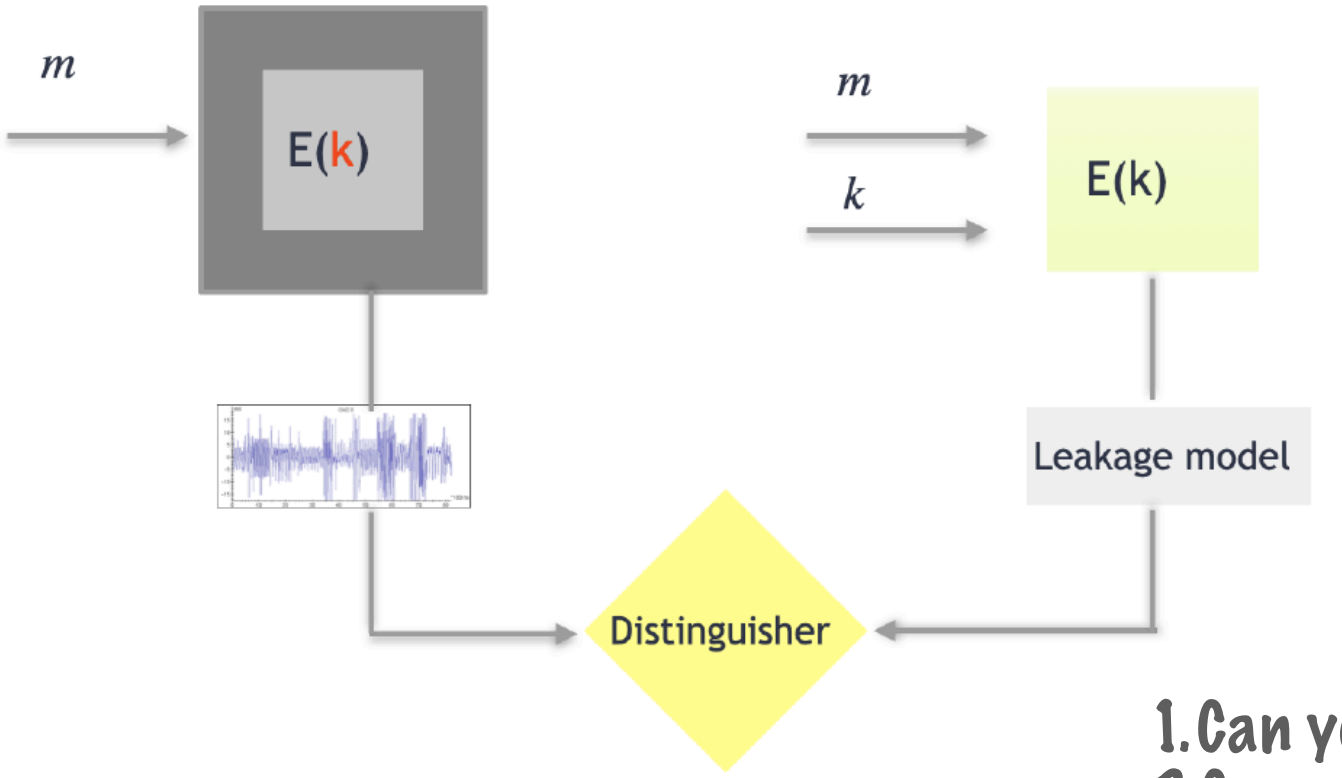
DPA Success Metrics and Leakage assessment

Ileana Buhan, March 2024
@ileanabuhan

Radboud University

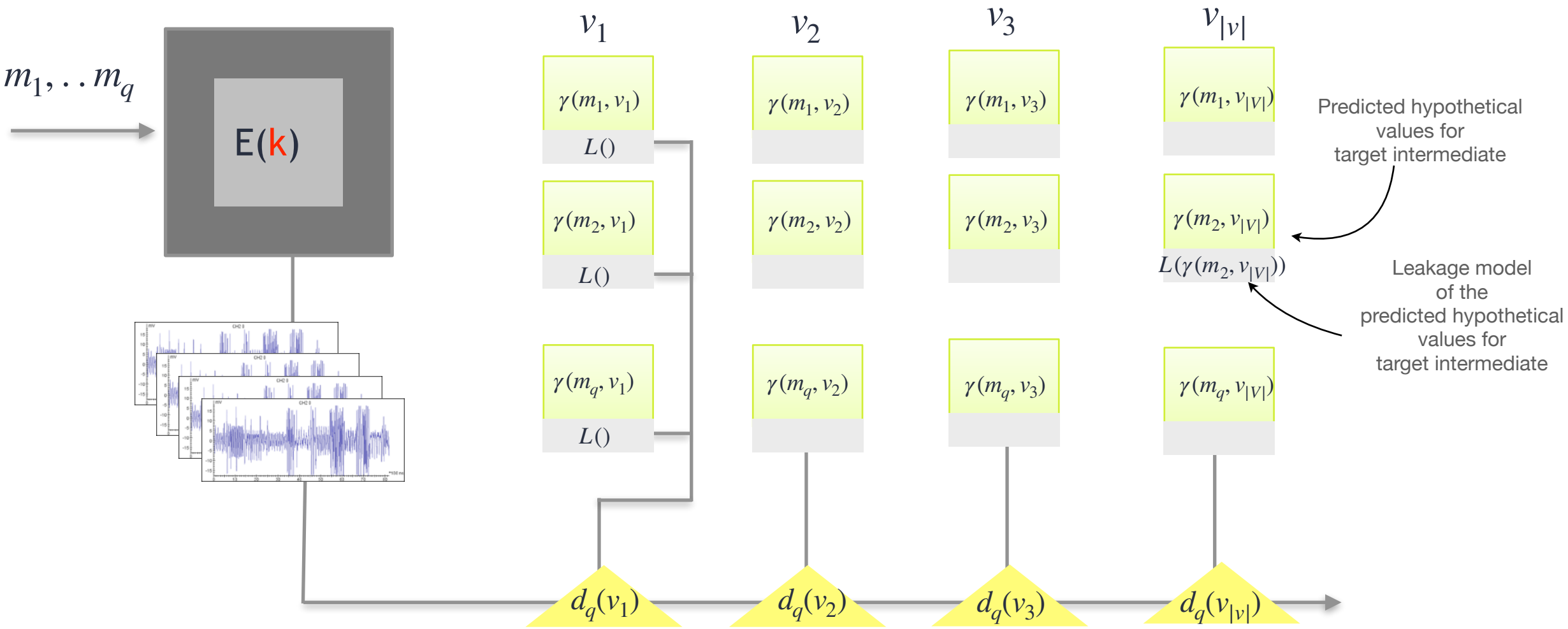
Recap

DPA attacks



1. Can you name three types of leakage model?
2. Can you name two distinguishers?
3. What is missing in this figure?

DPA attacks


$m_1, .. m_q$
E(k)
v_1
v_2
v_3
$v_{|v|}$
$\gamma(m_1, v_1)$
$\gamma(m_1, v_2)$
$\gamma(m_1, v_3)$
$\gamma(m_1, v_{|V|})$
$L()$
$\gamma(m_2, v_1)$
$\gamma(m_2, v_2)$
$\gamma(m_2, v_3)$
$\gamma(m_2, v_{|V|})$
$L()$
$L(\gamma(m_2, v_{|V|}))$
$\gamma(m_q, v_1)$
$\gamma(m_q, v_2)$
$\gamma(m_q, v_3)$
$\gamma(m_q, v_{|V|})$
$L()$
Predicted hypothetical values for target intermediate
Leakage model of the predicted hypothetical values for target intermediate
$d_q(v_1)$
$d_q(v_2)$
$d_q(v_3)$
$d_q(v_{|v|})$

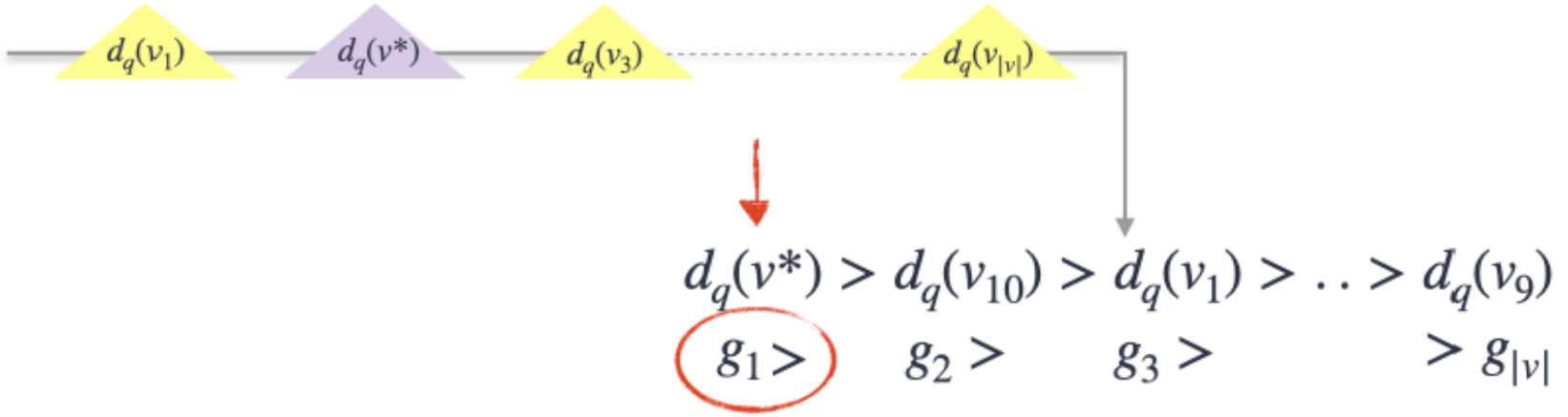
This lecture

1. 1. DPA success metrics
2. 2. Hypothesis testing: a primer
3. 3. Leakage Assessment

1. Success Metrics in SCA

Guessing Entropy/Key rank

Lets assume we have the results of a key recovery experiment with q queries. We know that the correct value is v^*:



The result is the guess vector:

Position of the correct key candidate = 1

$$g_q = [g_1, g_2, g_3, \cdots g_{|v|}]$$

Guessing Entropy/Key rank

Guessing entropy gives the correct key's average position in several experiments

How do we compute it?

Number of queries(traces)

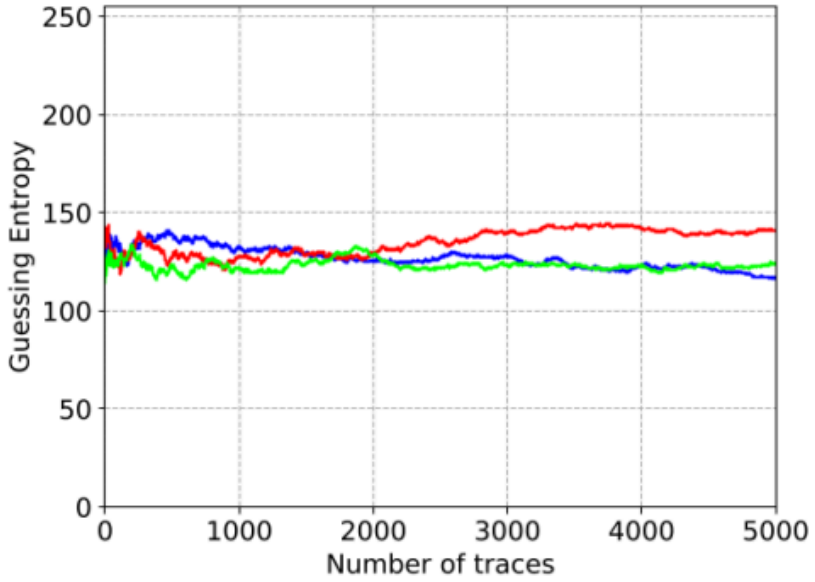
Position of the correct key candidate

$$GE(q) = E[i, g_i(v^*) \in g_q]$$

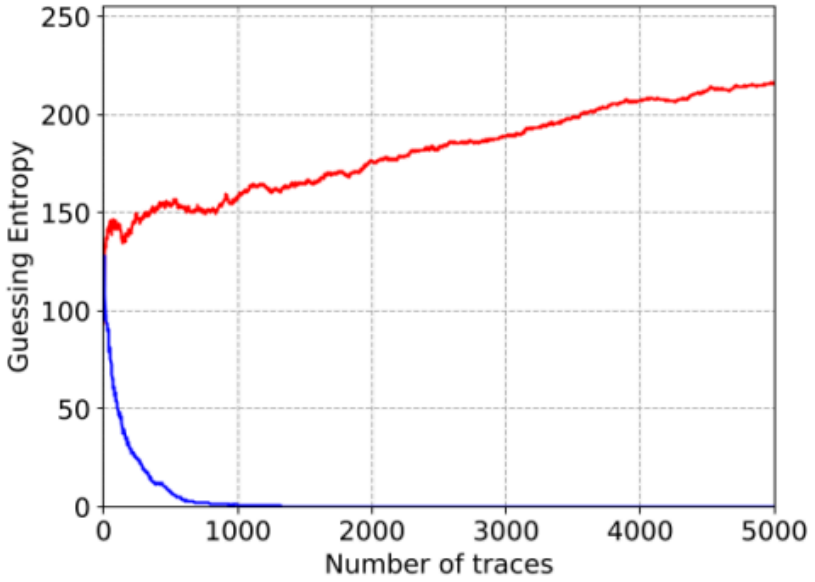
Expectation (average), from multiple experiments

