H2020 project 731591
http://www.reassure.eu/

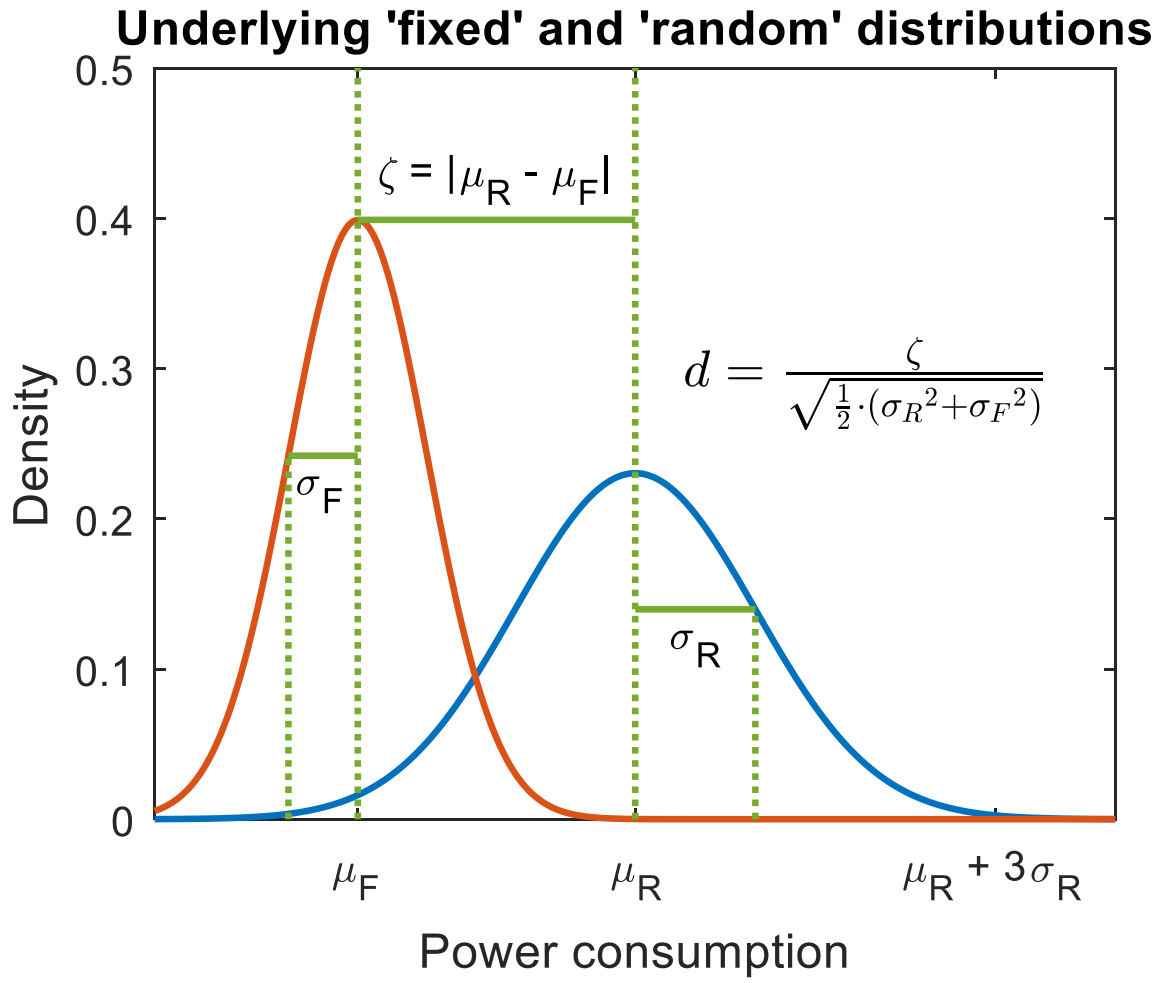
Understanding Leakage Detection

Part 3: Interpreting Outcomes

Aims

- We next revisit some of the intuitive concepts we have introduced and explain how they fit in to a formal statistical framework.
- This gives us tools to analyse the tests themselves in order to see how well they are fulfilling the criteria for thorough and fair evaluation.
- We will then show how to adapt tests to do the job better.
- Ultimately, though, we seek to highlight the *limitations* of testing methodologies and to emphasise the importance of realism in setting achievable goals and drawing measured conclusions.

