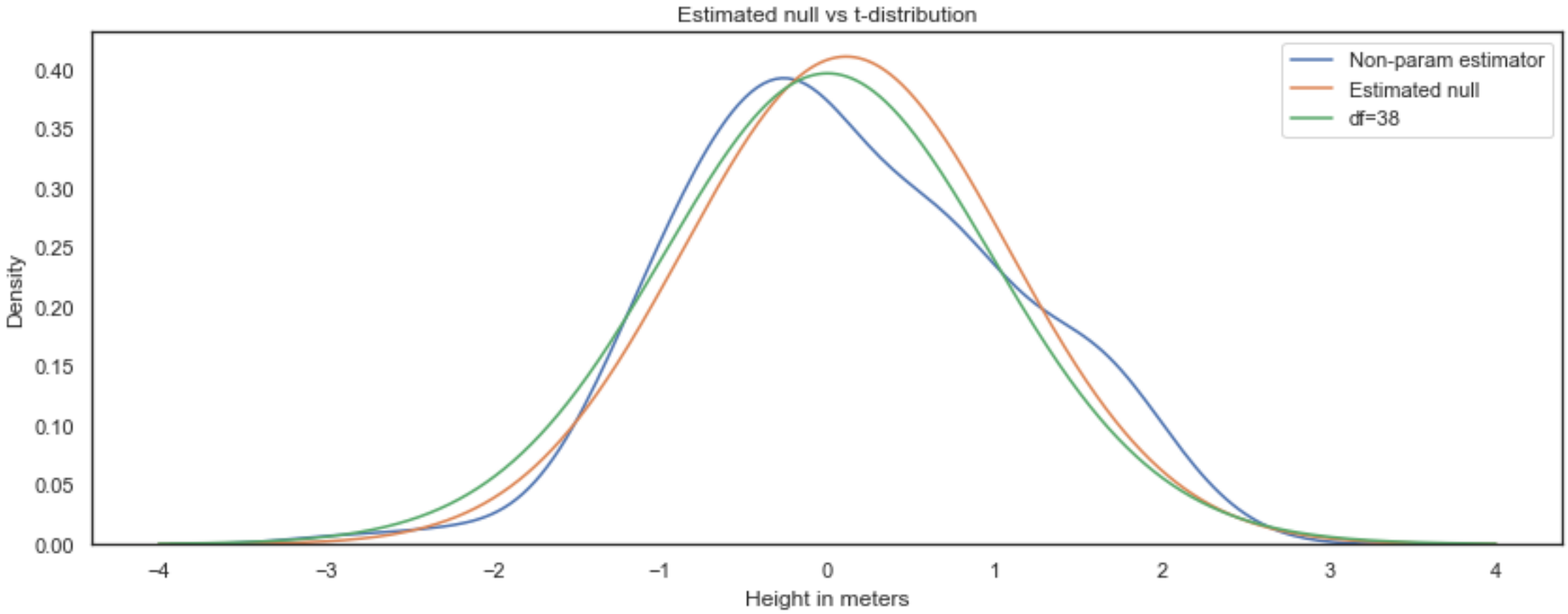


2.3 Two-sample mean test statistic





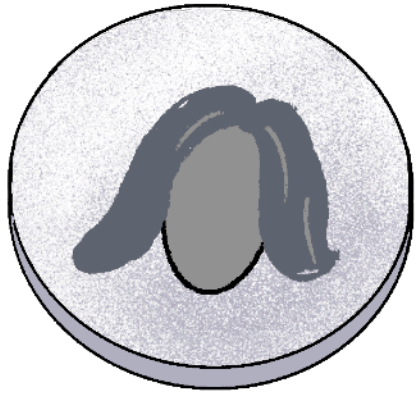
2.3 Two-sample mean test statistic



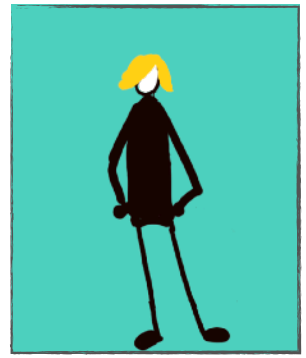
The p-value, a measure of suprize

The p-value, a measure of suprize

How likely* it is that flipping one coin will produce head?

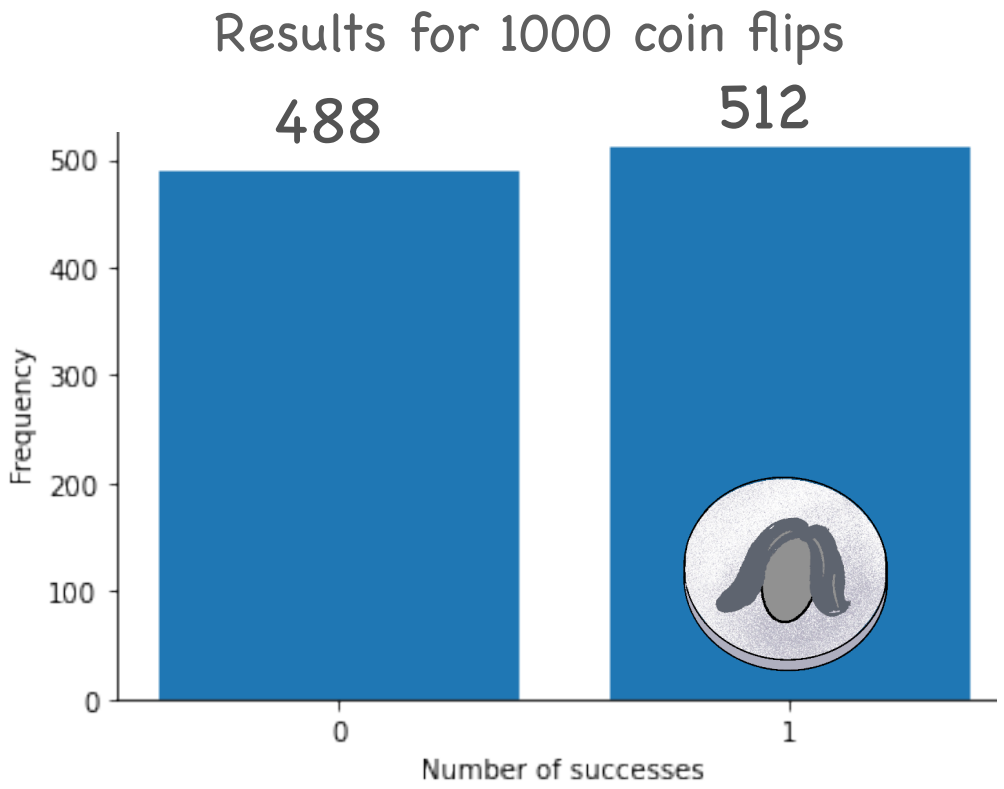


*in a universe where Ileana does not have magic powers

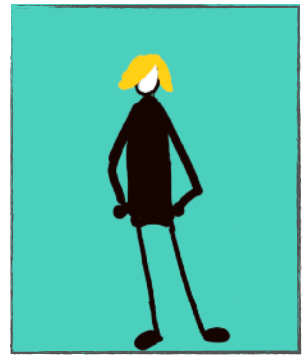


The p-value, a measure of suprize

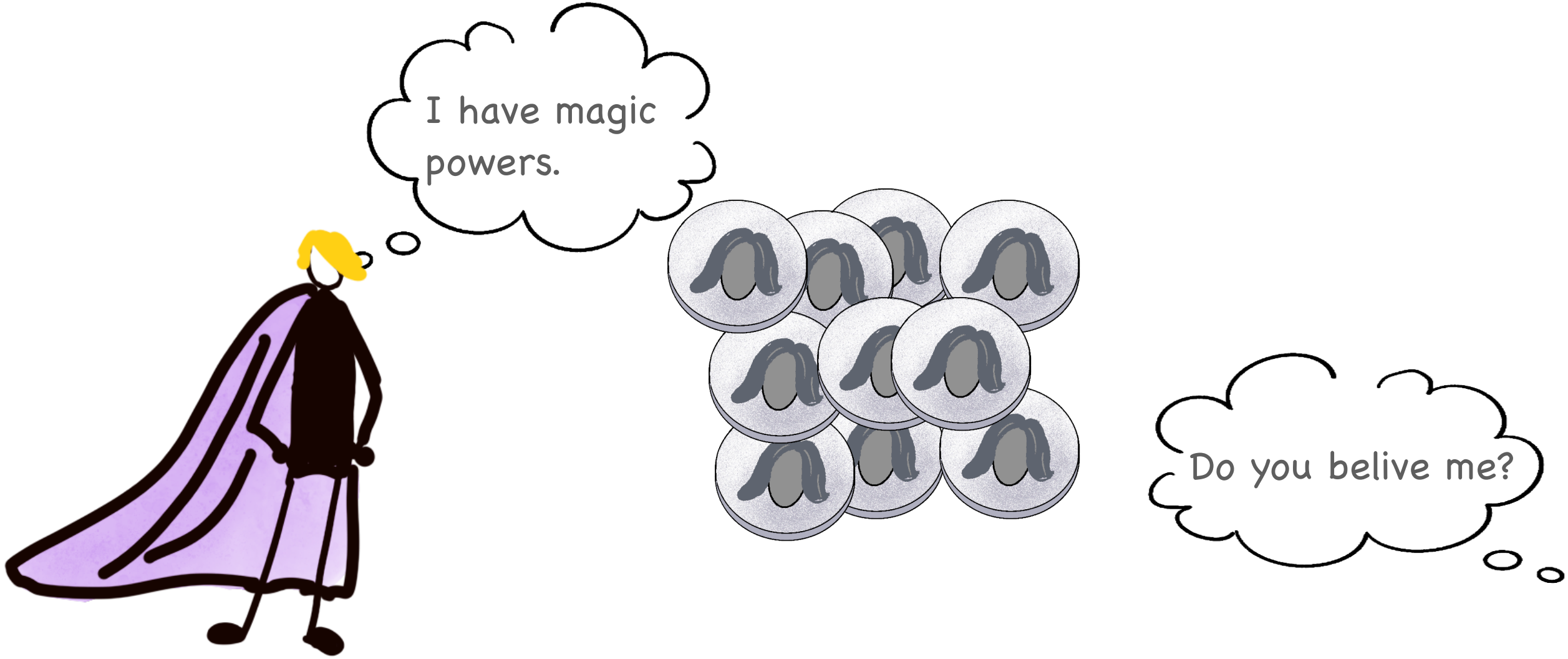
How likely* it is that flipping one coin will produce head?