Guess vector for q queries

Guessing Entropy/Key rank



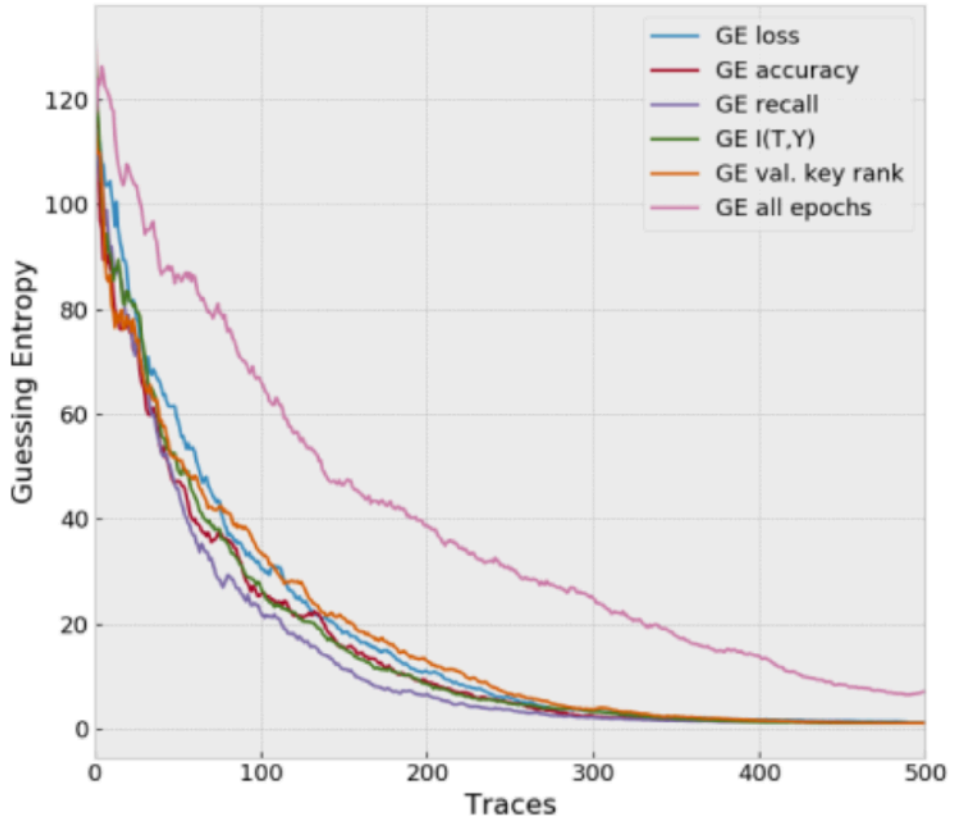
(a) CNN from [26] on the synchronized ASCAD dataset with random keys and the identity leakage model.



(b) CNN from [26] on the synchronized ASCAD dataset with a fixed key and the Hamming weight leakage model.

Guessing Entropy/Key rank

Figure 7: Results on ASCAD for the Hamming weight leakage model, CNN architecture.



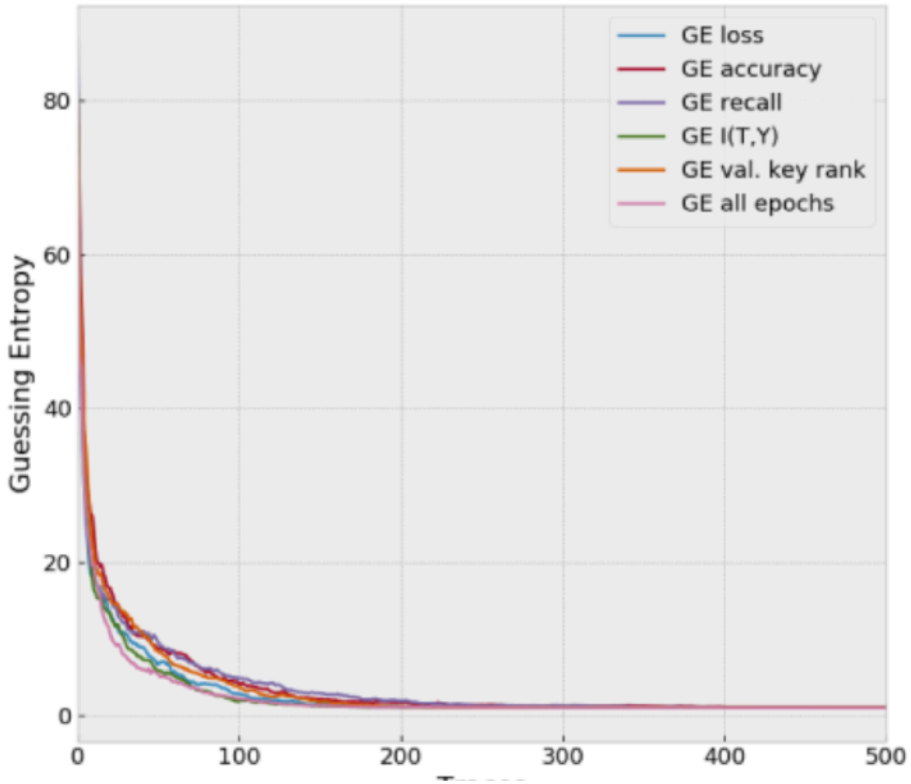


Figure 11: Results on ASCAD for the identity leakage model, CNN architecture.

Guessing Entropy/Key rank

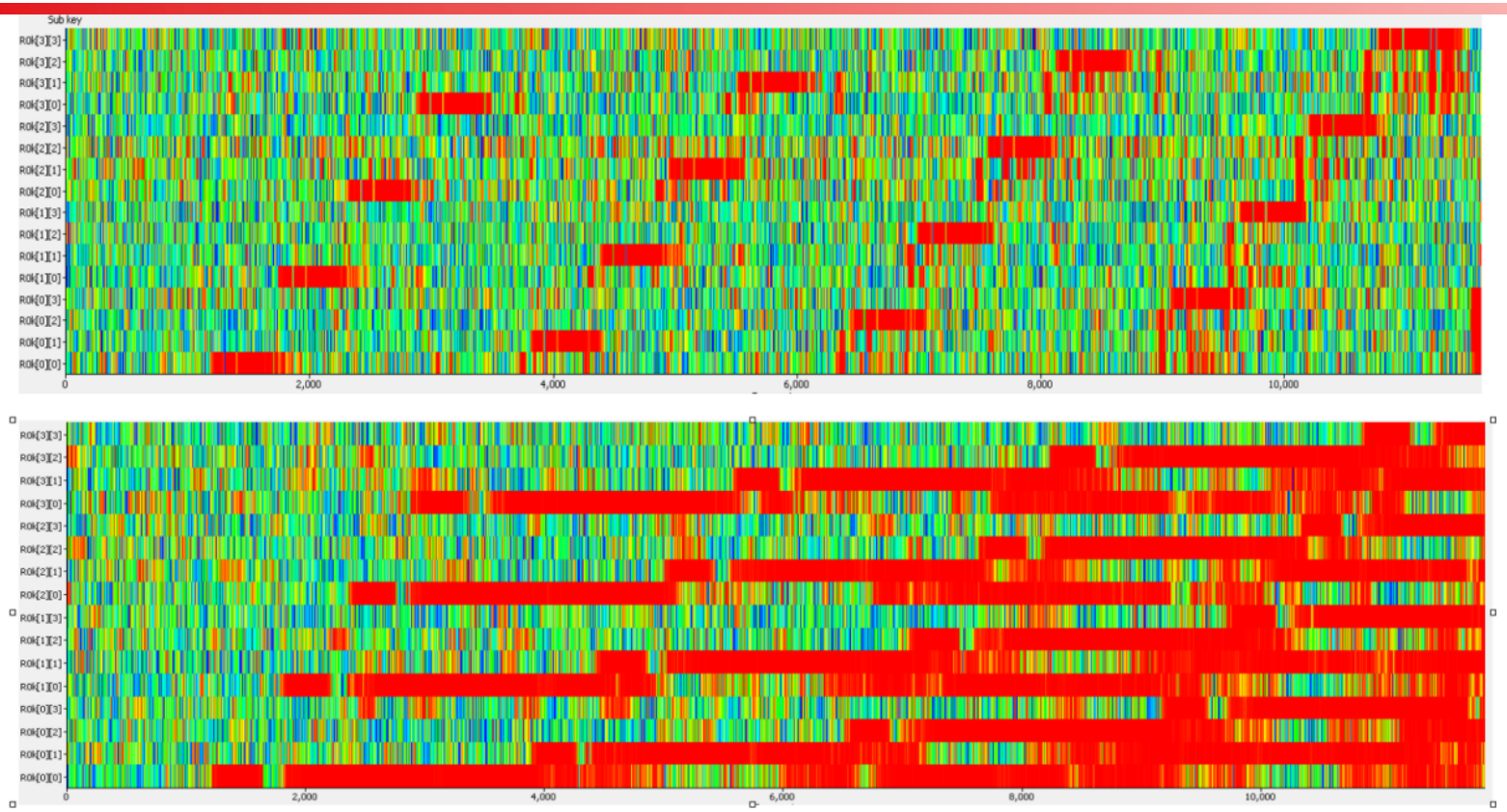


Fig. 6: (To be viewed in colors.) Key rank results for the `STM32` (top) and `NRFf51` (bottom) devices. The selection captures the s-box operation. We perform key ranking on all 16 bytes in round 1. The red color indicates strong leaks, where the correct key candidate is ranked in the first position, whereas other colors represents weak leaks.

Vipul Arora, Ileana Buhan, Guilherme Perin, Stjepan Picek:
A Tale of Two Boards: On the Influence of Microarchitecture on Side-Channel Leakage. CARDIS 2021: 80-96

Guessing Entropy/Key rank

Why is it useful?

It measures:

- the average number of key candidates to be tested after a side-channel attack

- how much a side-channel attack reduced the complexity of an exhaustive key search

Success Rate of order o

Is the correct target intermediate ranked within the first o positions?

$$g_q = [g_1, g_2, g_3, \cdots g_{|v|}]$$

Order = 3

Order

$$SR^o(q) = Pr\{v^* \in [g_1, \cdots g_o]\}$$

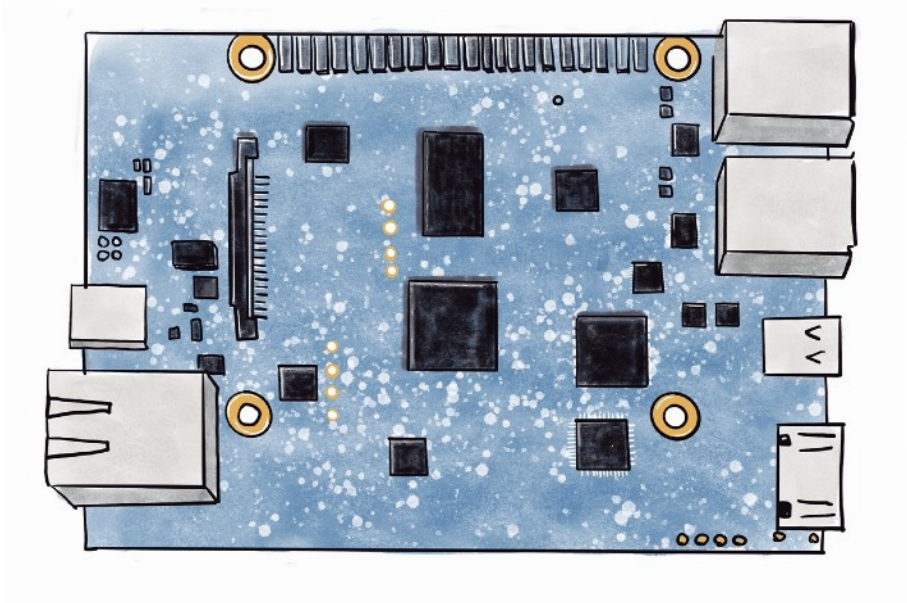
Number of queries

Probability

2. Hypothesis testing: a primer

Main question

Can we check vulnerability to side-channel attacks without doing an attack?
How do we decide that a device is leaking information?
We put the device on trial!