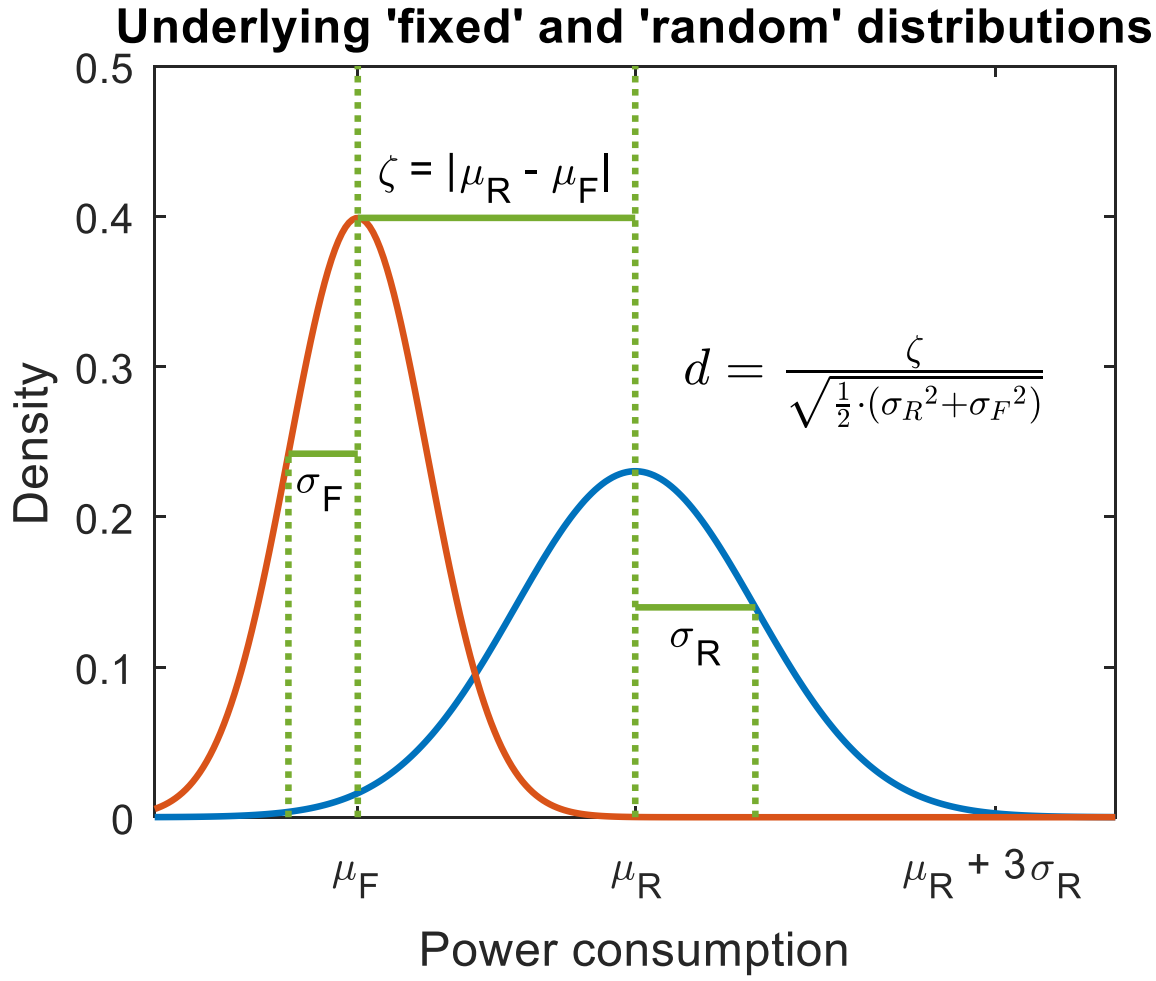
Effect sizes



- Remember the distance between the means that had an impact on the fixed-versus-random outcome? This is called the *effect size* (denoted ζ).
- If you divide it by the pooled standard deviation of the two distributions, you get the *standardised effect size* (denoted *d*) [Coh1988].

[Coh1988] J. Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 1988.