*in a universe where Ileana does not have magic powers

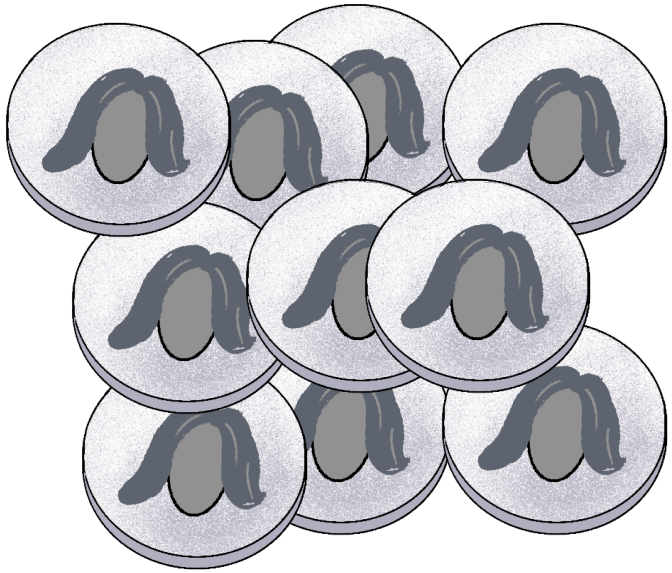


The p-value, a measure of suprize

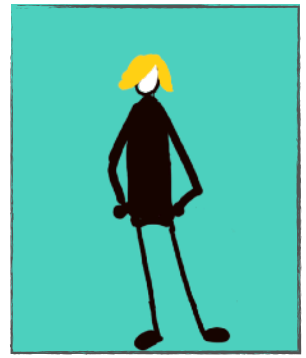


The p-value, a measure of suprize

How likely* it is that flipping ten coins will produce all heads?

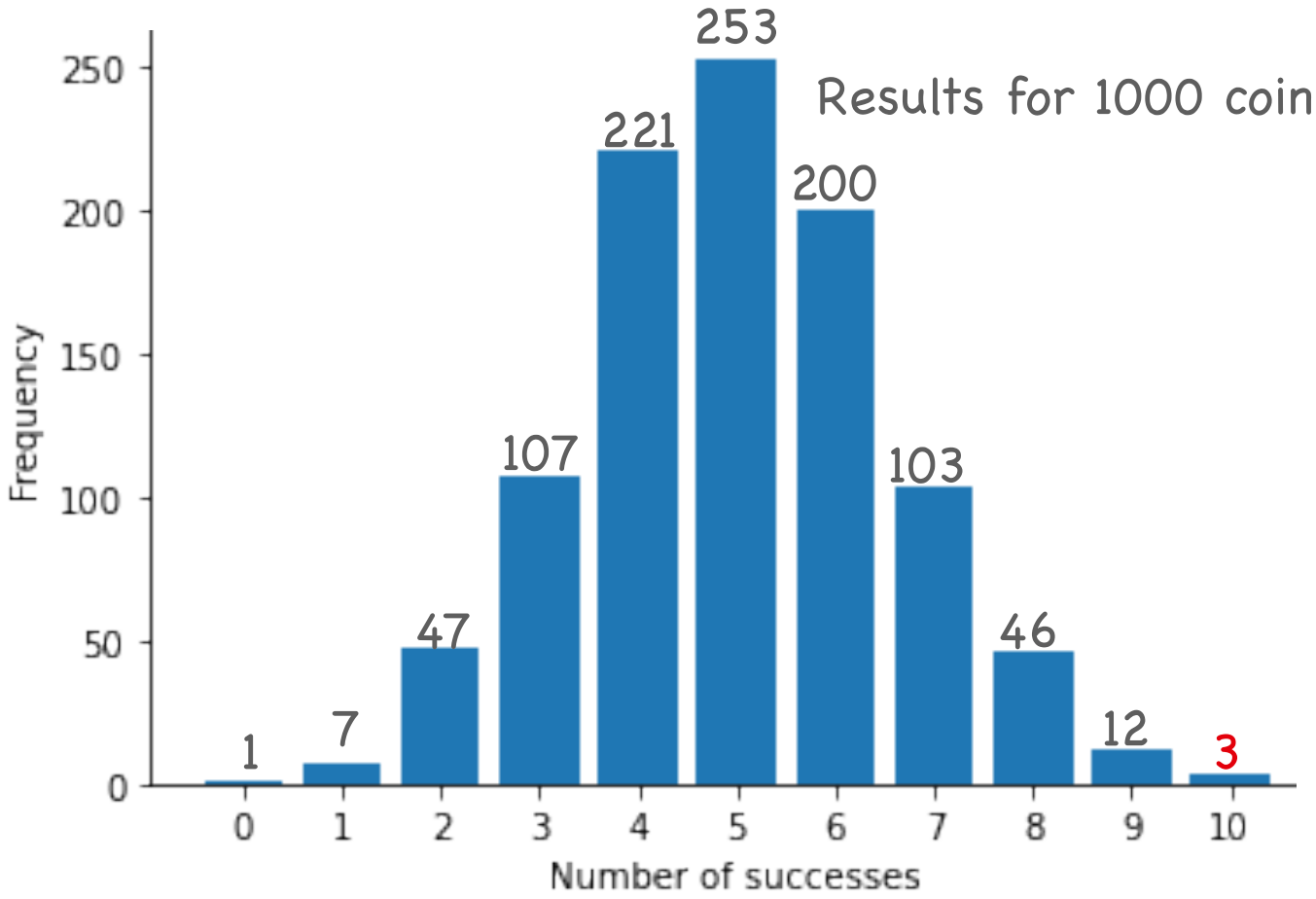


*in a universe where Ileana does not have magic powers

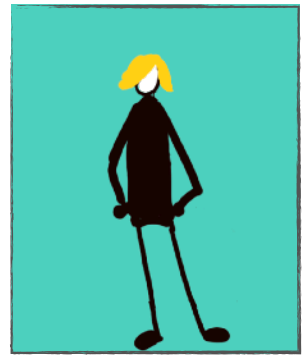


The p-value, a measure of suprize

How likely* it is that flipping ten coins will produce all heads?

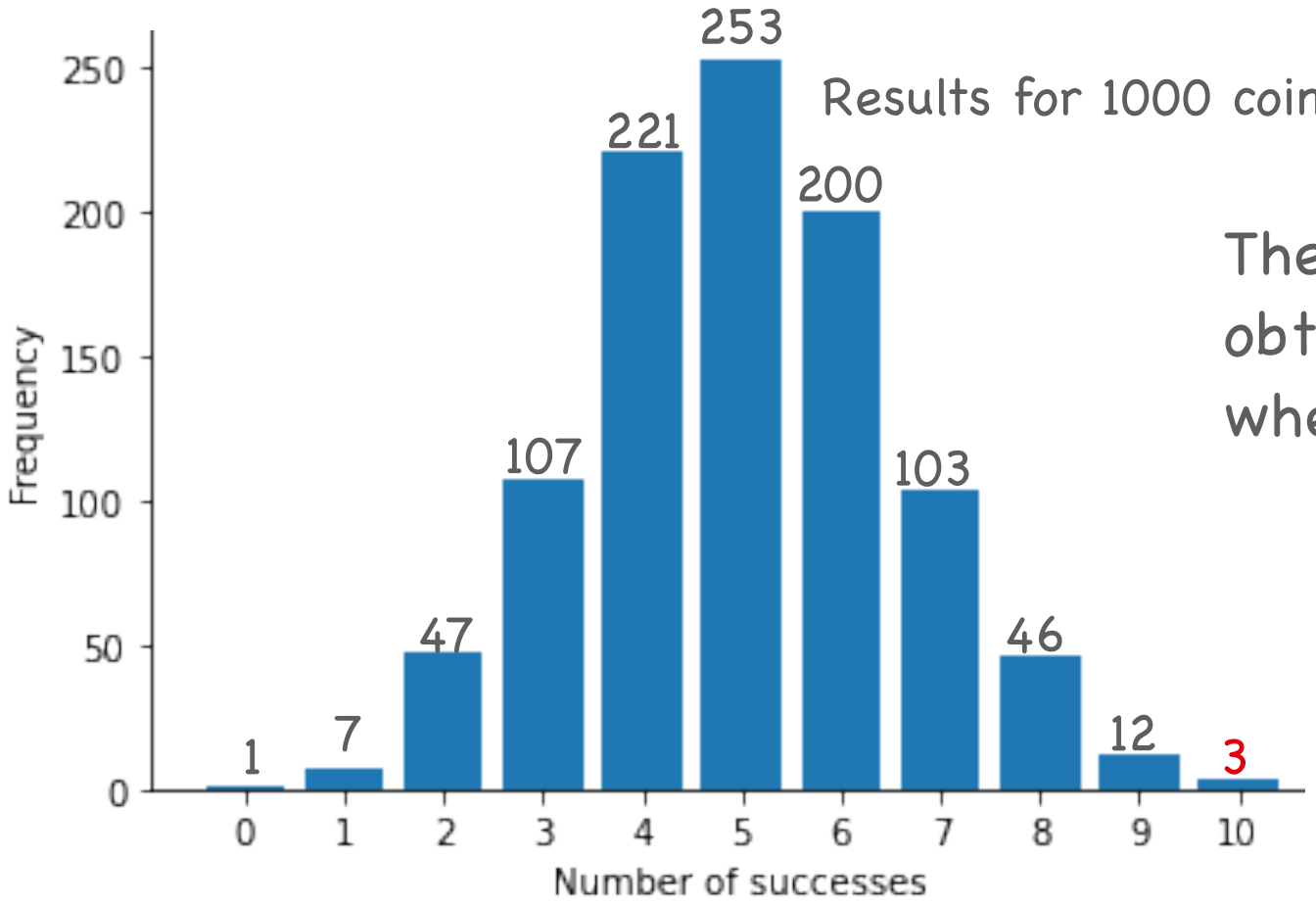


*in a universe where Ileana does not have magic powers



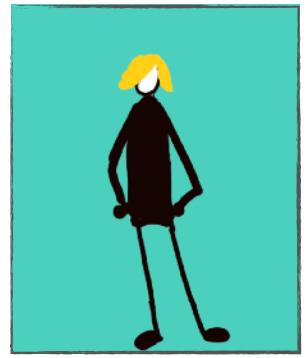
The p-value, a measure of suprize

How likely* it is that flipping ten coins will produce all heads?