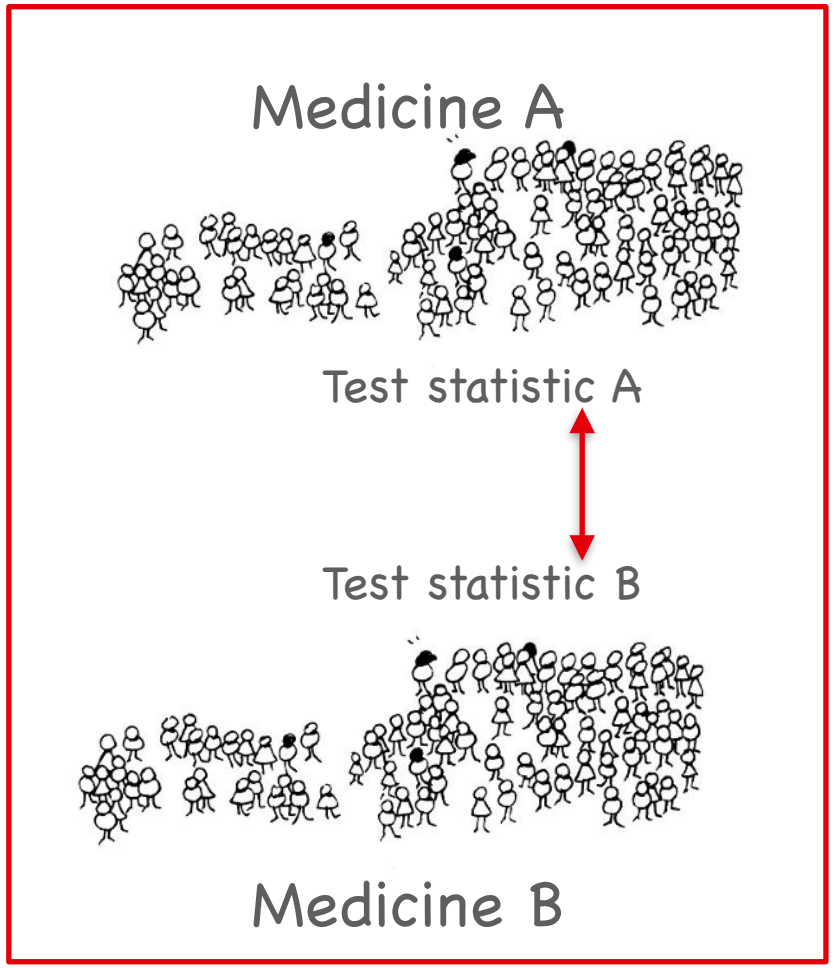


H0: device is NOT guilty of leaking information

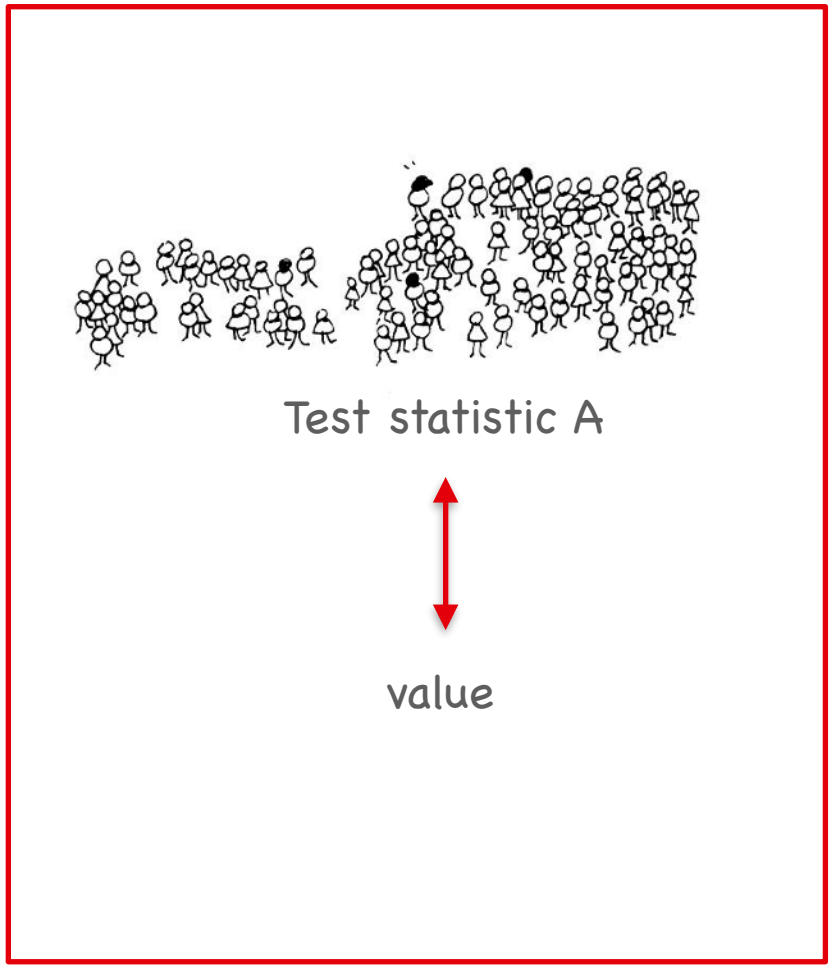
H1: device IS guilty of leaking information

We use Null Hypothesis Significance Testing!

Two types of questions



Two-sample test



One-sample test

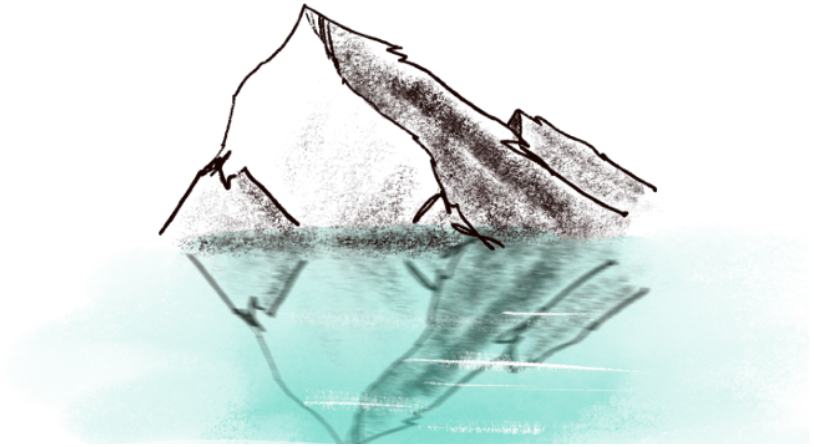
What is a hypothesis?

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

Population vs sample data

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



Population

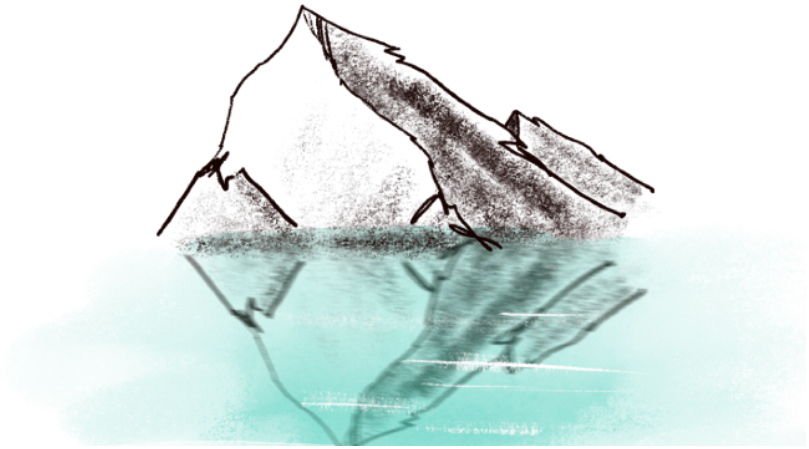
μ, σ

Sample data

$\bar{x}, s$

Recall

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



Population

μ, σ

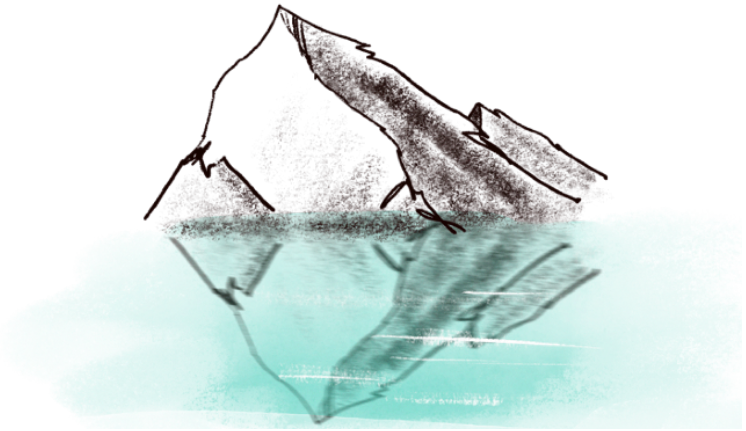
Is this question:

(A)Two-sample test

(B) One-sample test

What is hypothesis testing?

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



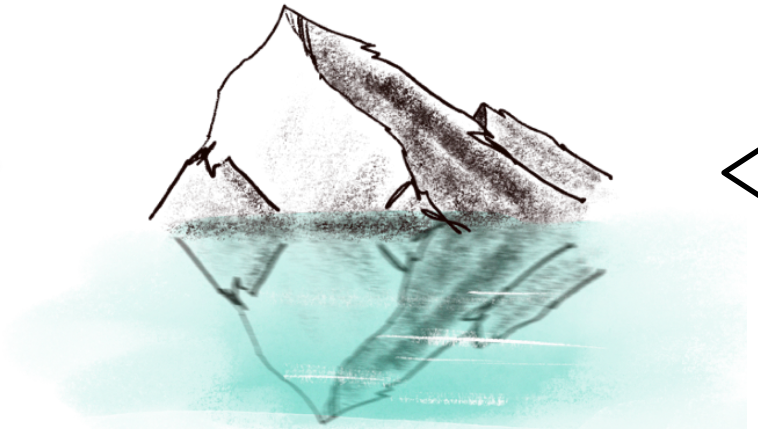
Hypothesis test

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 how compatible the experimental results are with the hypothesis test.

population parameter



Hypothesis test

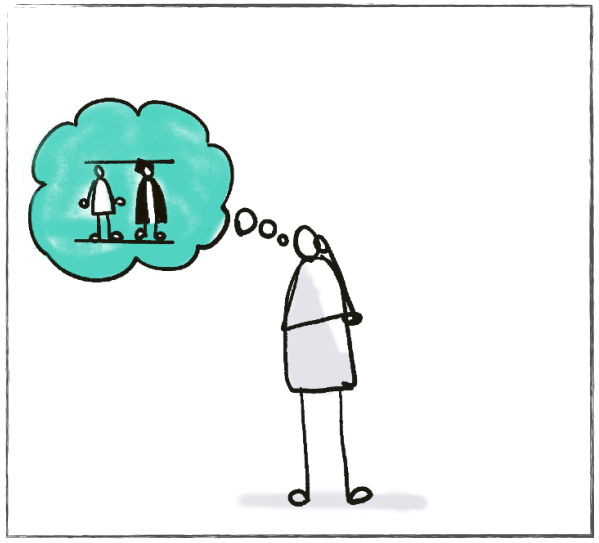
test statistic

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%.