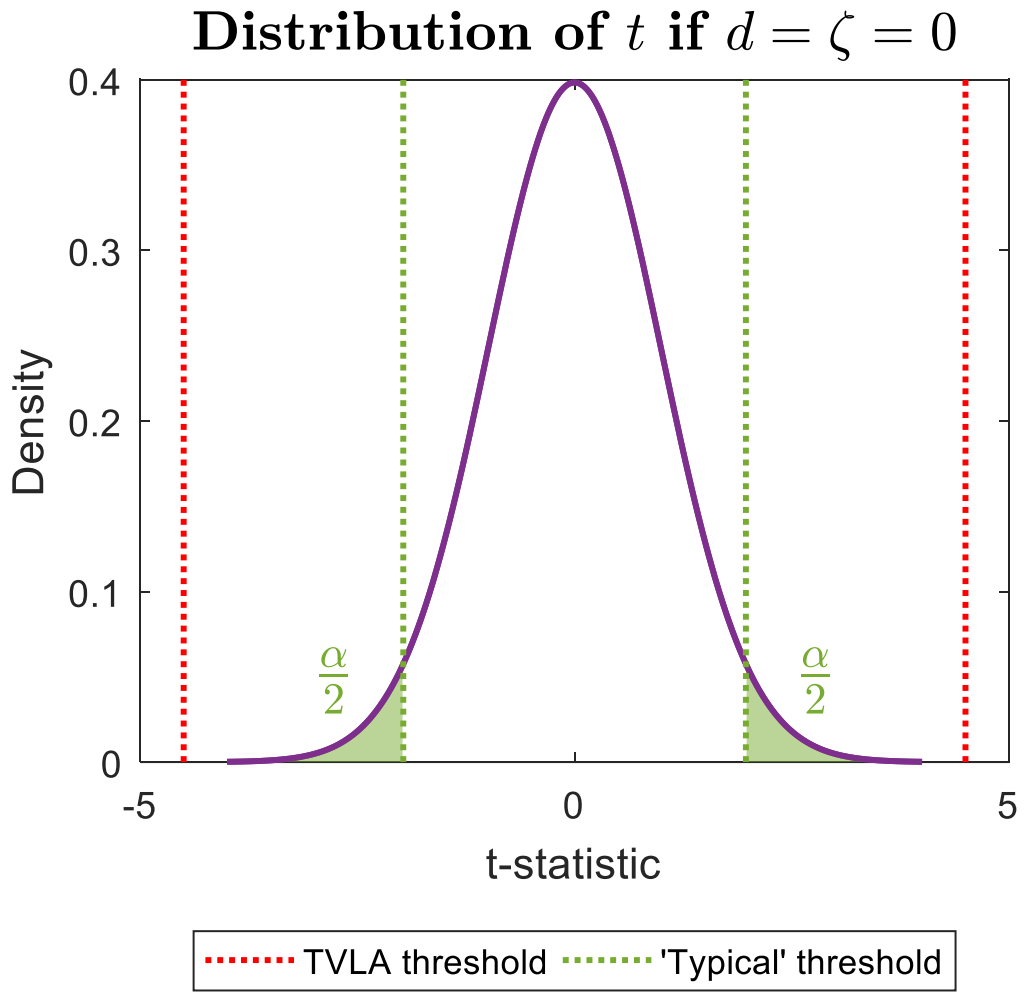
Null hypothesis and the test statistic



- The scenario that $d = \zeta = 0$ is called the *null hypothesis*, and we are looking for evidence to reject it.
- Recall the test statistic *t* that we compute from *samples* of the two distributions:

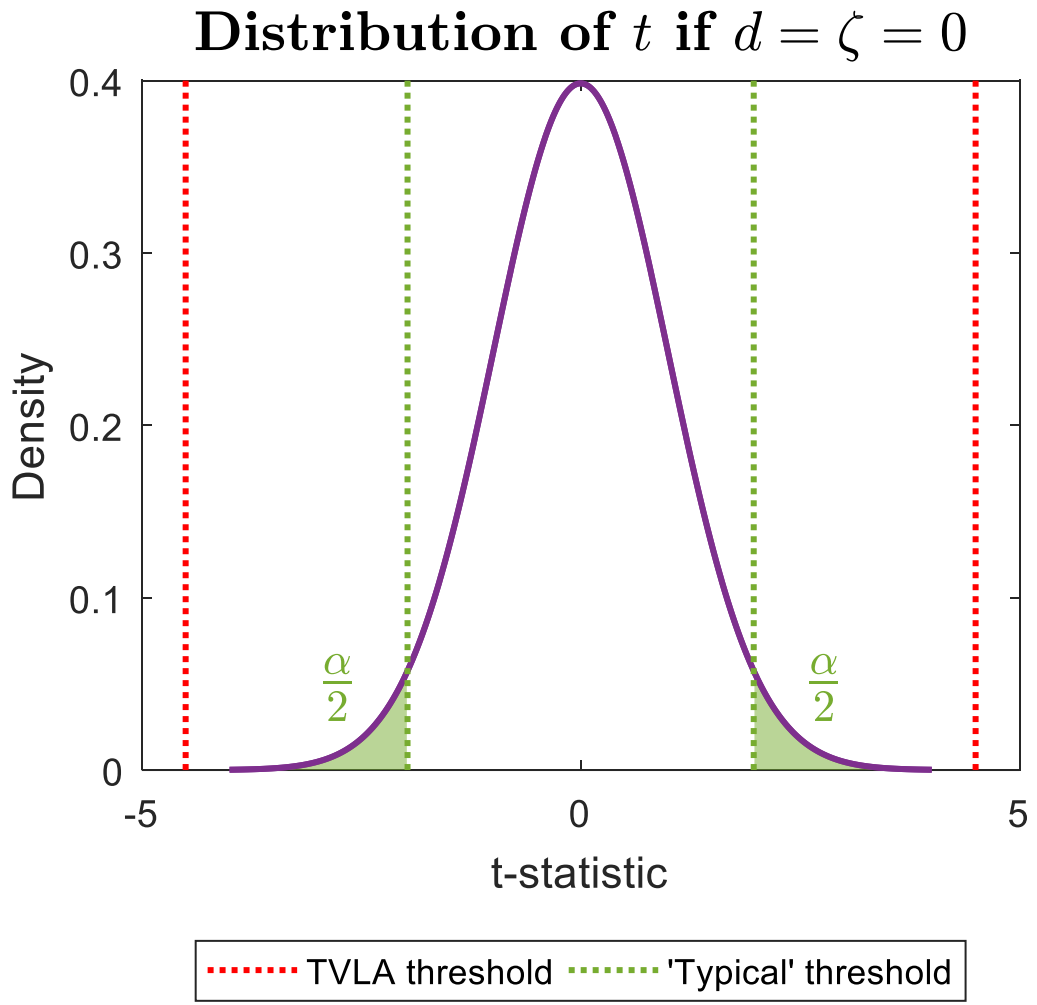
$$t = \frac{mean(random) - mean(fixed)}{\sqrt{\frac{var(random)}{N_{random}} + \frac{var(fixed)}{N_{fixed}}}}$$