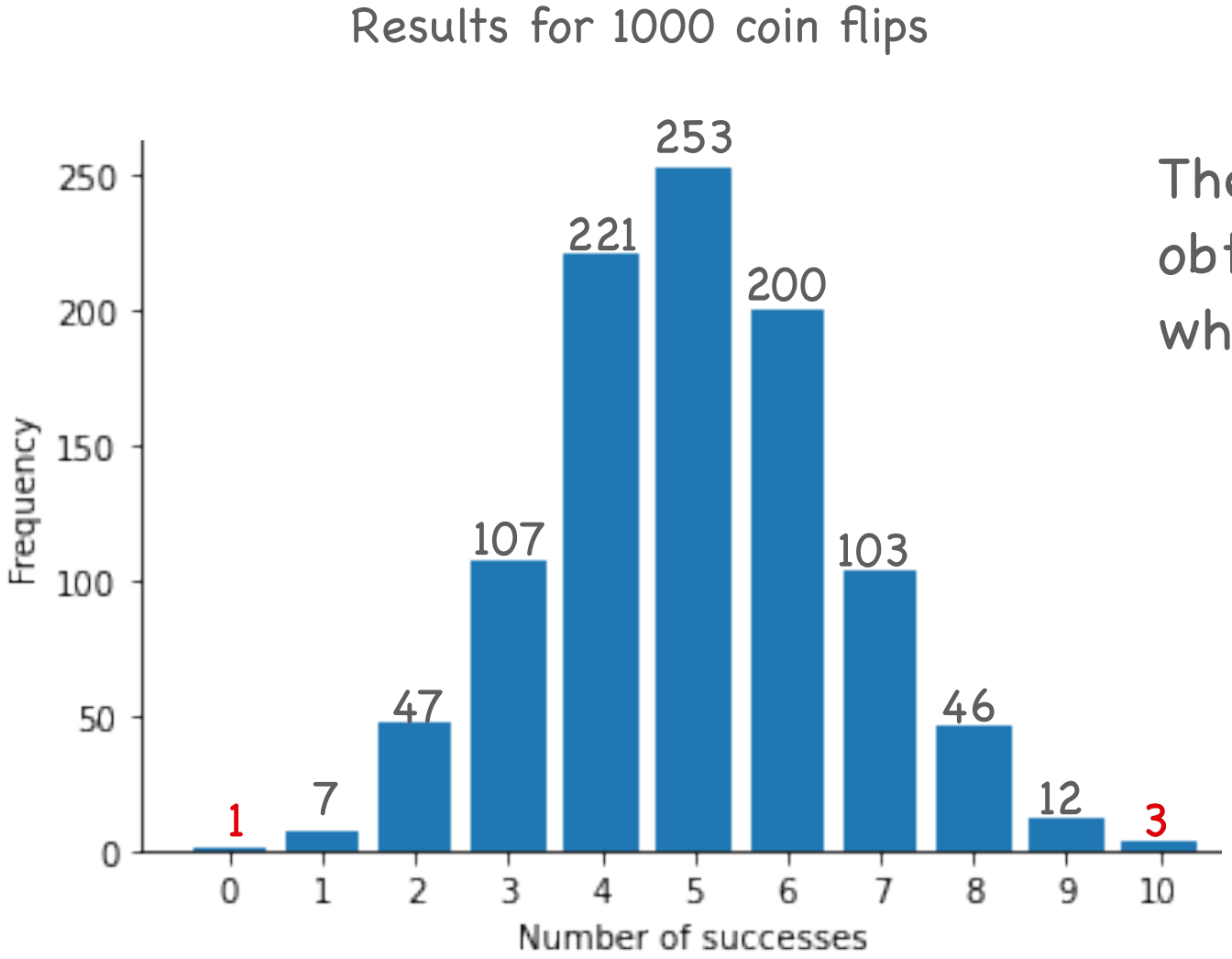


The p-value is the probability of obtaining the data* in universe where the null-hypothesis is true.

*in a universe where Ileana does not have magic powers



The p-value, a measure of suprize

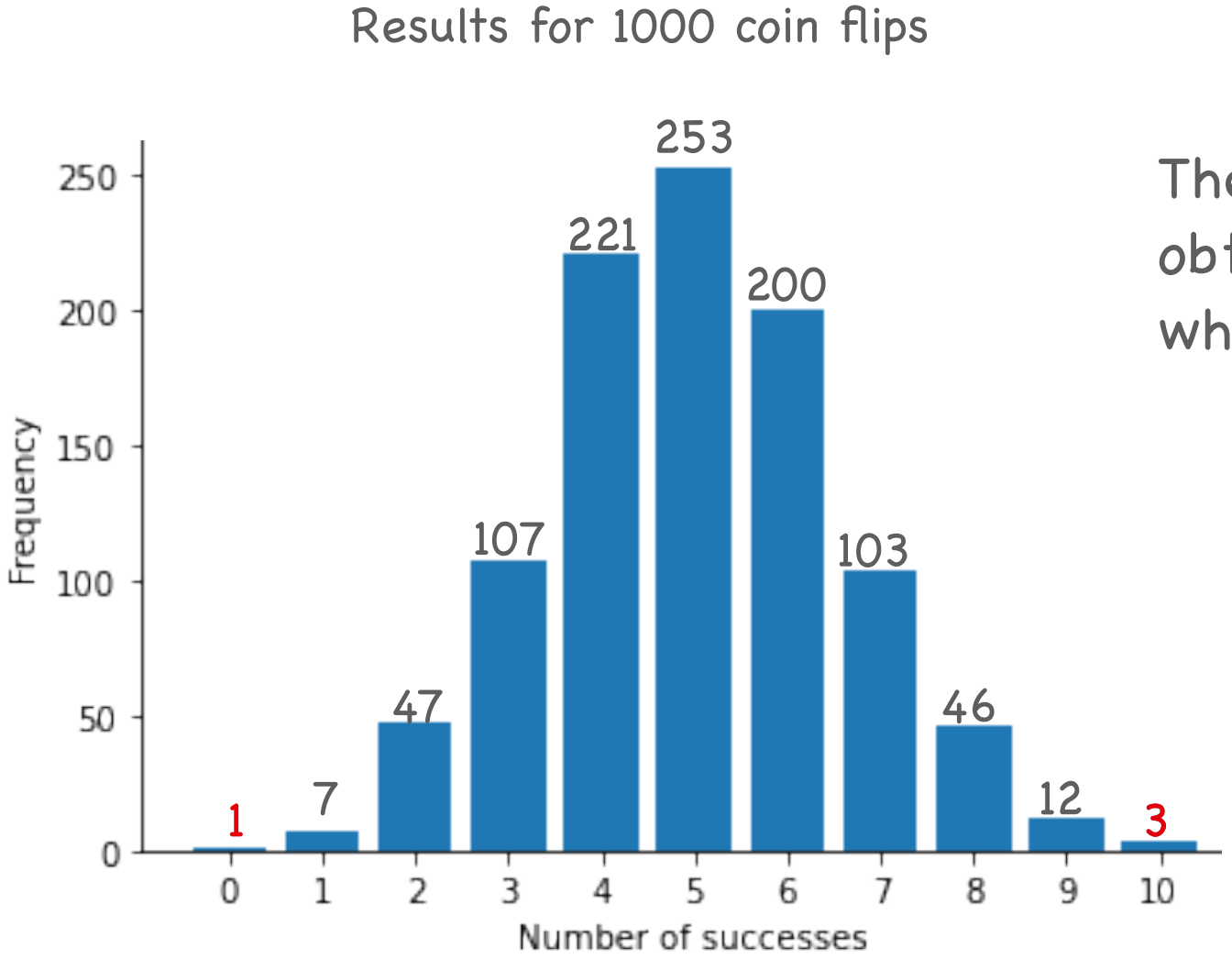


The p-value is the probability of obtaining the data* in universe where the null-hypothesis is true.

$$p = \frac{4}{1000} = 0.004$$

*or data showing as great or greater difference than the null

The p-value, a measure of suprize

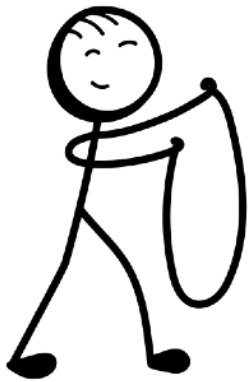


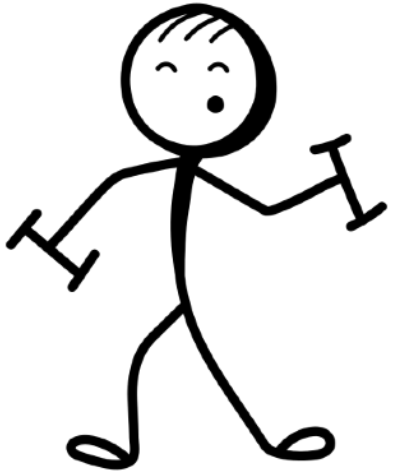
The p-value is the probability of obtaining the data* in universe where the null-hypothesis is true.