NHST in three steps

1. Select the H_0, α;



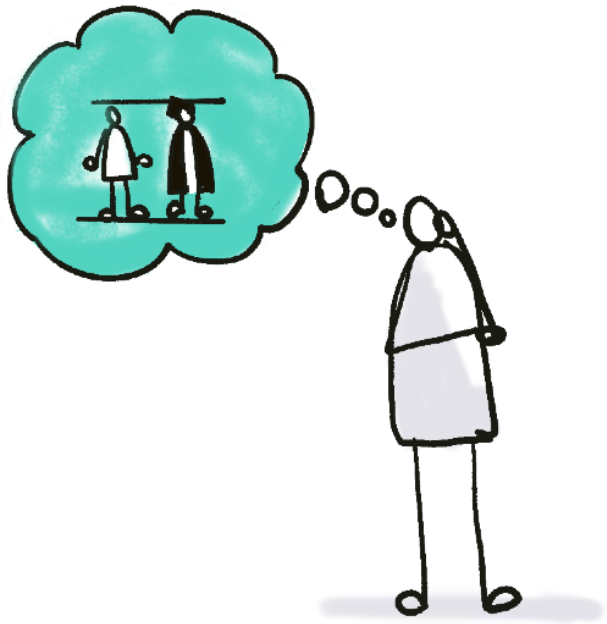
2. Collect data

3. Test

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

In this type of testing, we cannot accept H_0, only reject it! Why?

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.

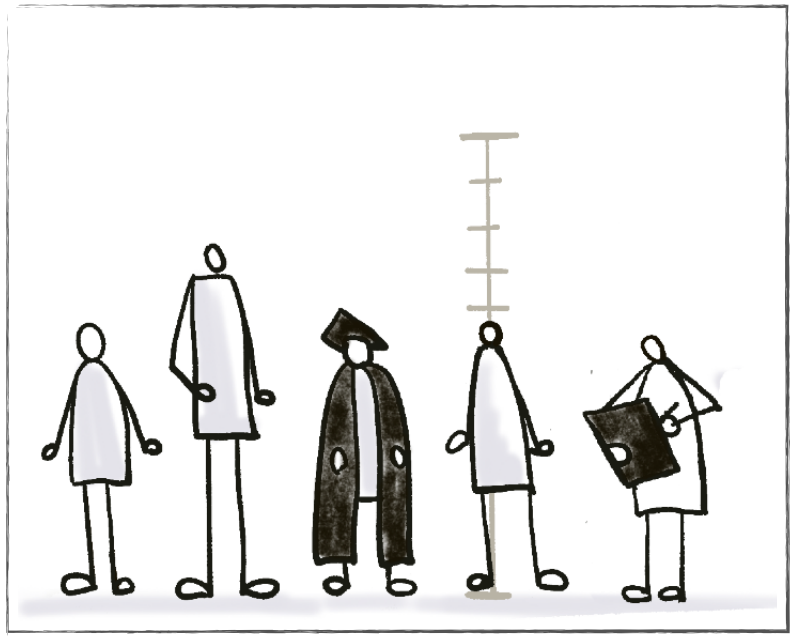
A good null-hypothesis is interesting to reject and must be specific.

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.

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.

3. Test the significance

A. 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?

B. p-values

C. Significance level

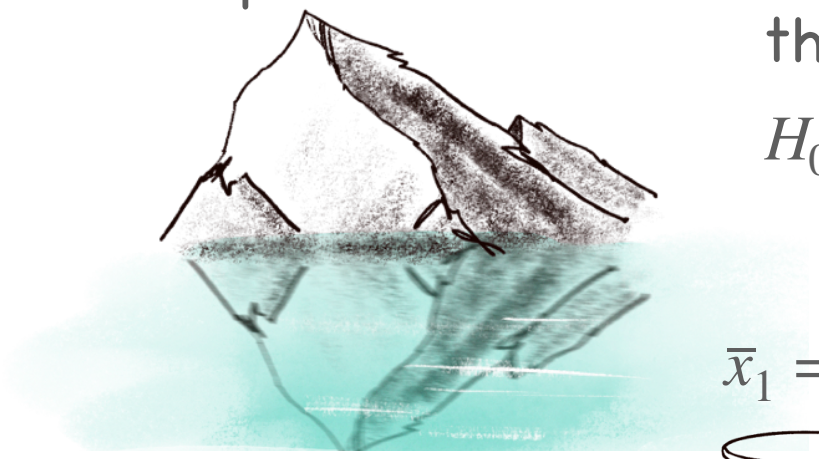
3A. The null-distribution

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

3A. 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%.

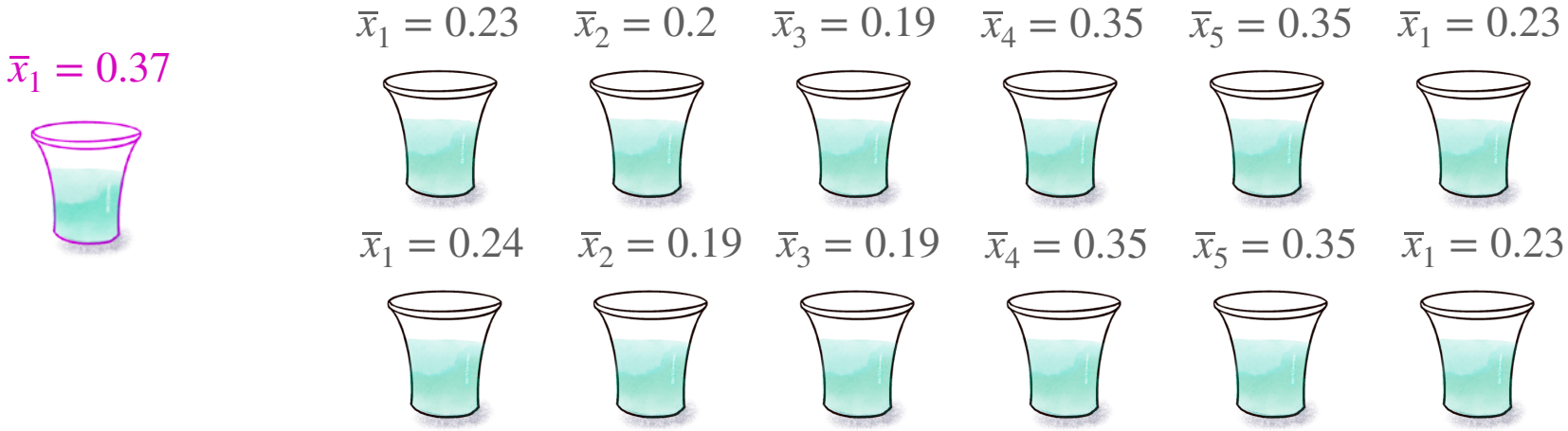
$H_0 : \mu_{salt} = 0.3$

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

*one-sample t-test

3.B P-value

The p-value is the probability that a value at least as extreme as our test statistics ($\bar{x}_1 = 0.37$) is observed when $H_0 : \mu_{salt} = 0.3$

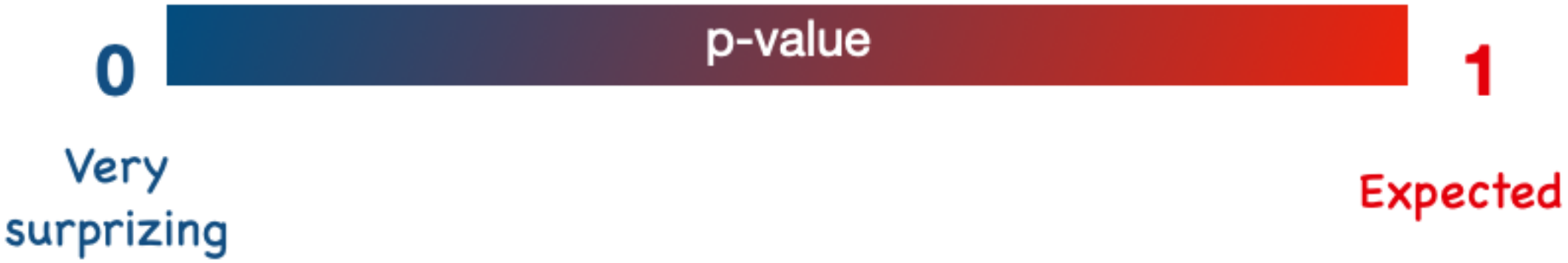


For which case would you reject H_0?

- small p-value - (the observed sample test is unlikely)
- large p-value - (the observed test is very likely)

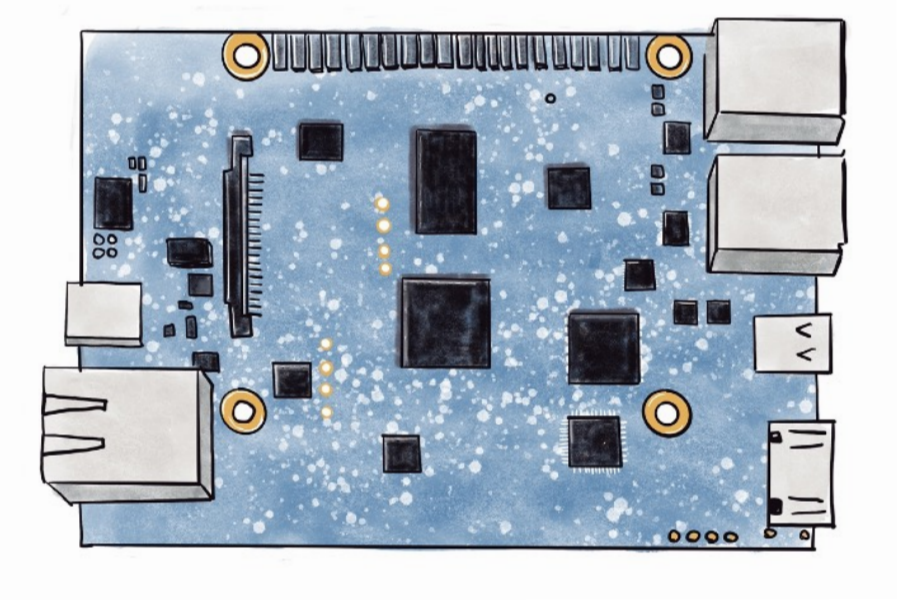
3.B P-value

For which case would you reject H_0?



Note: to compute a p-value we need:
(1) null-distribution and
(2) an observed sample statistic.

3c. The significance level α



H_0: device is NOT guilty of leaking information

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.

3. Leakage assessment

Leakage detection in action

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

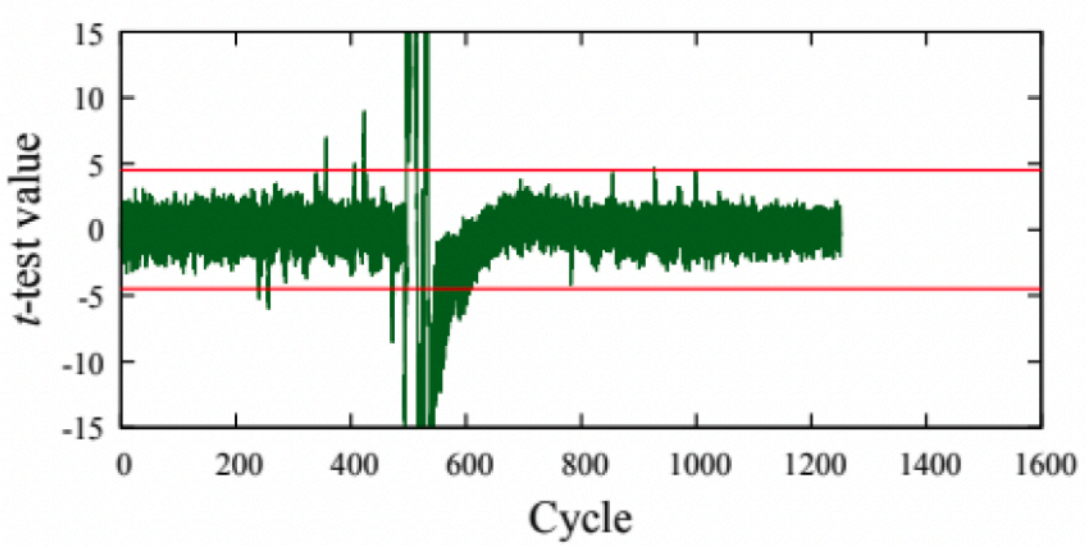
- Non-specific or general test: aims to detect any leakage that depends on input data (or key);

a.k.a fixed - vs - random;