Critical values



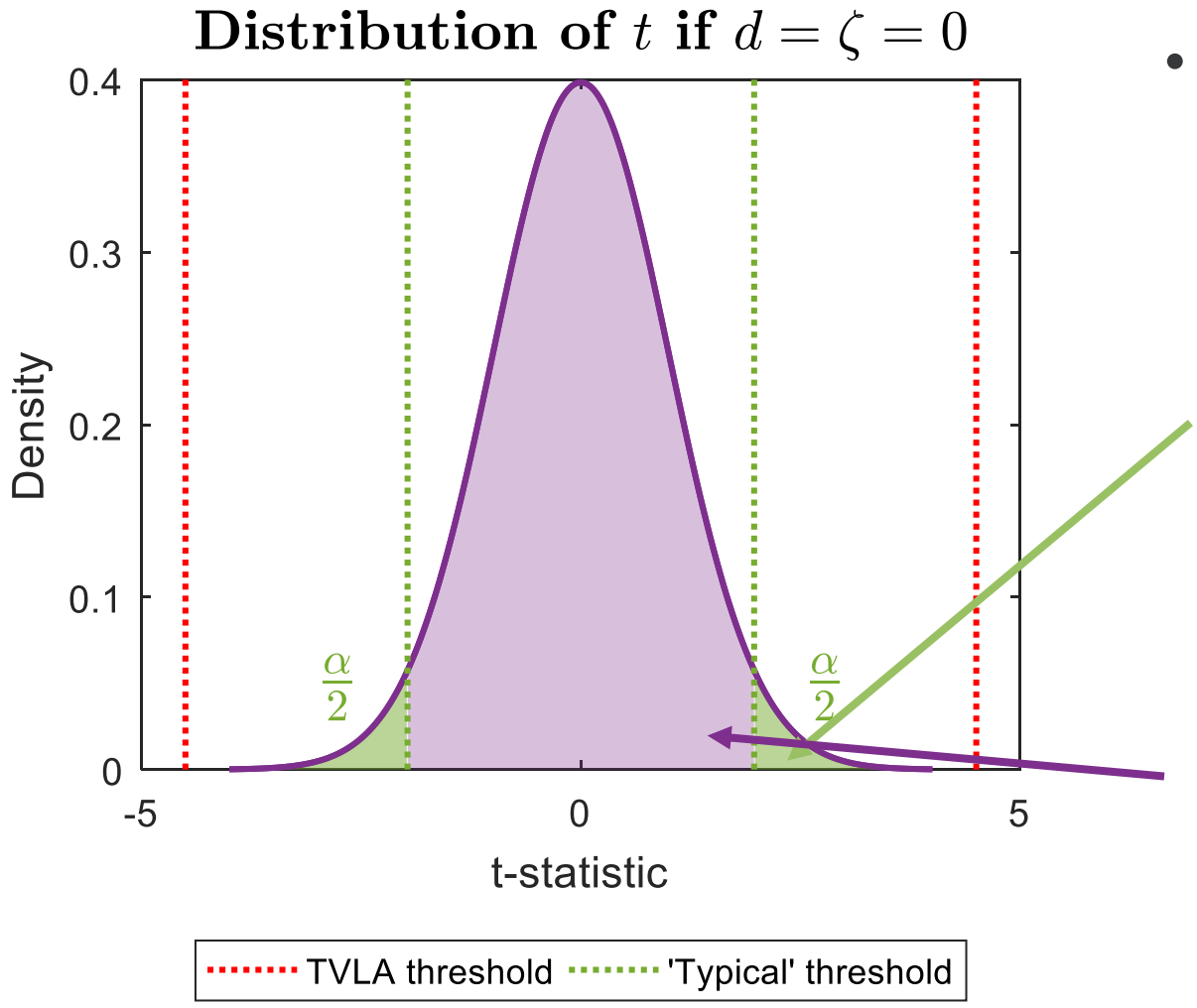
- *t* has a known distribution in the case that the null hypothesis is true. So…
- If the value of *t* that we compute from the data is much bigger or smaller than we expect under this distribution, we conclude that ζ is non-zero, (i.e. that there *is* leakage).
- What do we mean by ‘much bigger or smaller’? How do we set the decision criteria?
- The threshold of 4.5 that we took for granted according to the TVLA framework is a *critical value*. It is chosen to minimise the risk of a certain type of wrong decision.