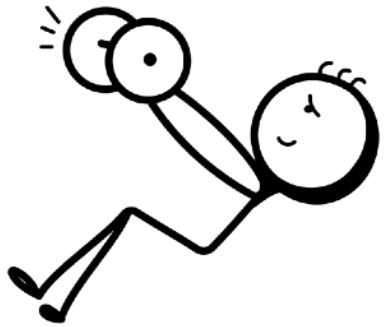
$$p = \frac{4}{1000} = 0.004$$

*or data showing as great or greater difference than the null

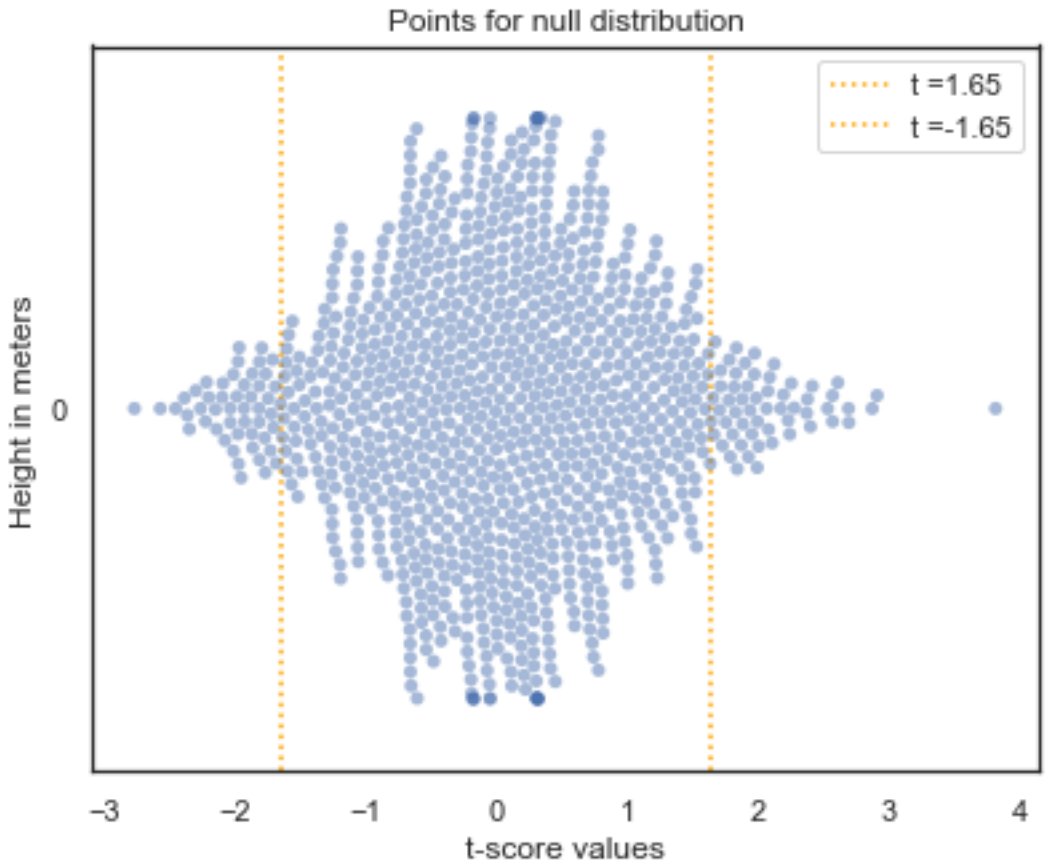
Exercise 3 (p-values)