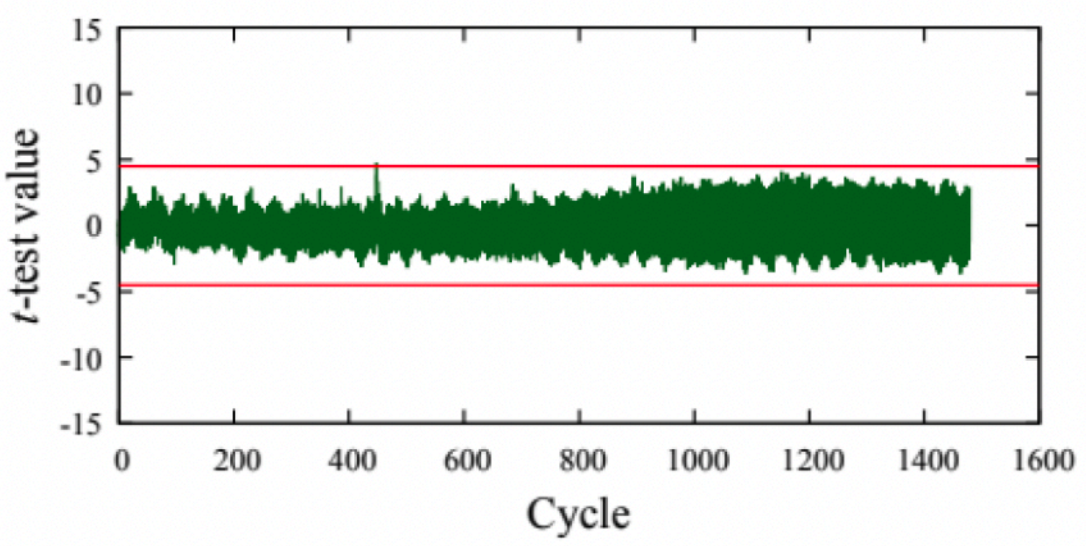
- Specific-test: targets a specific intermediate value of the cryptographic algorithm that could be exploited to recover keys or other sensitive information.

a.k.a fixed - vs - fixed;

TVLA in action



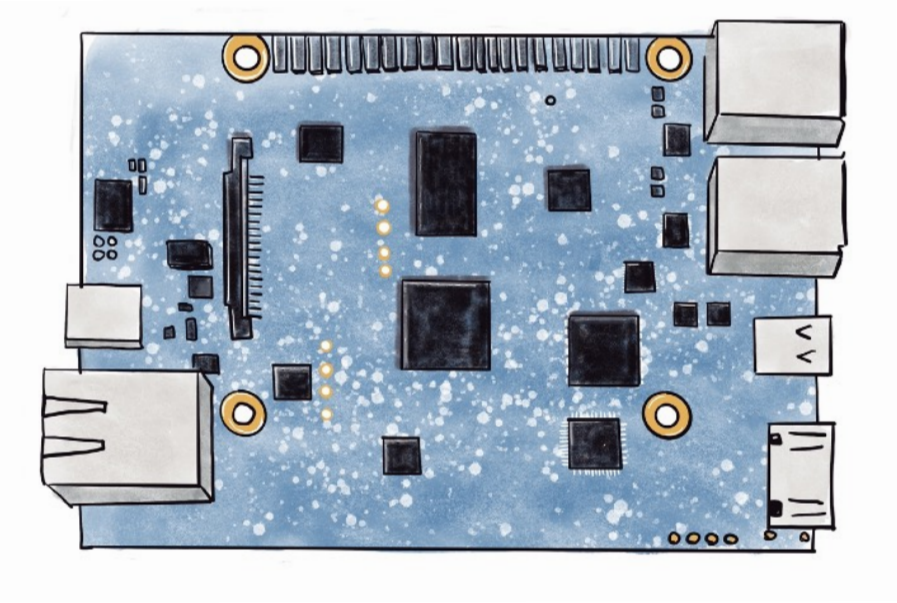
(a) AES original implementation.



(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

TVLA 1. H_0



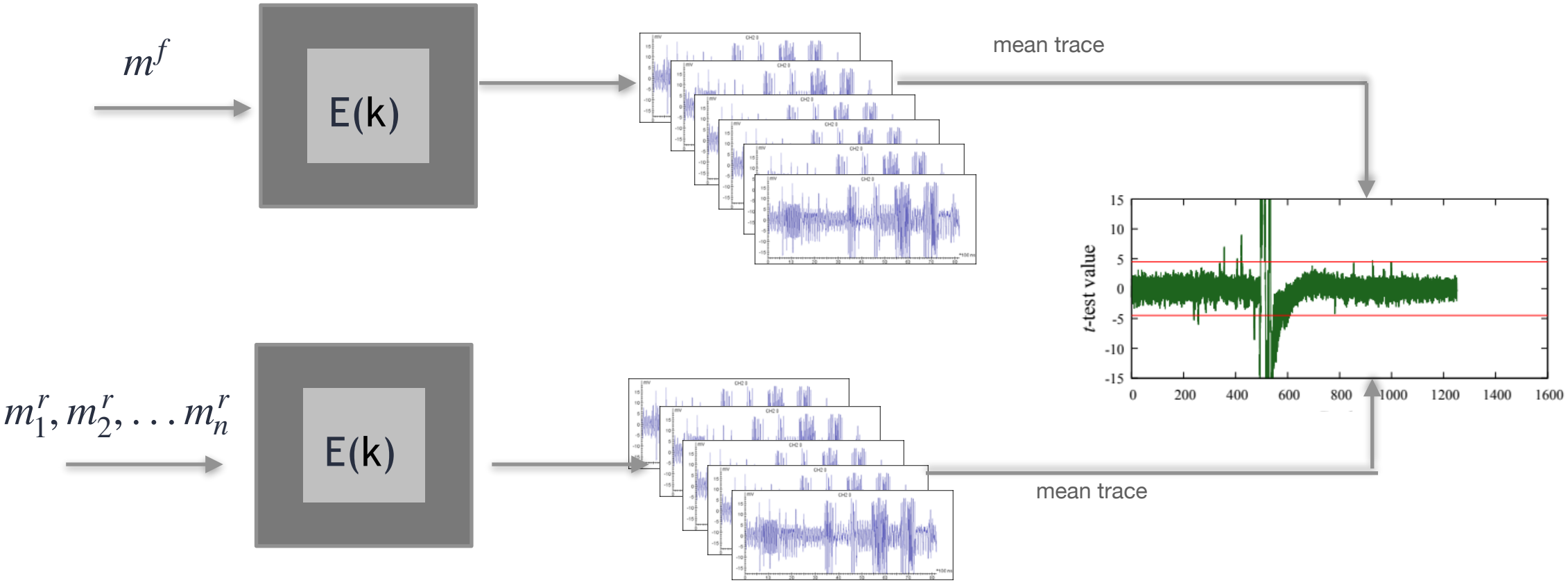
H_0: device is NOT guilty of leaking information

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

H_a: device is guilty of leaking information

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

TVLA 2. Collect data



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 3. Test significance of observed results

(1) Compute the sample statistic (standardized difference or t-score), $t = \frac{\bar{x}_f - \bar{x}_r}{\sqrt{s_f^2 + s_r^2}}$

where $s^2 = \frac{1}{n}\sum_{i=0}^{n}(x_i - \bar{x})^2$ is the sample standard deviation and $\bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i$ is the sample mean;

(2) Produce the appropriate null distribution (a t-distribution):

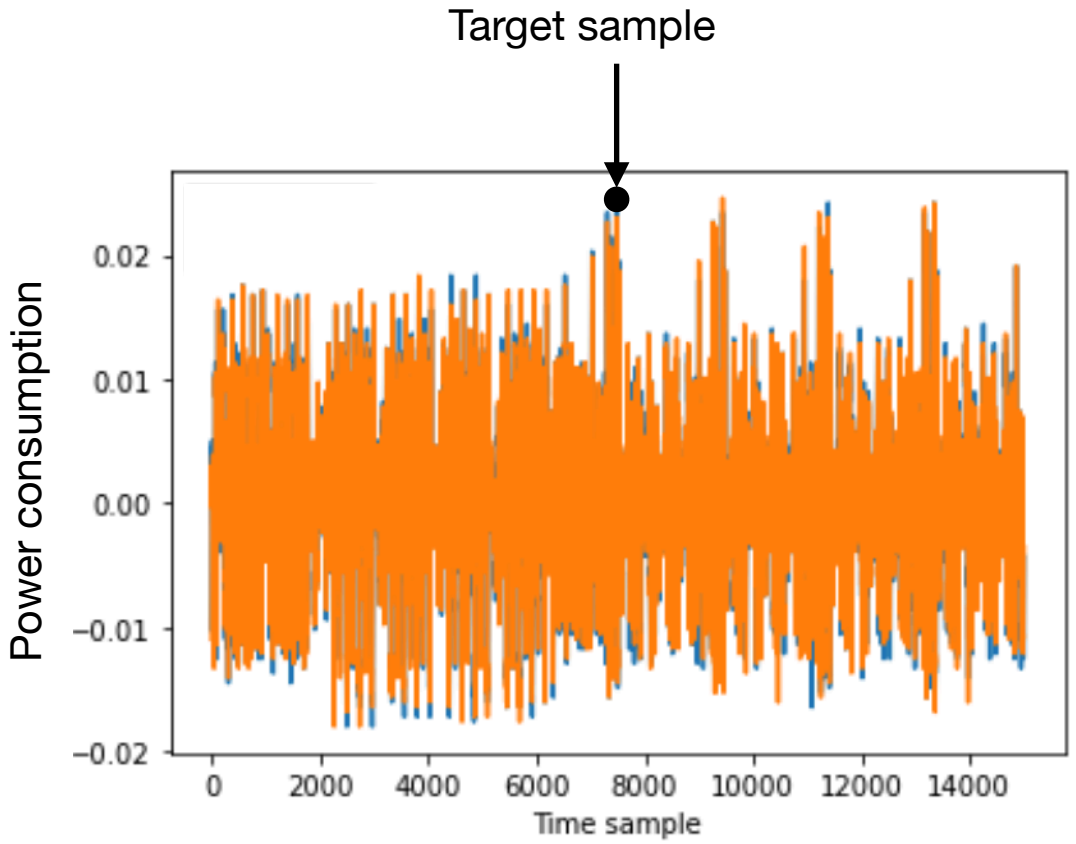
$$f(x, v) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi v}\Gamma(\frac{v}{2})}(1 + \frac{x^2}{v})^{-\frac{v+1}{2}}$$

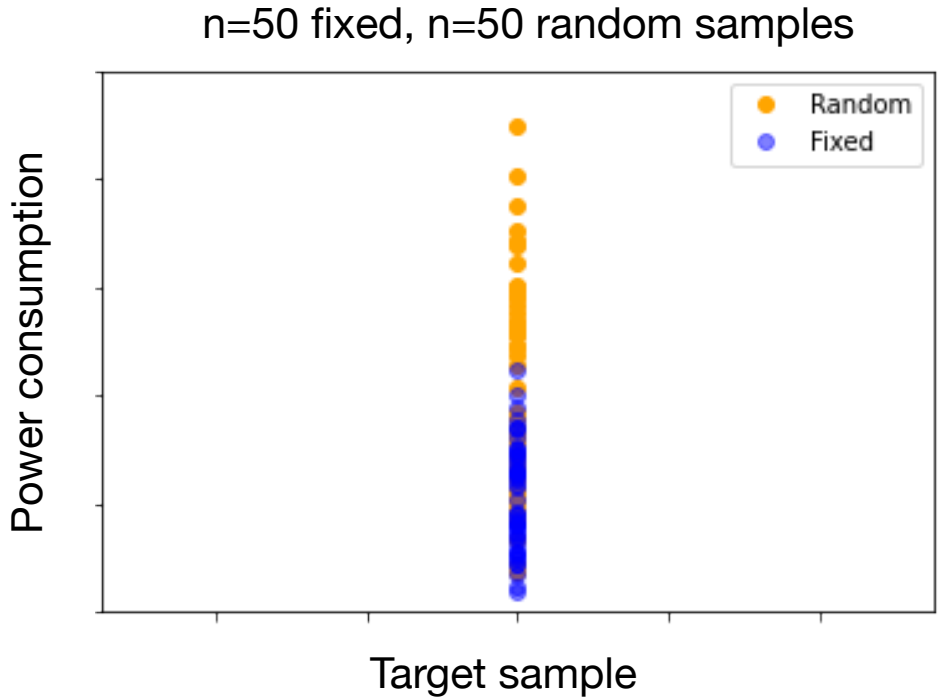
where $v = N_f + N_r - 2$ represents the degrees of freedom and Γ is the Gamma function.