Controlling false positives



- If the null hypothesis is true, the probability that *t* will be between -4.5 and 4.5 is about 0.99999.
- So the probability that the null hypothesis (of no leak) is true *but* that we decide from the data that it isn't, is very small (α = 0.00001).
- This probability is called the *significance level* and is fixed by the analyst to control the rate of *false positives* (aka *Type I errors*) at some acceptable level.
- In most statistical applications a much larger significance level is considered acceptable, typically α = 0.05.

However…

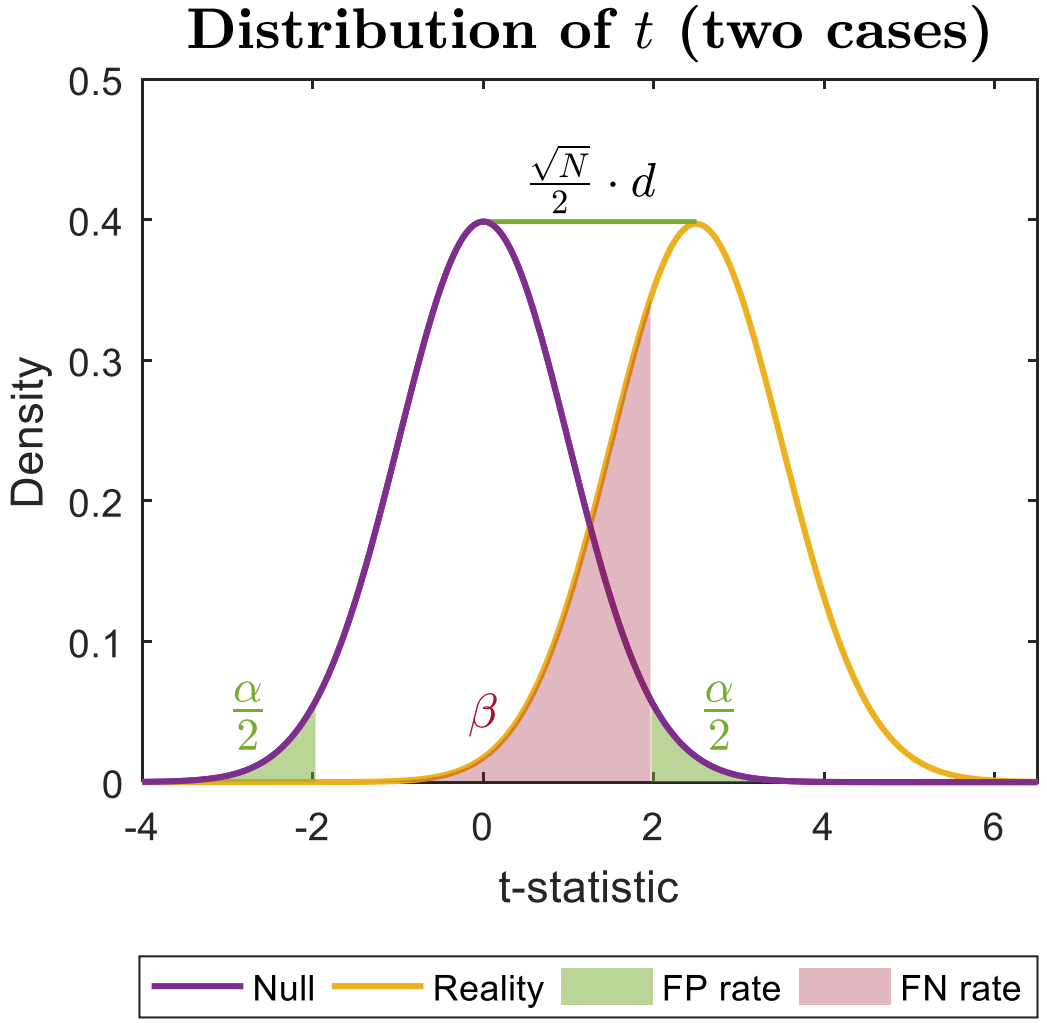


- If the observed value of *t* is *not* larger in magnitude that the derived critical value, *this does not prove the absence of leakage.*

  Conclusion: "The evidence supports the decision to reject the null hypothesis at significance level α (equivalently, with *confidence* 1 - α)".