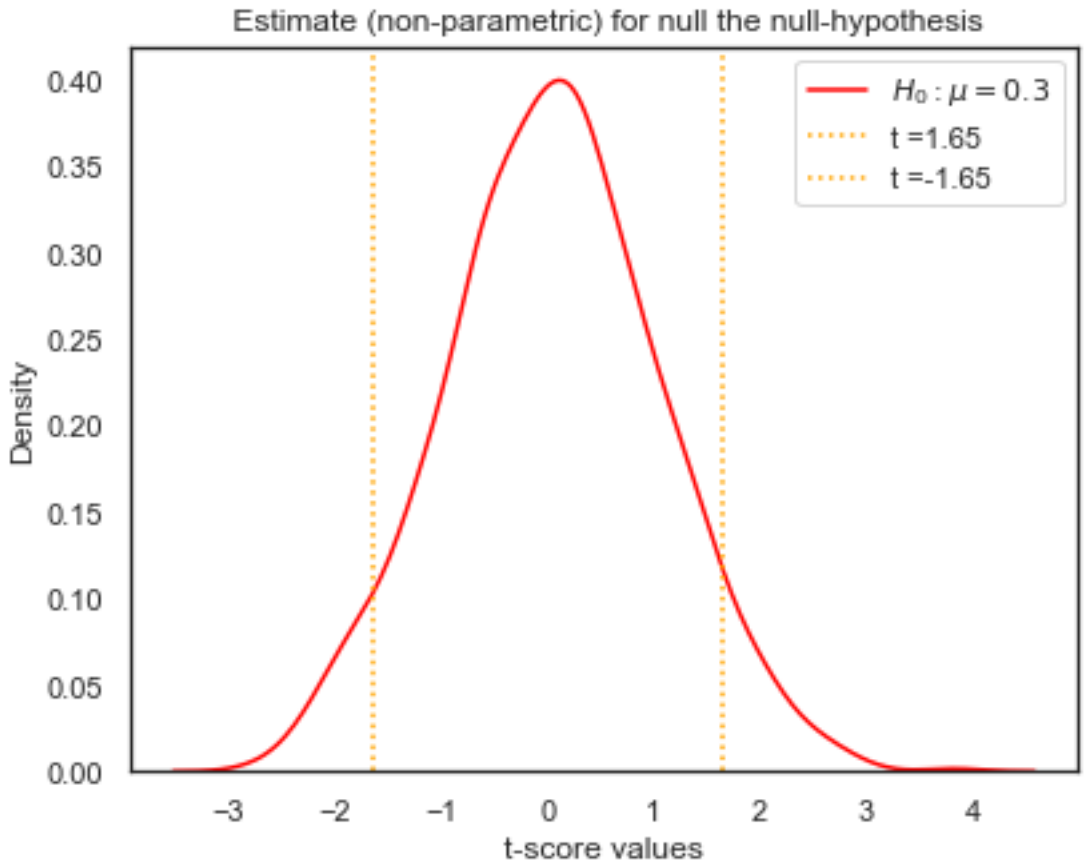






Reporting the results

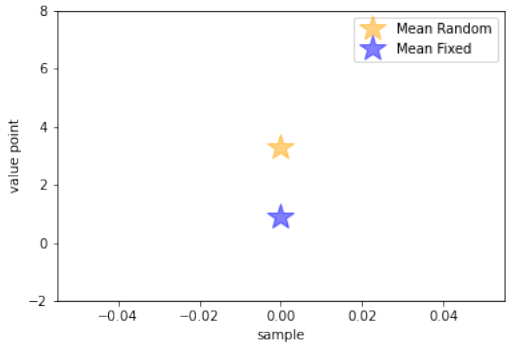


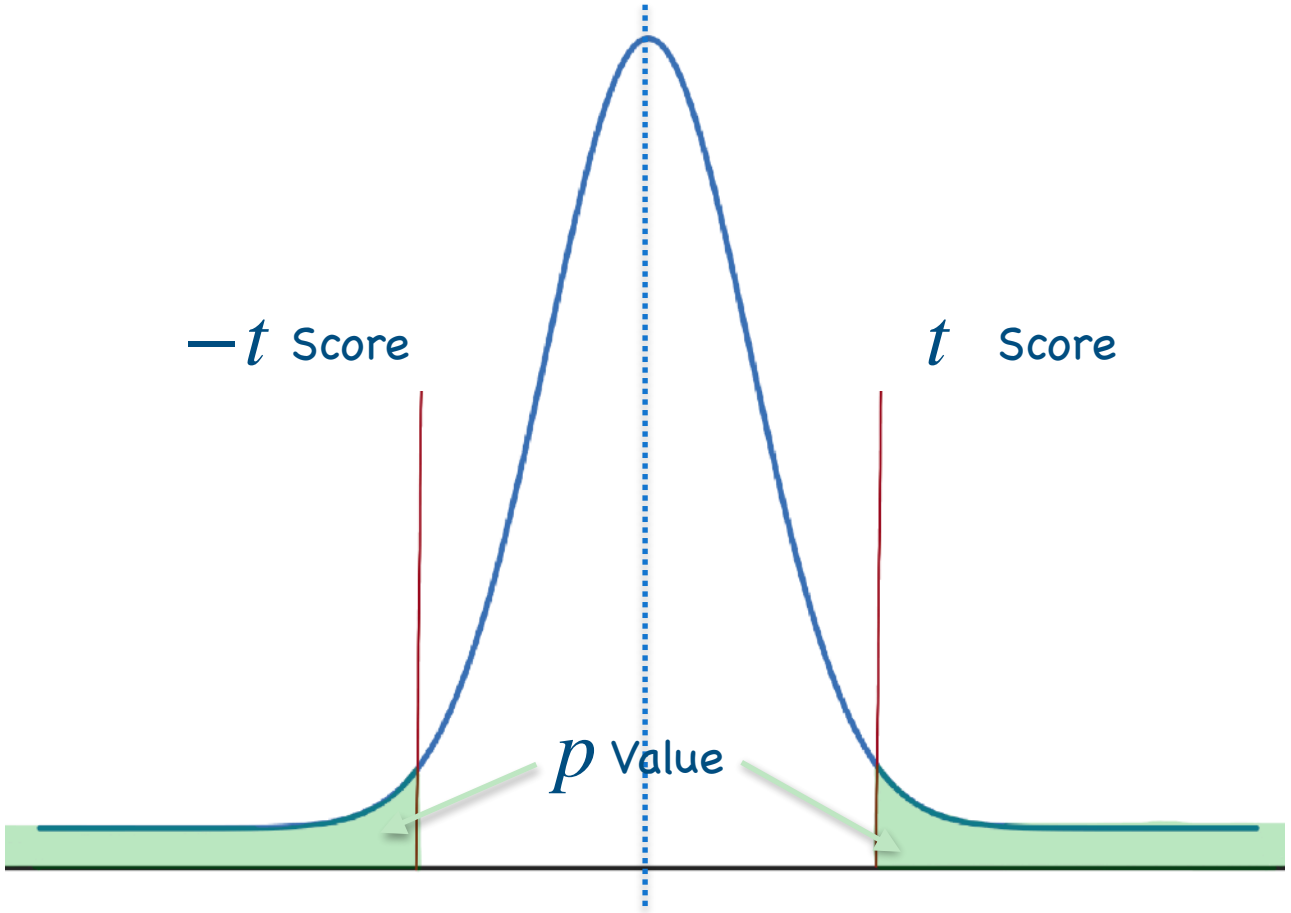


p-value and α in a nutshel

t-score= -2.019

p-value= 0.345

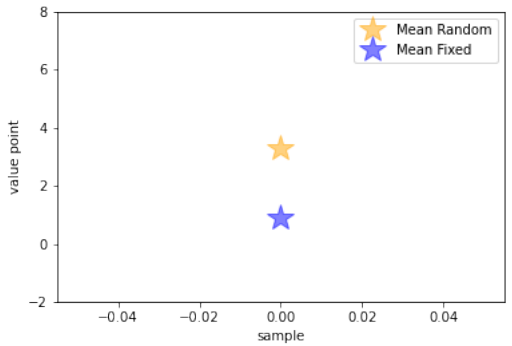




p-value and α in a nutshel

t-score= -2.019

p-value= 0.345



$\mu_r = \mu_f$

P- value answers the question:

If I live in a universe where **H_0 is true**, how surprizing is to measure a t-score of -2.019?

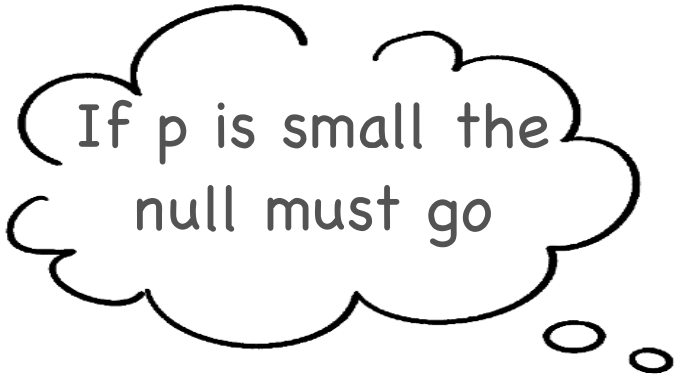
0

Very surprizing

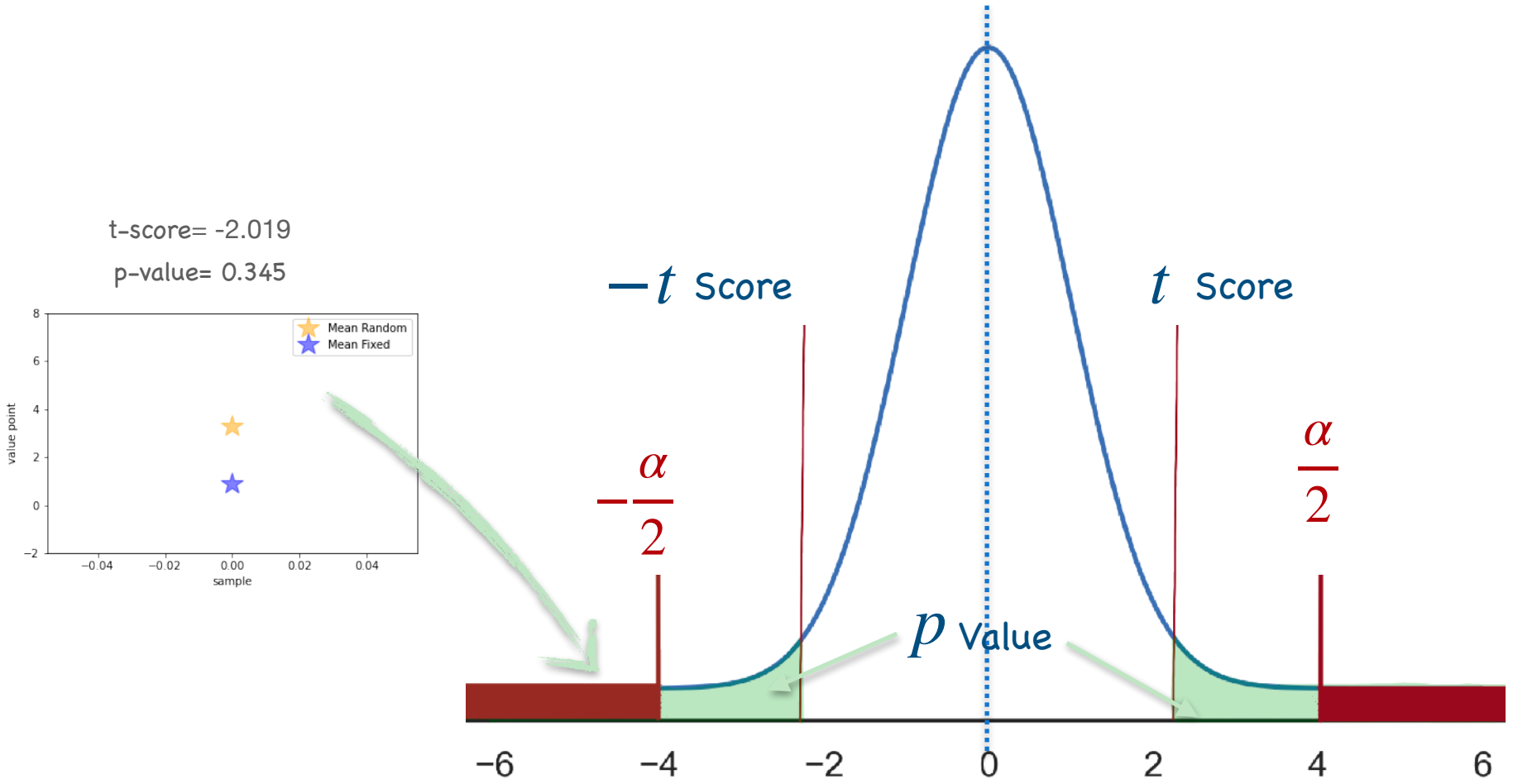
p-value

1

Expected



p-value and α in a nutshel



Reporting the results

When writing up the reuslts you should always include the following information:

- the value of the test statistic
- The sample size
- The p-value

Finally, do your results provide evidence aagingt the null hypothesis?

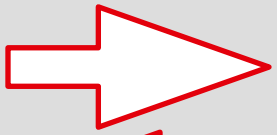
"We reject/fail to reject the null-hypothesis at a significance level of α

A note on p-values

The p-value indicates strength of evidence against H_0. Its the probability of getting the observed results or more extrem results IF H_0 is true.

Red flag: the p-value is NOT the probability of H_0 being true.

Inverse probability fallacy:

(A: it rains)  (B: I take my umbrella)

(A: it rains)  (B: I take my umbrella)

Why use p-values?

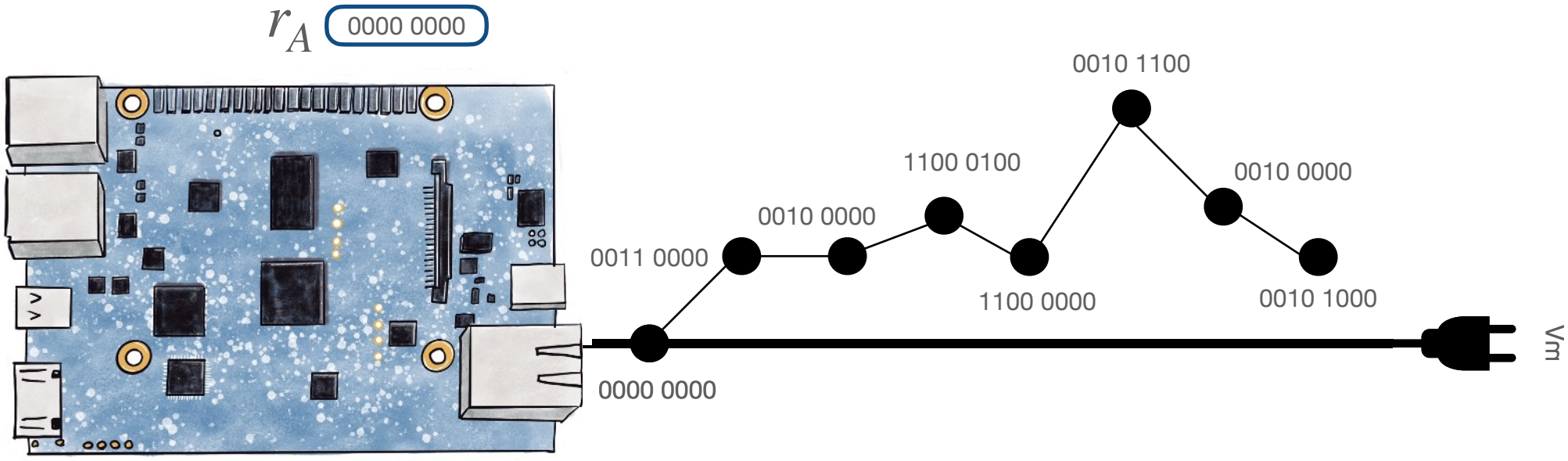
If its so hard to interpret p-values, why do we use the? Testing means that at each step, we critically assess our data. A high p-value means that maybe there is no pattern in the data, even if it seems to us to be.

Confirmation bias:

After buying a red car, I tend to notice them everywhere.

TVLA on real traces

What is leakage?