(3) Compute the p-value as $p = 2\int_t^{\infty} f(x, v)dx$

(4) Compare the p-value to the chosen significance level (α)

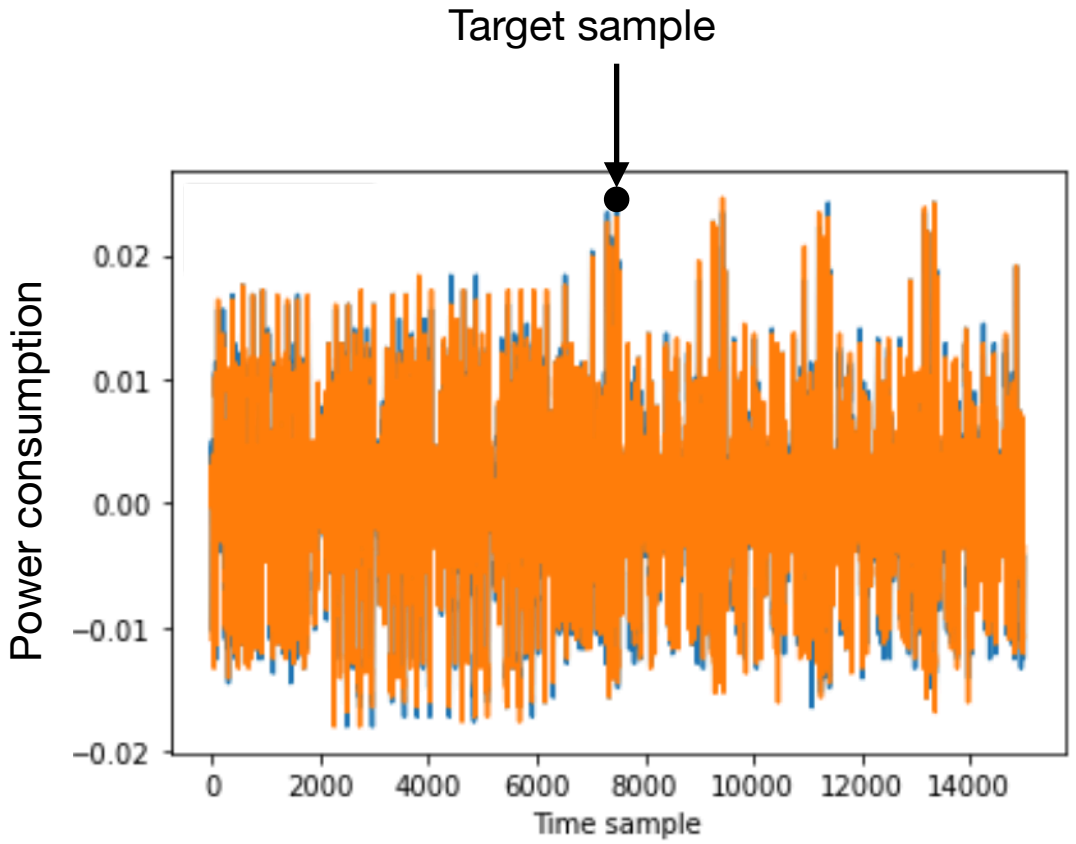
TVLA - two-sample t-test

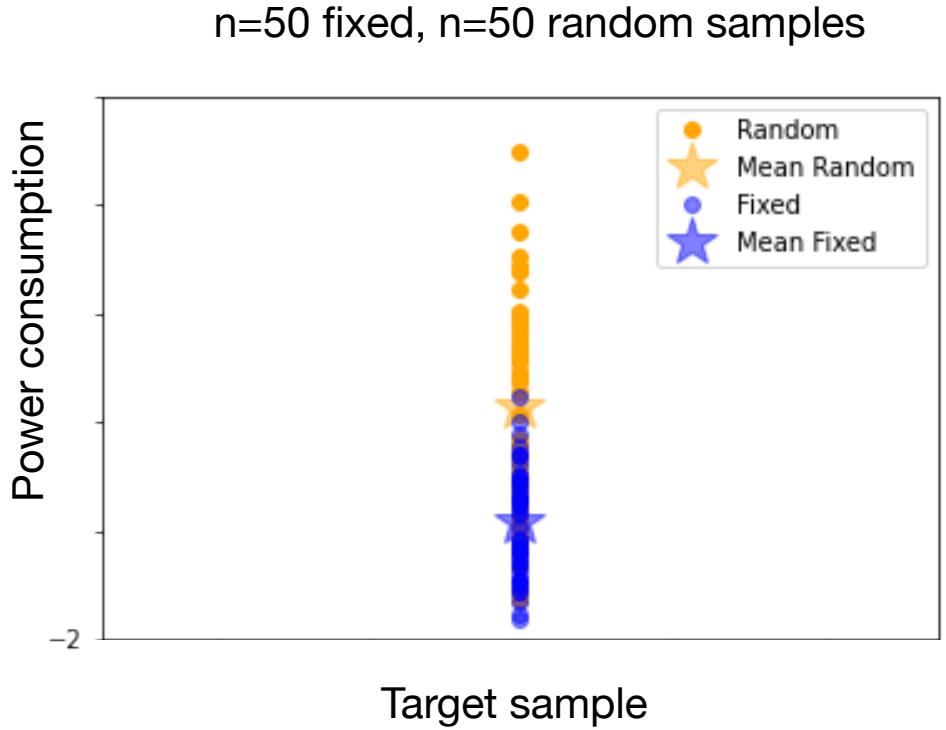




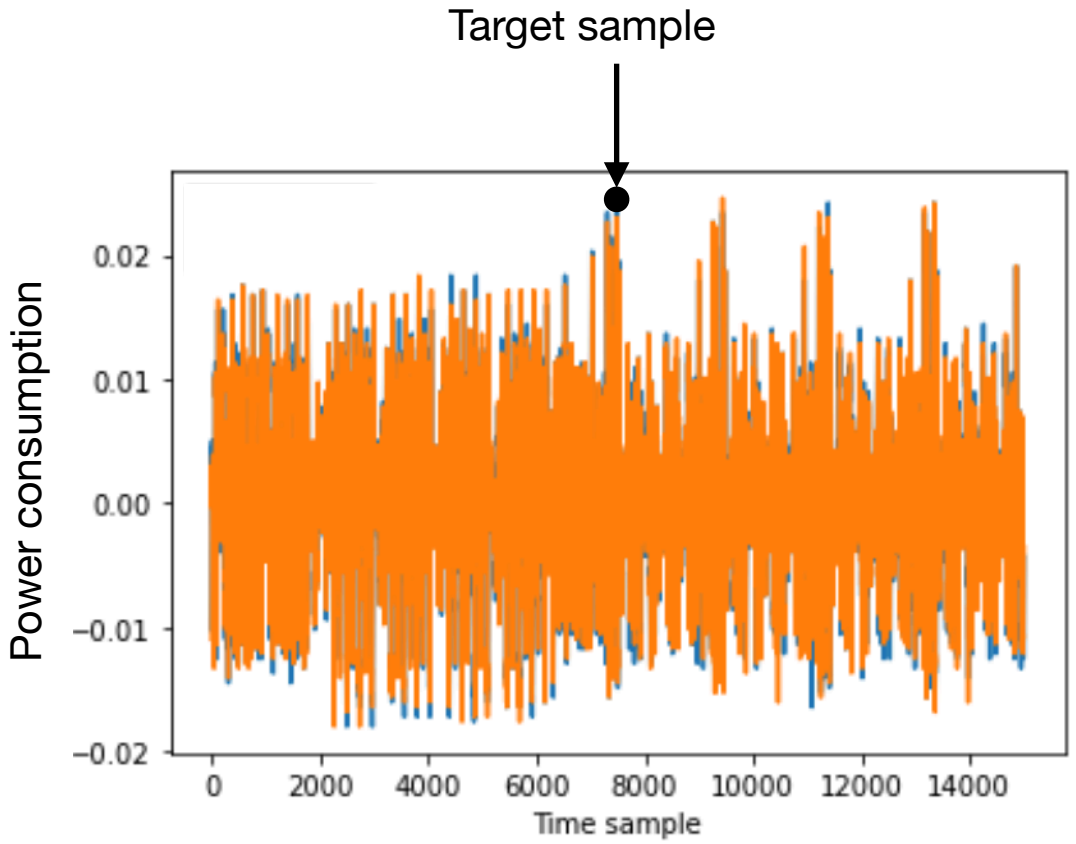
First order t-test, we analyze each sample independently