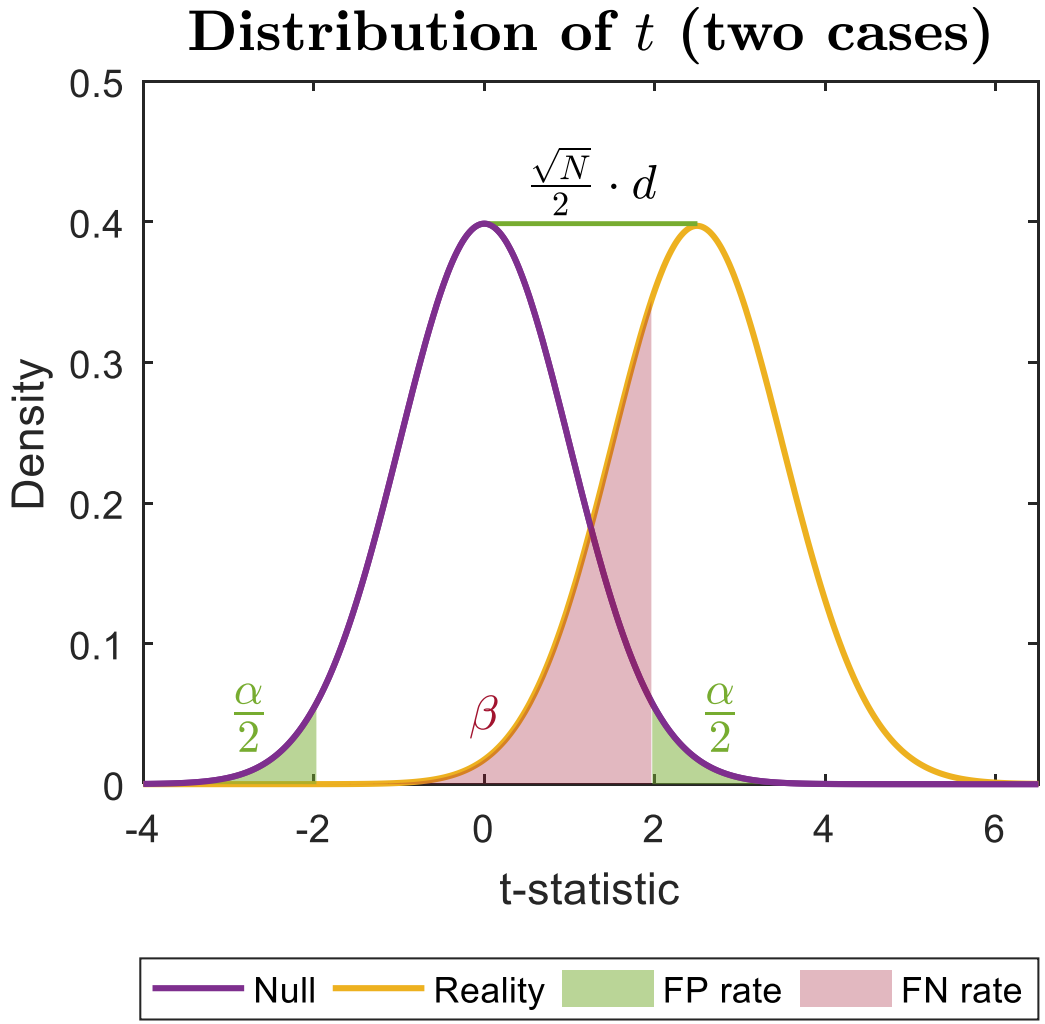
  Conclusion: "There is not enough evidence to reject, at significance level α, the null hypothesis that the two sampled distributions have the same population mean."

False negatives are also concern



- We have 'controlled' the probability of a false positive but we haven't even thought about the possibility of a *false negative*.
- A false negative happens when the true (unknown) effect size is non-zero but there isn't enough evidence to confidently decide this.
- If *t* doesn't follow the null (zero-mean) distribution then it follows some other distribution which depends on the 'true' effect size and the sample size.
- In this figure, the pink shaded region is the probability of *failing to reject* the null when the true standardised effect size is actually *d*.
- Note that you can make this area smaller by choosing a larger α and/or increasing the sample size *N*.