Leakage is the dependence between power consumption and the sensitive data.

Leakage detection in action

You are a developer who wants to ensure that your implementation does not leak, but not perform full attack. Ideas?

- Hint 1: we know that any dependency between the measured side-channel and the sensitive data is a potential side channel vulnerability;
- Hint 2: using the reverse logic, if there are no dependencies, there is no side-channel vulnerability;

Can we check vulnerability to side-channels without doing an attack?

- yes! measure the side channel for di erent input values and see if they are different;
- complicating fact: side channel measurements are influenced by many factors, not always straightforward;

Leakage detection in action

Test Vector Leakage Detection (TVLA) most popular leakage detection test.

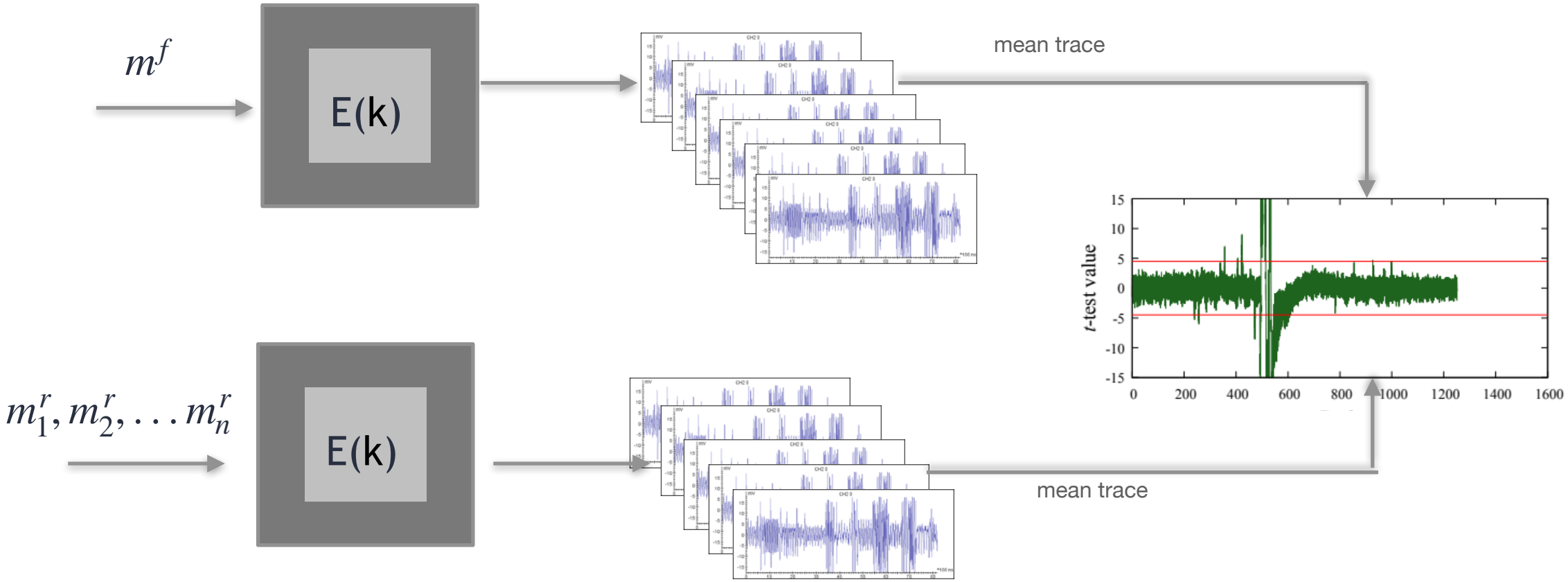
- Non-specific or general test: aims to detect any leakage that depends on input data (or key);

a.k.a fixed - vs - random;

- Specific-test: targets a specific intermediate value of the cryptographic algorithm that could be exploited to recover keys or other sensitive information.

a.k.a fixed - vs - fixed;

Collecting data for TVLA

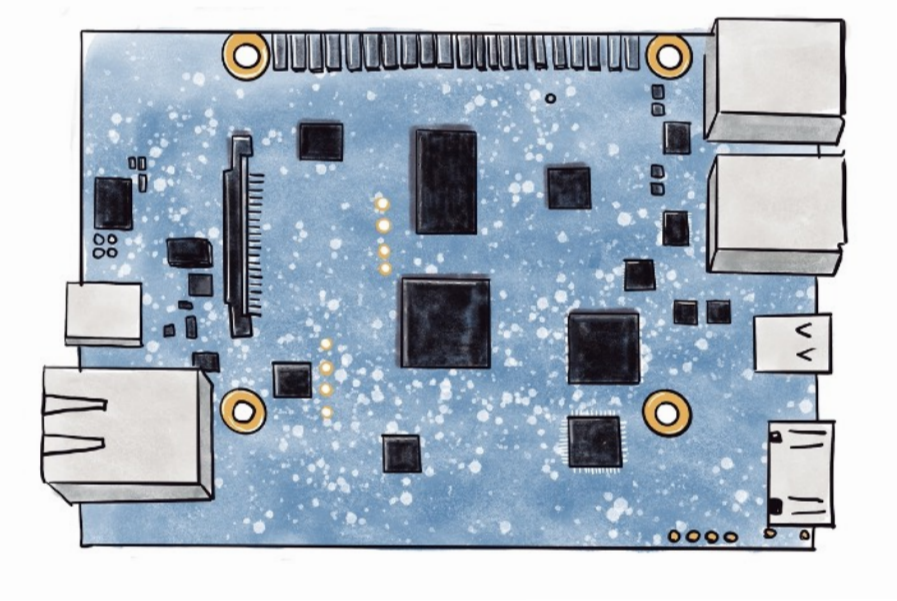


Suggested reading:

A testing methodology for sidechannel resistance validation Gilbert Goodwill, Benjamin Jun, Josh Jaffe, Pankaj Rohatgi: Cryptography Research Inc.

https://csrc.nist.gov/csrc/media/events/non-invasive-attack-testing-workshop/documents/08_goodwill.pdf

TVLA - two-sample t-test



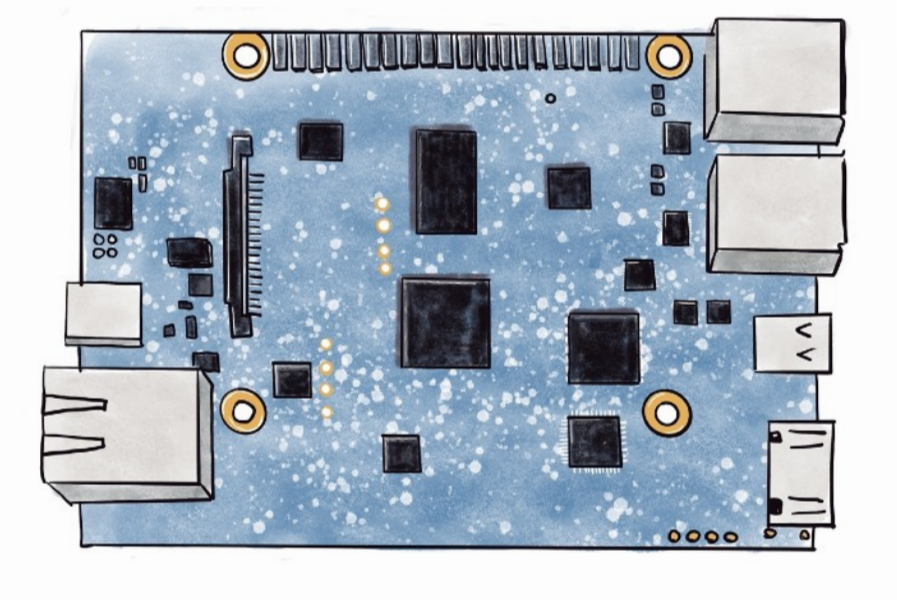
H_0: device is NOT guilty of leaking information