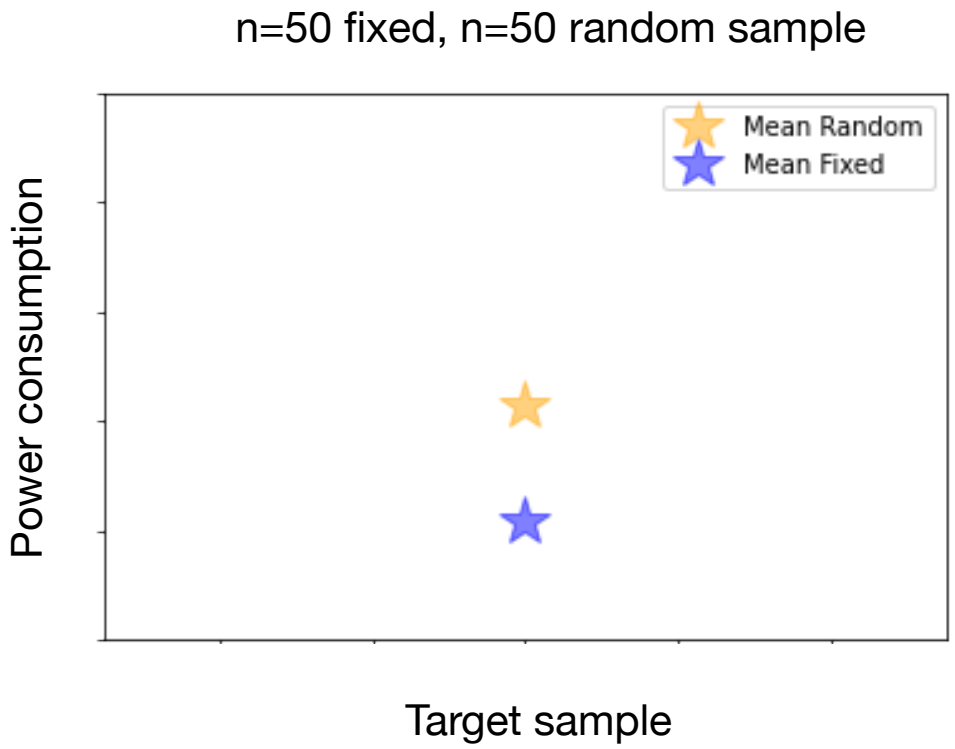
TVLA - two-sample t-test



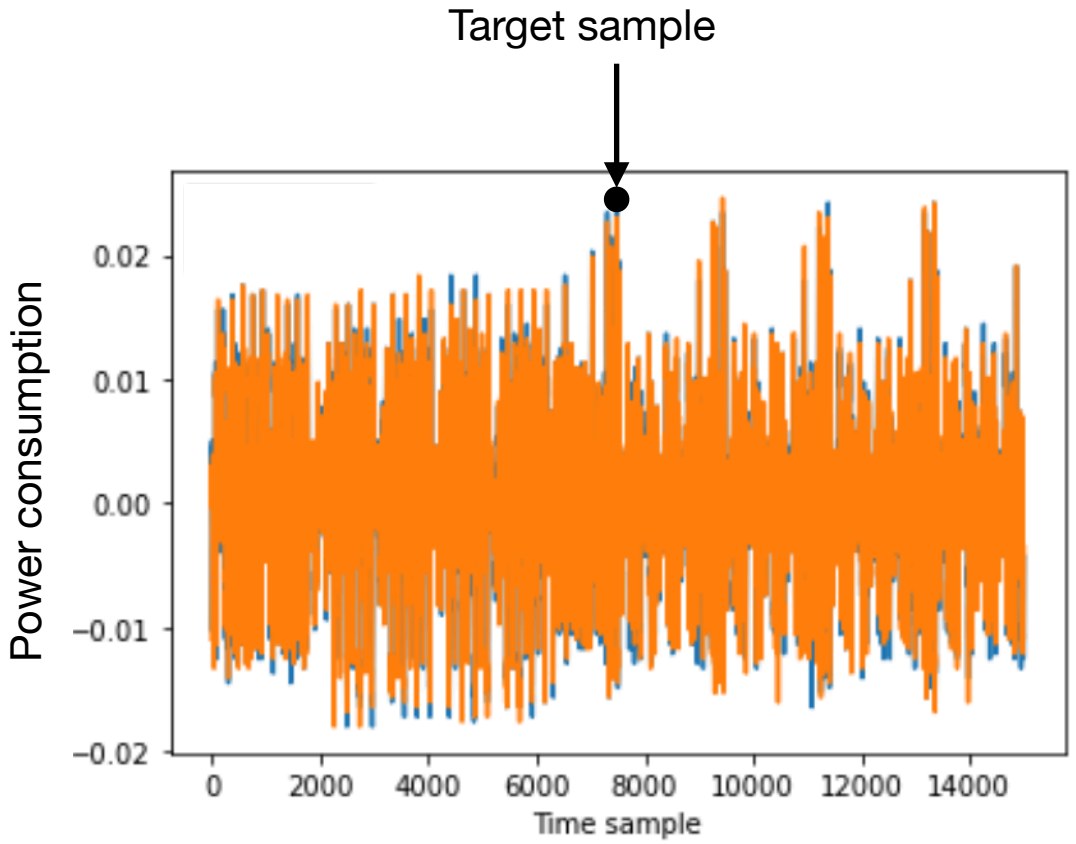


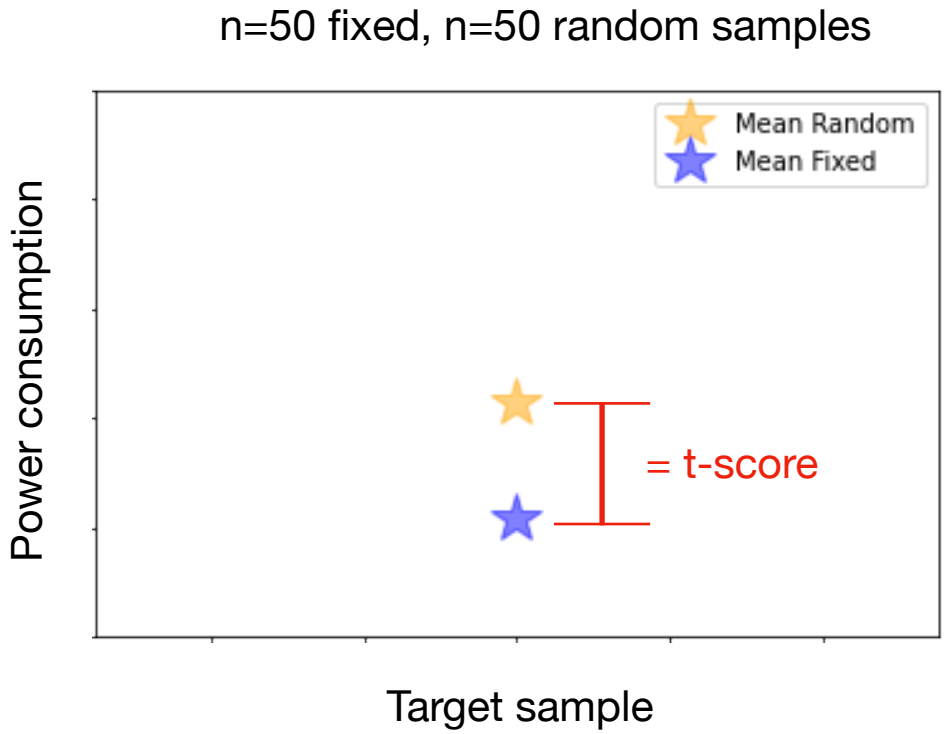
TVLA - two-sample t-test





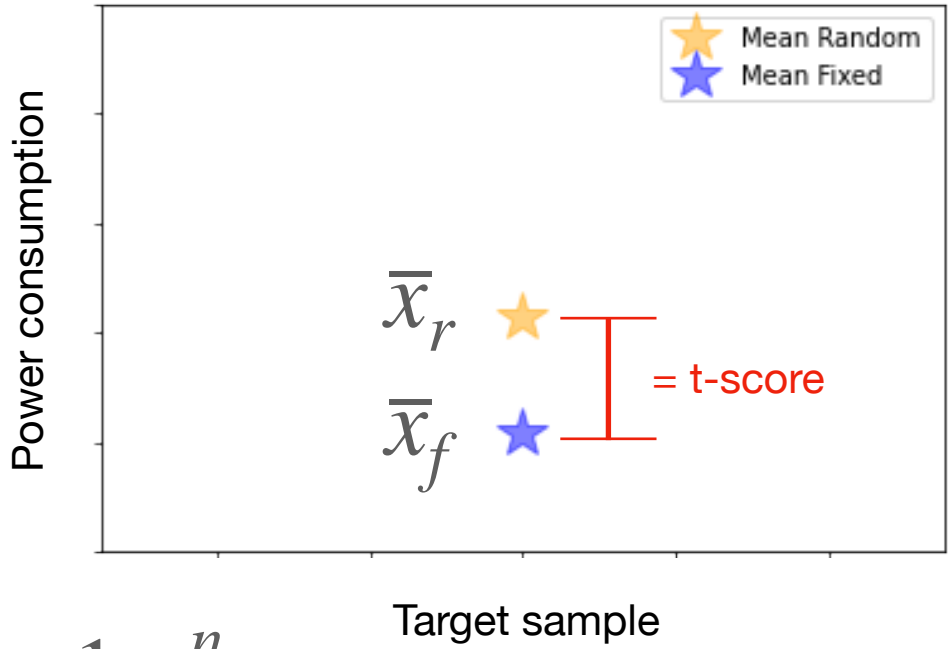
TVLA - two-sample t-test





TVLA - two-sample t-test

$$t = \frac{\bar{x}_f - \bar{x}_r}{\sqrt{\left(\frac{s_f^2 + s_r^2}{n-1}\right)}}$$

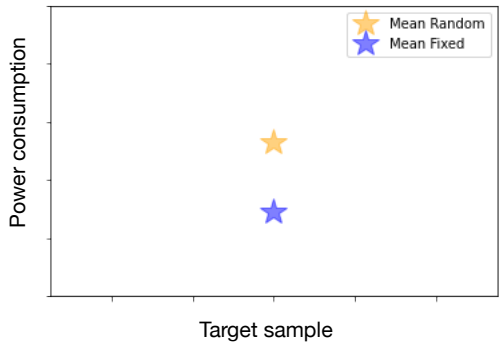


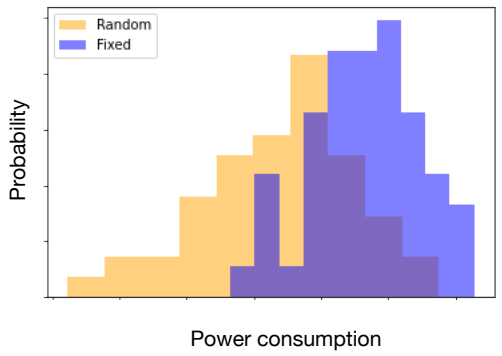
$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \bar{x} = \frac{1}{n}\sum_{i=0}^{n} x_i$$

Experiment 1 ($\mu_f \neq \mu_r$)

T-score= -6.019

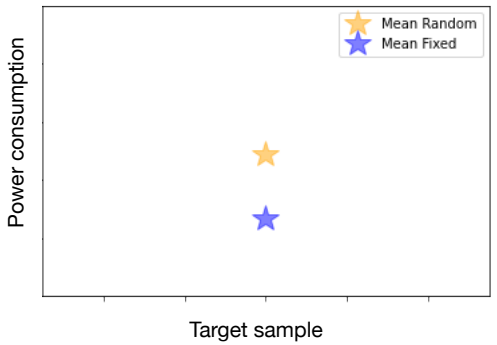
p-value= $3.3*10^{-8}$

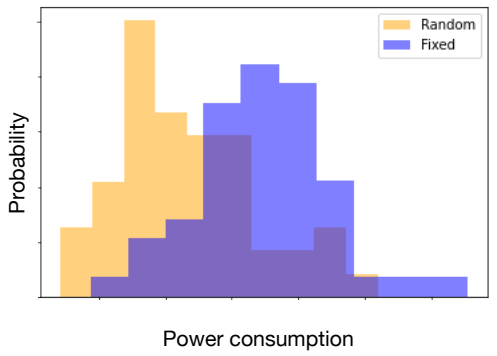




T-score= -5.13

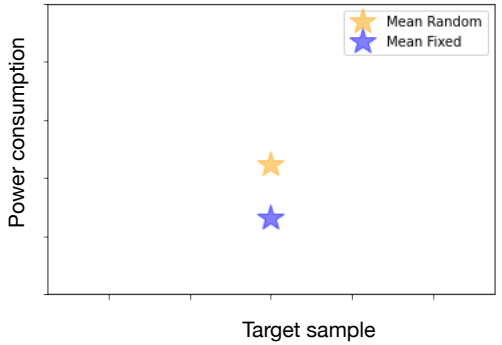
p-value= $1.40*10^{-6}$

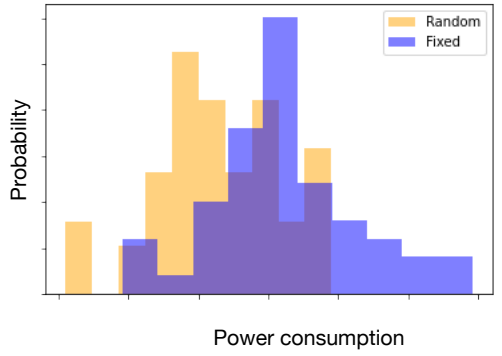




T-score= -4.62

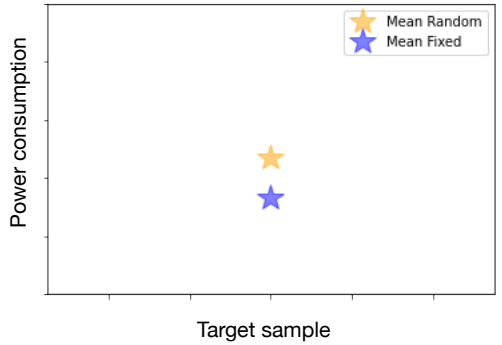
p-value= $1.15*10^{-5}$

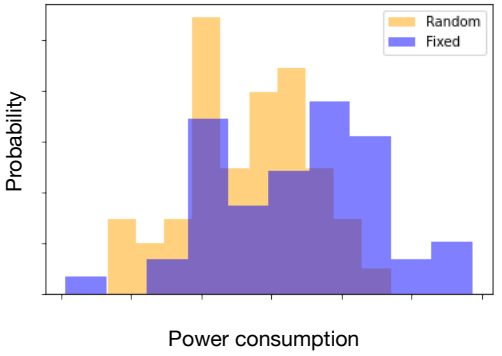




T-score= -3.25

p-value= 0.001



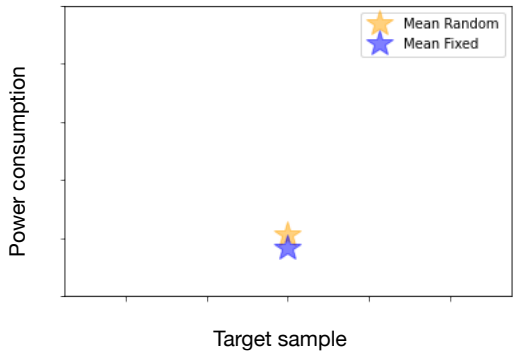


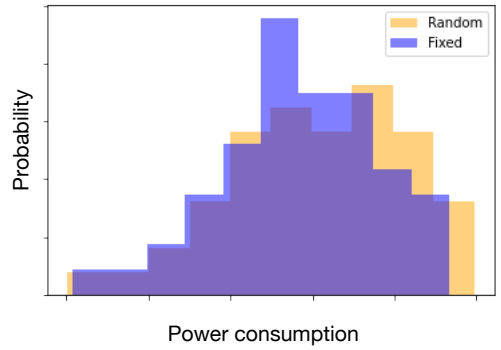
$\mathcal{N}(1,2)$ $\mathcal{N}(3,2)$

Experiment 2 ($\mu_f = \mu_r$)