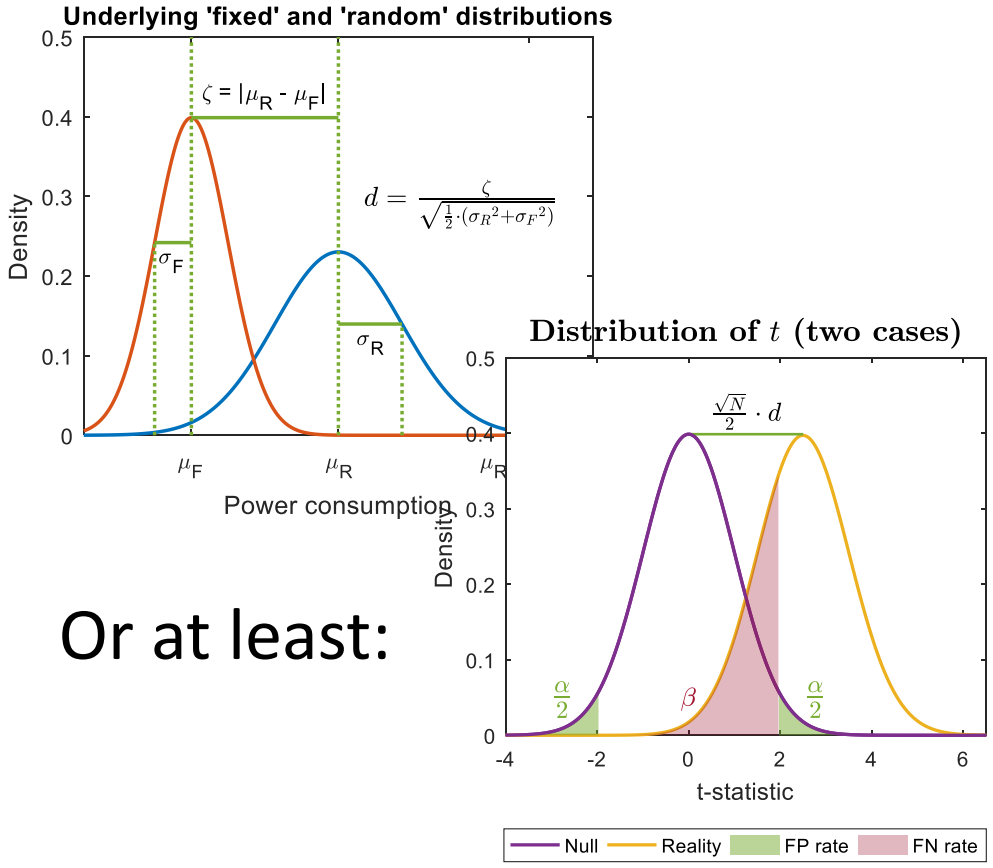
So there's a trade-off



- We know the false positive rate because we fix this in advance. But the actual false negative rate depends on the true underlying distribution.
- There's a trade-off: it's not necessarily a good strategy to set the strictest possible false positive rate, as this increases the risk of false negatives.
- Both types of errors are reduced by collecting more data.
- A well-designed statistical test achieves high 'power' as well as high 'confidence'.

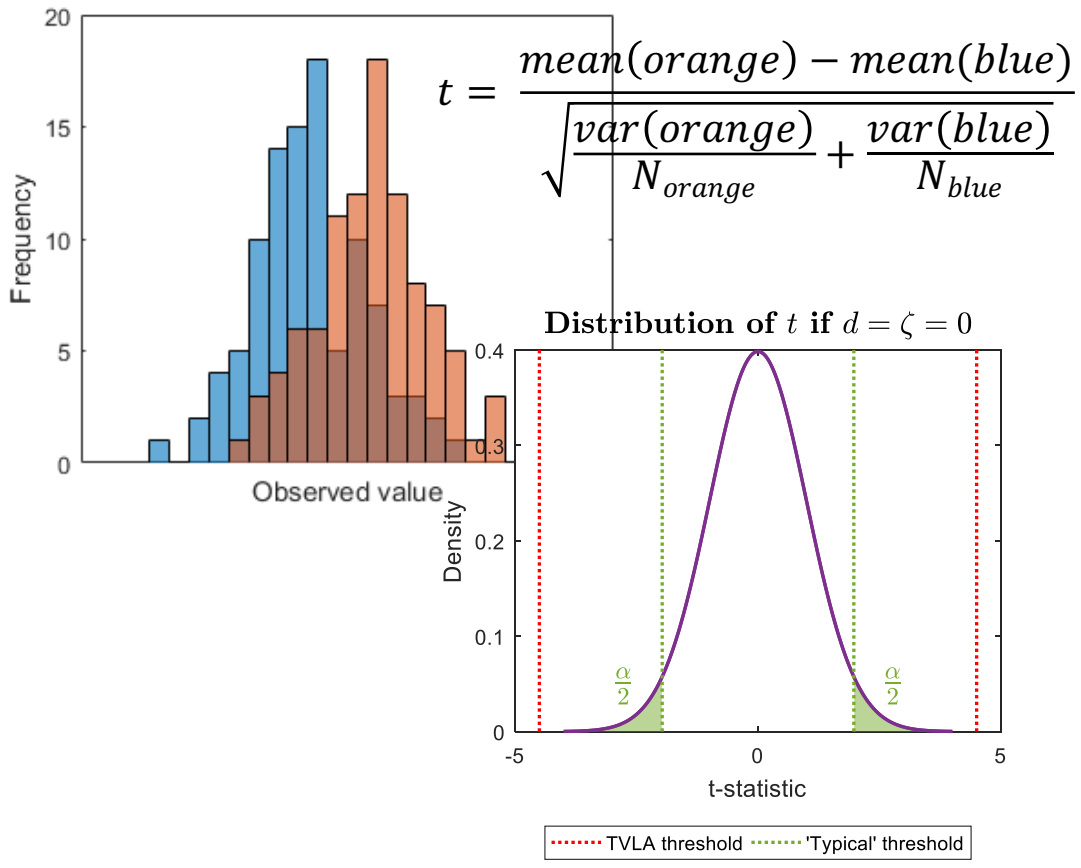
Don’t forget how much we don’t know!