$$\mu_{fixed} = \mu_{random}$$

H_a: device is guilty of leaking information

$$\mu_{fixed} \neq \mu_{random}$$

Selecting the significance level

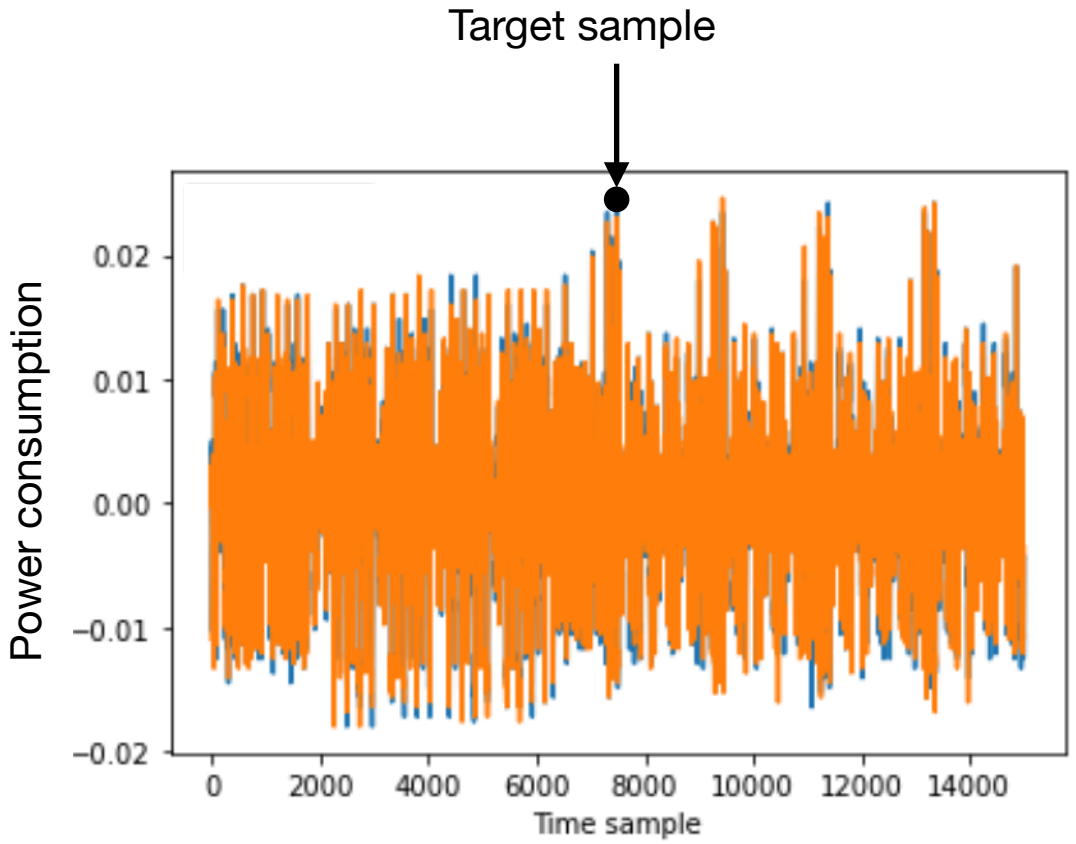


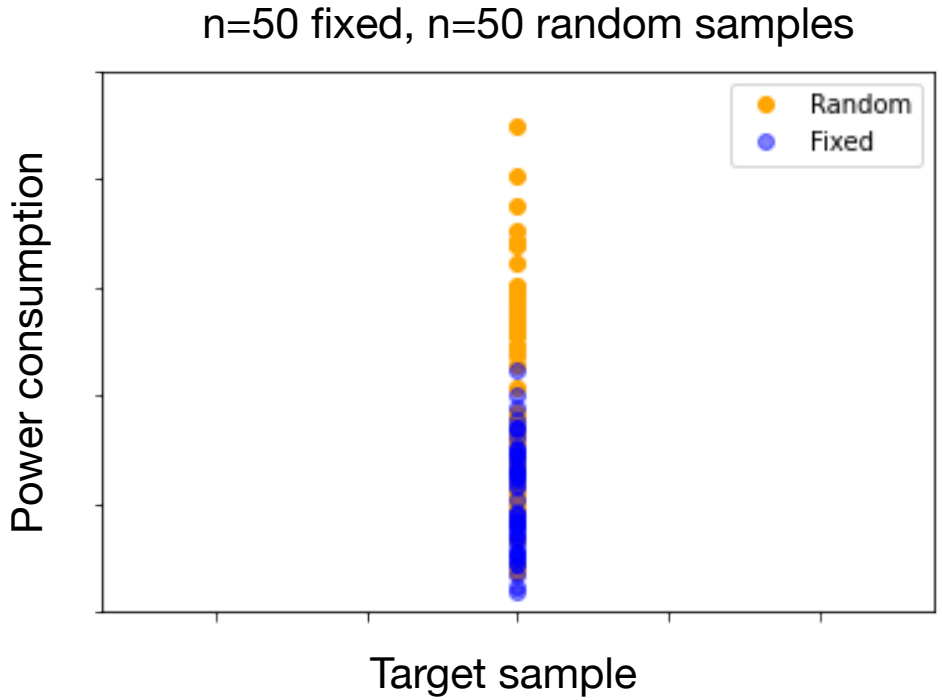
H_0: device is NOT guilty of leaking information

$$\mu_{fixed} = \mu_{random}$$

The significance level (α) is the probability at which we are prepared to reject the null hypothesis and conclude that the effect is statistically significant.

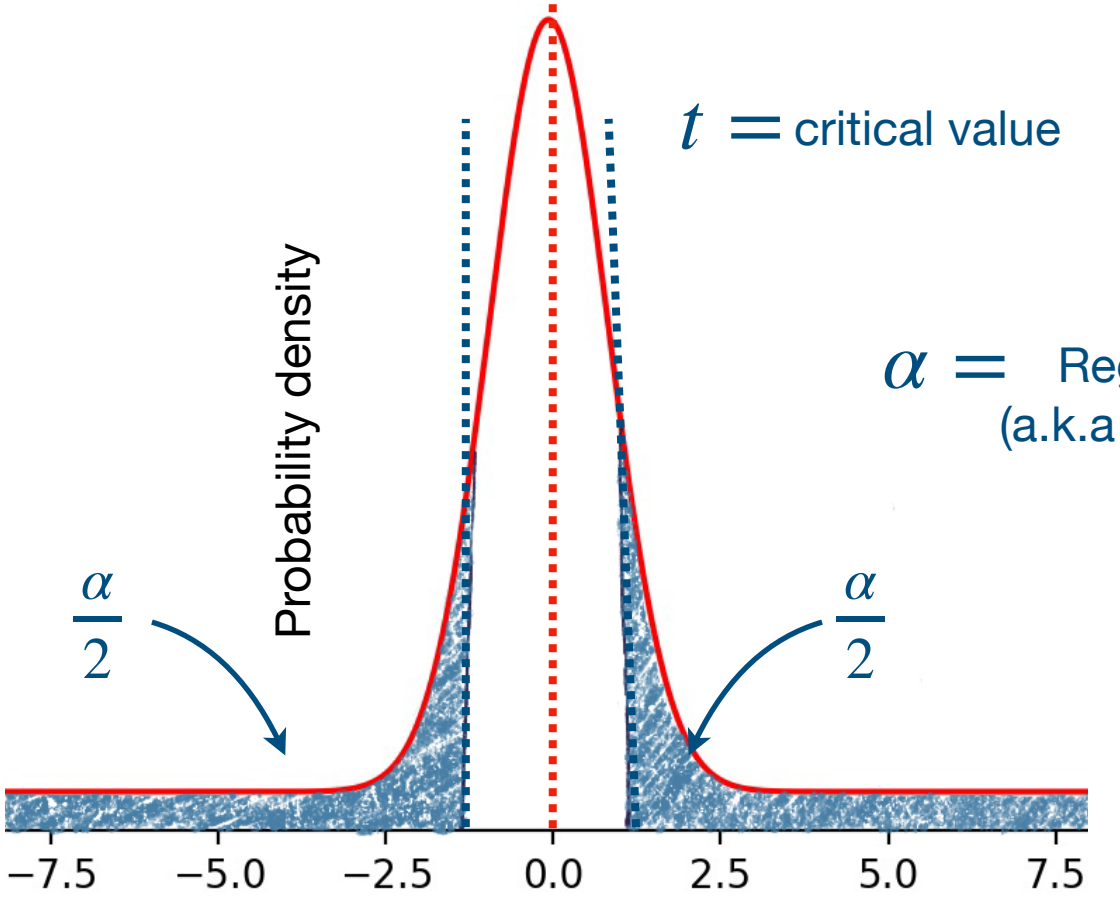
TVLA - two-sample t-test





First order t-test, we analyze each sample independently

Why the 4.5 value?



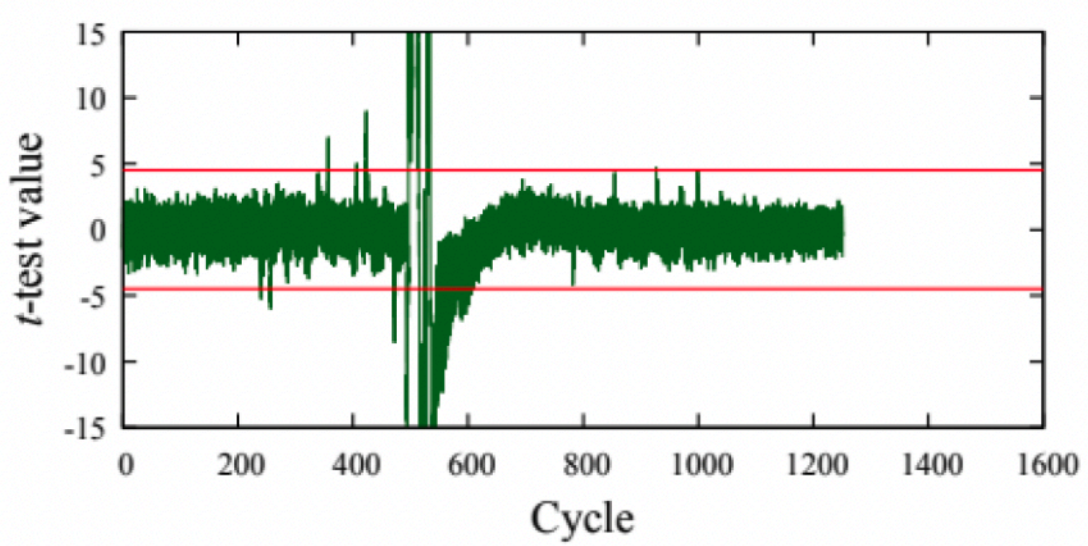
t = critical value

α = Region of rejection
(a.k.a significance level)

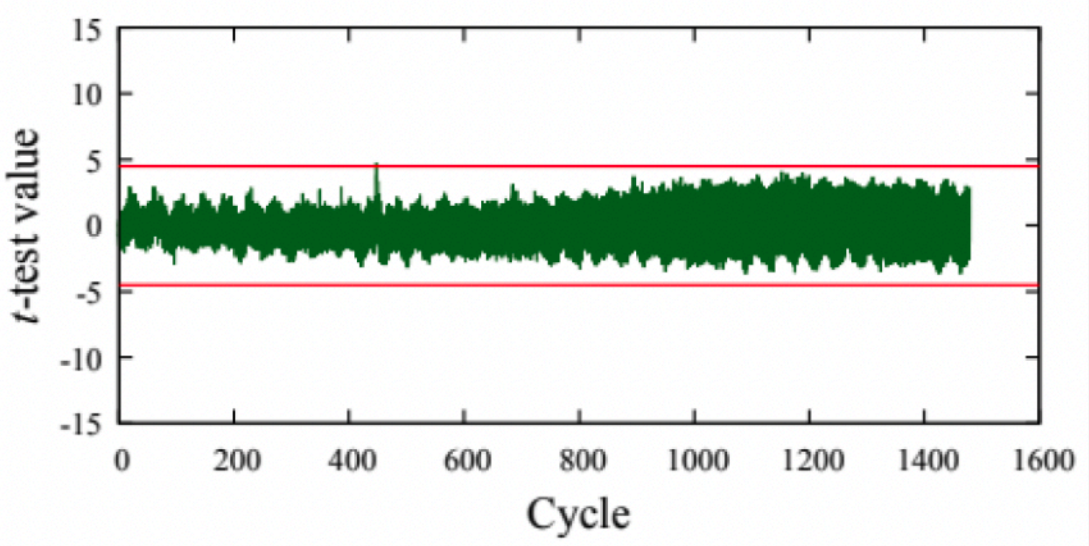
— T-distribution for $\mu_r = \mu_f$

$\mu_F - \mu_R$

What is the magic number 4.5?