T-score= -1.034

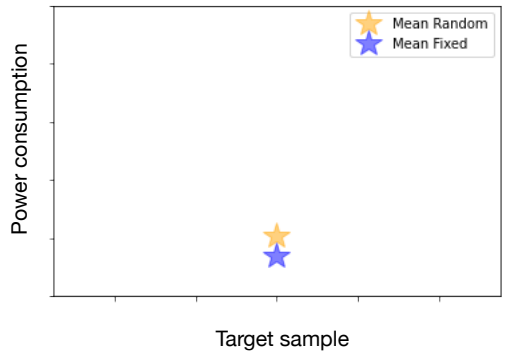
p-value= 0.303

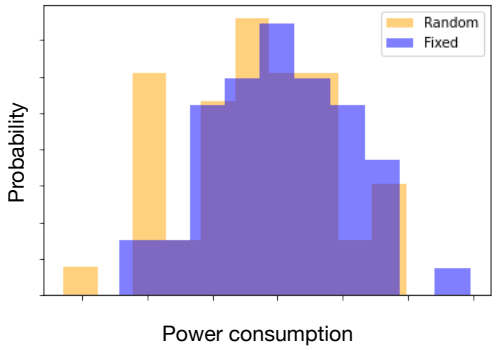




T-score= -1.540

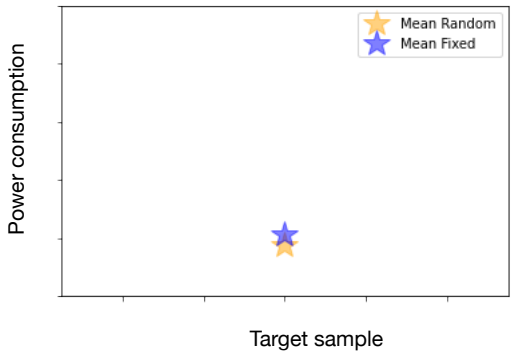
p-value= 0.126

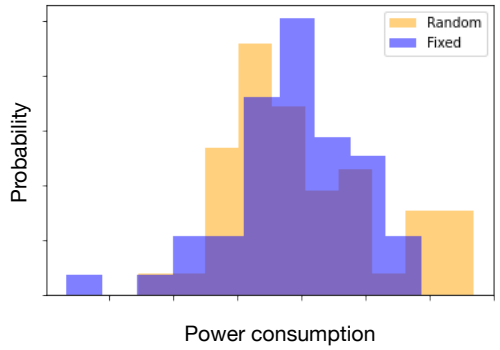




T-score= 0.875

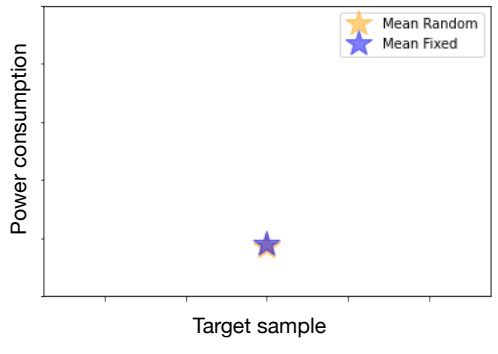
p-value= 0.3836

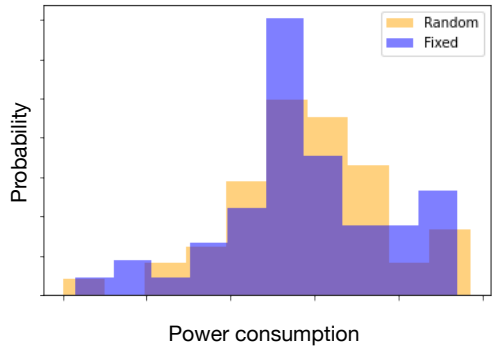




T-score= 0.254

p-value= 0.8



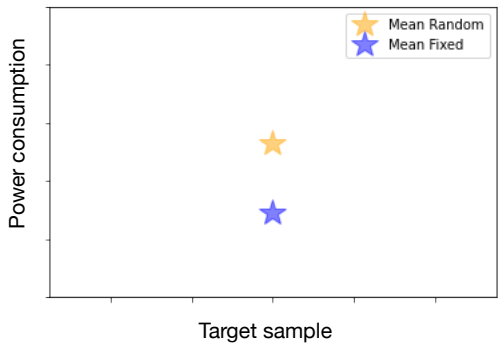


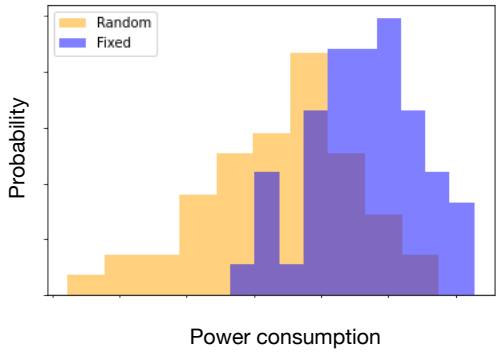
$\mathcal{N}(0,2)$ $\mathcal{N}(0,2)$

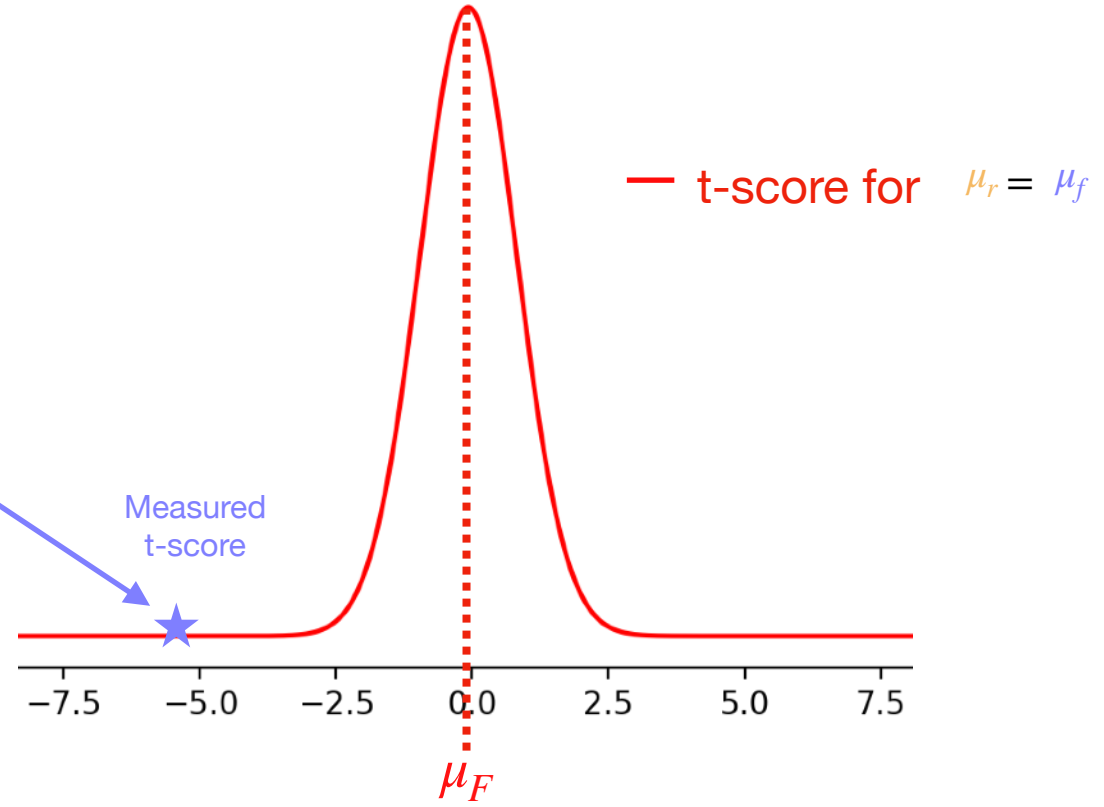
T-scores and P-values

T-score= -6.019

p-value= $3.3*10^{-8}$







P- value answers the question:

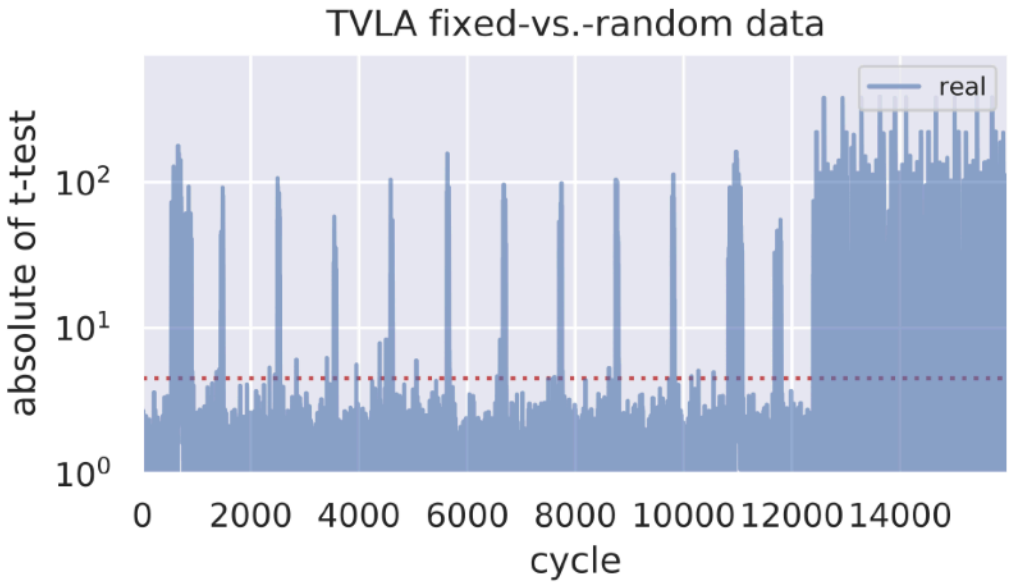
*If I live in a world where **H_0 is true**, how surprizing is to measure a t-score of -6.019?*

Notes

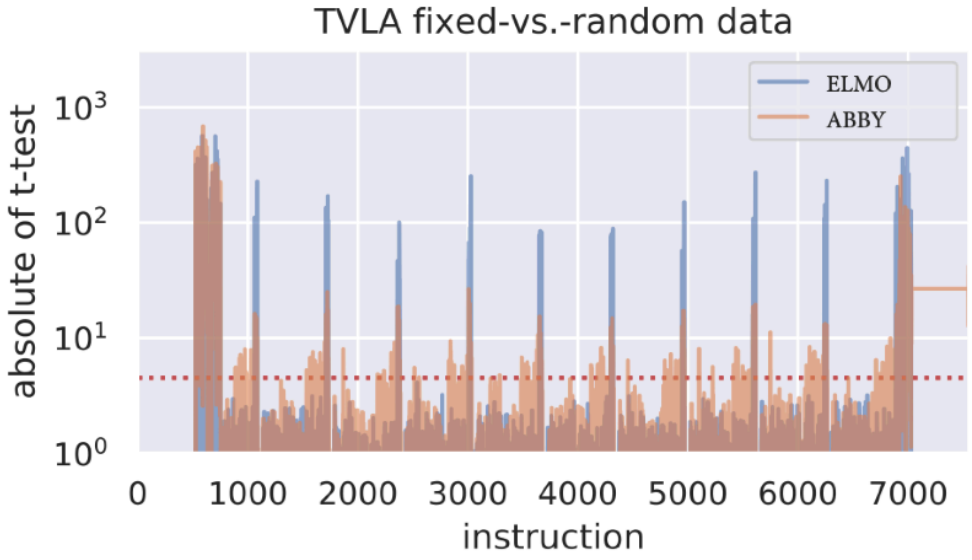
IMPORTANT: TVLA test is qualitative measure of leakage, and NOT a quantitative measure.

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

TVLA in action



(a) Measured power trace



(b) ELMO model vs ABBY model

Figure 8: Byte-Masked-AES TVLA result real vs. simulati

- Omid Bazangani, Alexandre Iooss, Ileana Buhan and Lejla Batina, *ABBY: Automating the creation of fine-grained leakage models*, ASIACCS (to appear) Link to paper

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

THANK YOU