(a) AES original implementation.

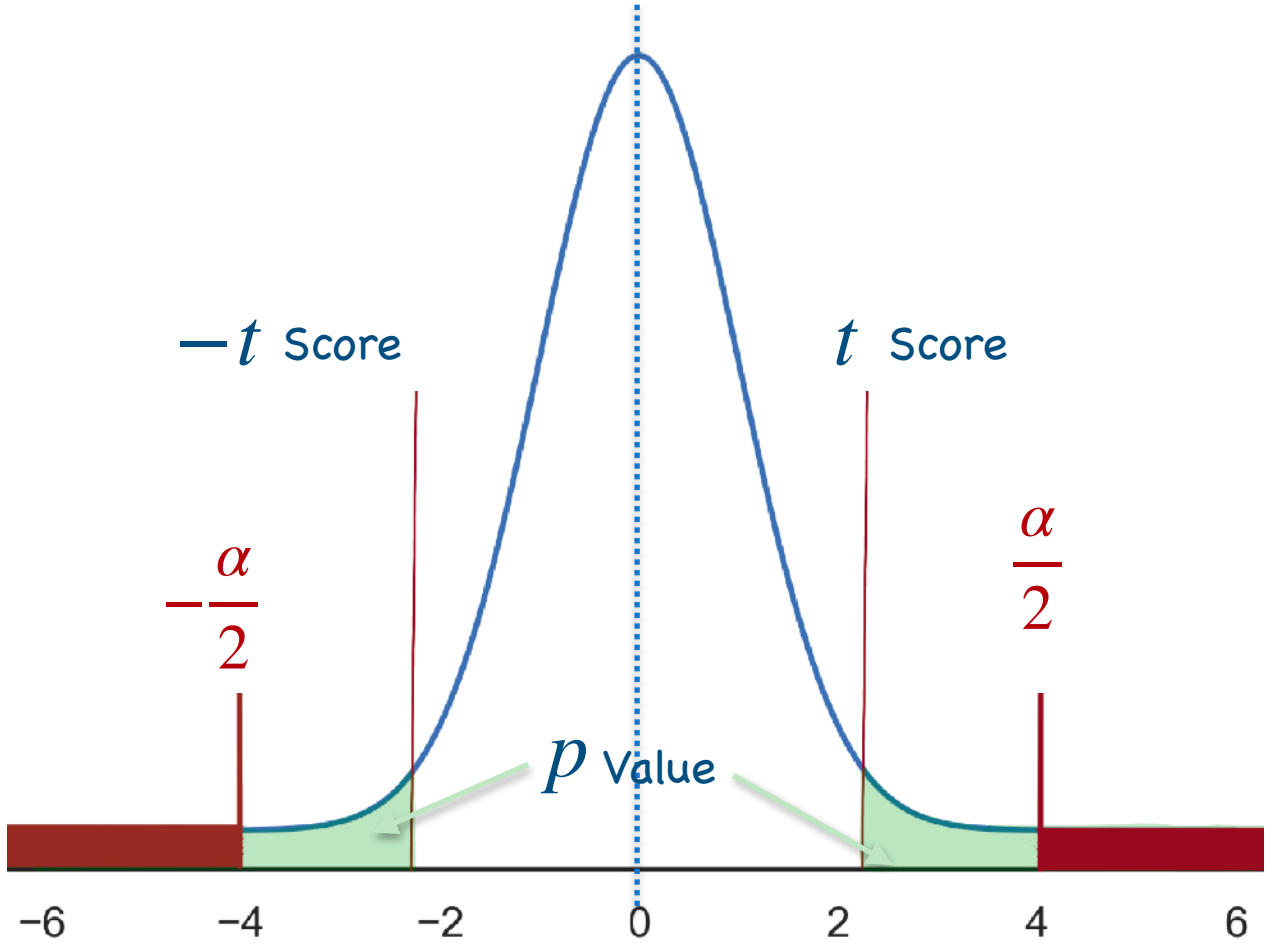


(d) AES fixed with ROSITA.

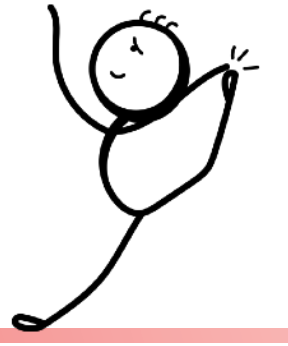
Source for the figure: Madura A Shelton and Niels Samwel and Lejla Batina and Francesco Regazzoni and Markus Wagner and Yuval Yarom *Rosita: Towards Automatic Elimination of Power-Analysis Leakage in Ciphers*, NDSS 2021

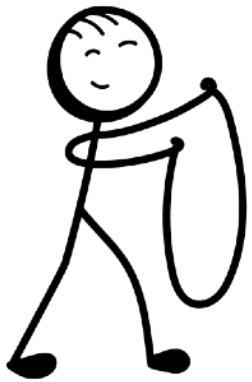
What is the magic number 4.5?

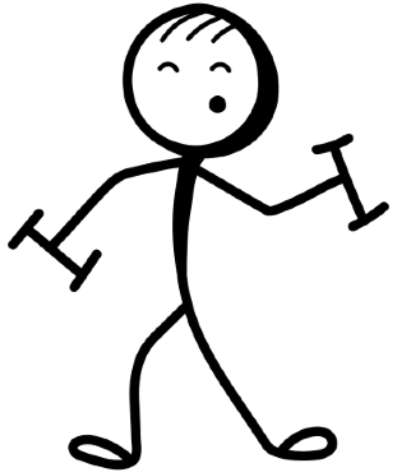
$\alpha = 0.001$

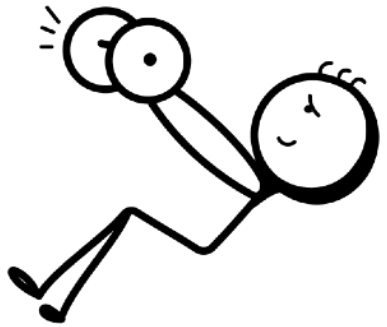


Exercise 4

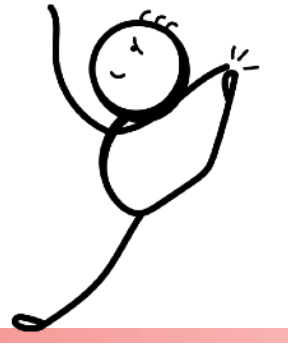


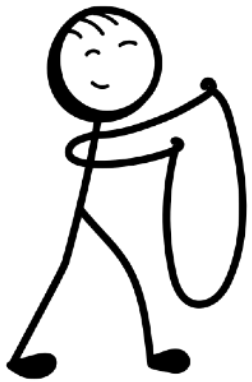


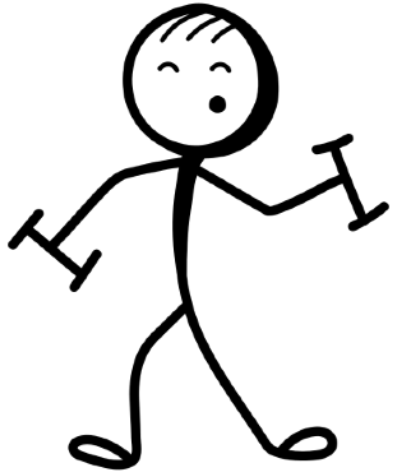


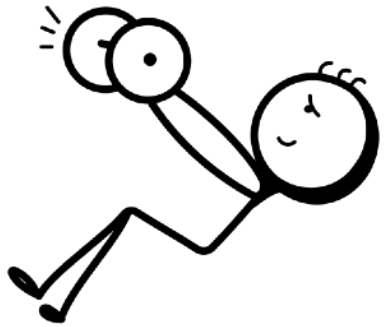


Exercise 5









Notes

IMPORTANT: TVLA test is qualitative measure of leakage, and NOT a quantitative measure.

If we are dealing with a high-order implementation, we always need to check if lower orders leak, there might be surprizes;

Final Notes

A lack of evidence to support the guilty verdict, does not mean the device is "innocent"; We say: "We fail to reject H0" and NOT "we accept H0"

Alternatively we say:

"The evidence supports the decision to reject H0 at significance level α".

Final Notes

A lack of evidence to support the guilty verdict, does not mean the device is "innocent"; We say: "We fail to reject H0" and NOT "we accept H0"

Alternatively we say:

"The evidence supports the decision to reject H0 at significance level α".

Why could TVLA to fail?

- Sample size - too small
- Effect size (the difference between the two means) is too small, because:
  - wrong fixed input;
  - too much noise (variance) in the sample data;
- Bad luck: statistical tests are probabilistic

Recommended reading

Carolyn Whitnall, Elisabeth Oswald:
A Critical Analysis of ISO 17825 ('Testing Methods for the Mitigation of Non-invasive Attack Classes Against Cryptographic Modules'). ASIACRYPT (3) 2019: 256-284

François-Xavier Standaert:
How (Not) to Use Welch's T-Test in Side-Channel Security Evaluations. CARDIS 2018: 65-79

Tobias Schneider, Amir Moradi:
Leakage Assessment Methodology - A Clear Roadmap for Side-Channel Evaluations. CHES 2015: 495-513

http://reassure.eu/leakage-detection-tutorial/

THANK YOU