What we would like to know:

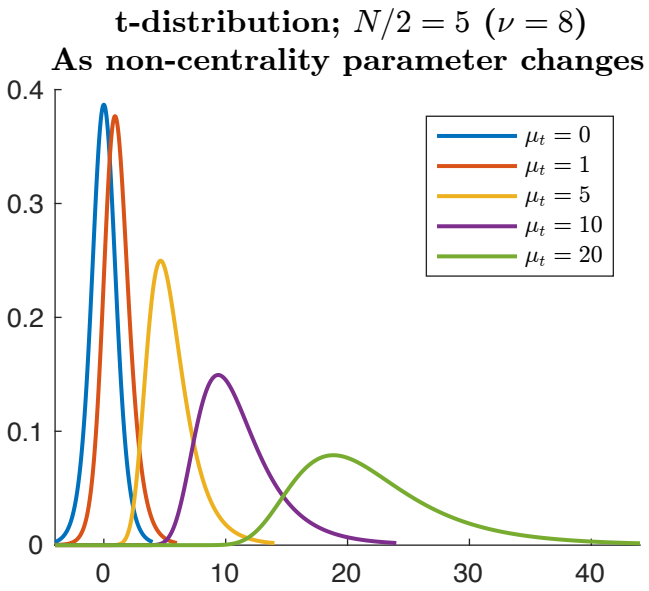


Or at least:

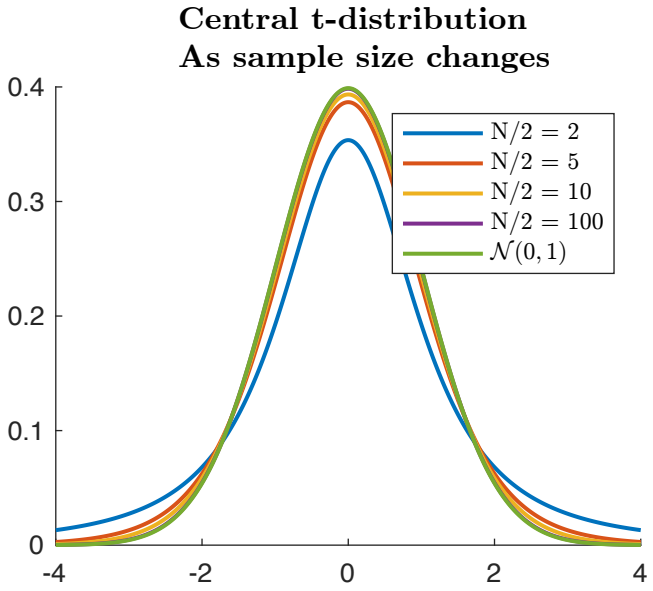
What we actually know:



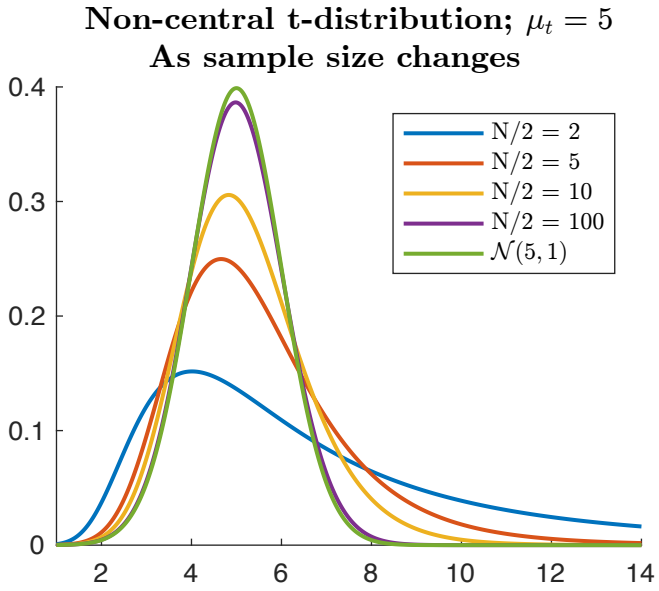
Making things a tiny bit simpler



The non-central *t*-distribution for a given sample size is more skewed and more spread out the larger the non-centrality parameter.



The central *t*-distribution tends towards the **standard normal** as the sample size increases.



The non-central *t*-distribution for a given non-centrality parameter tends towards a **normal with standard deviation 1** as the sample size increases.

So, what if a test fails to find leakage?

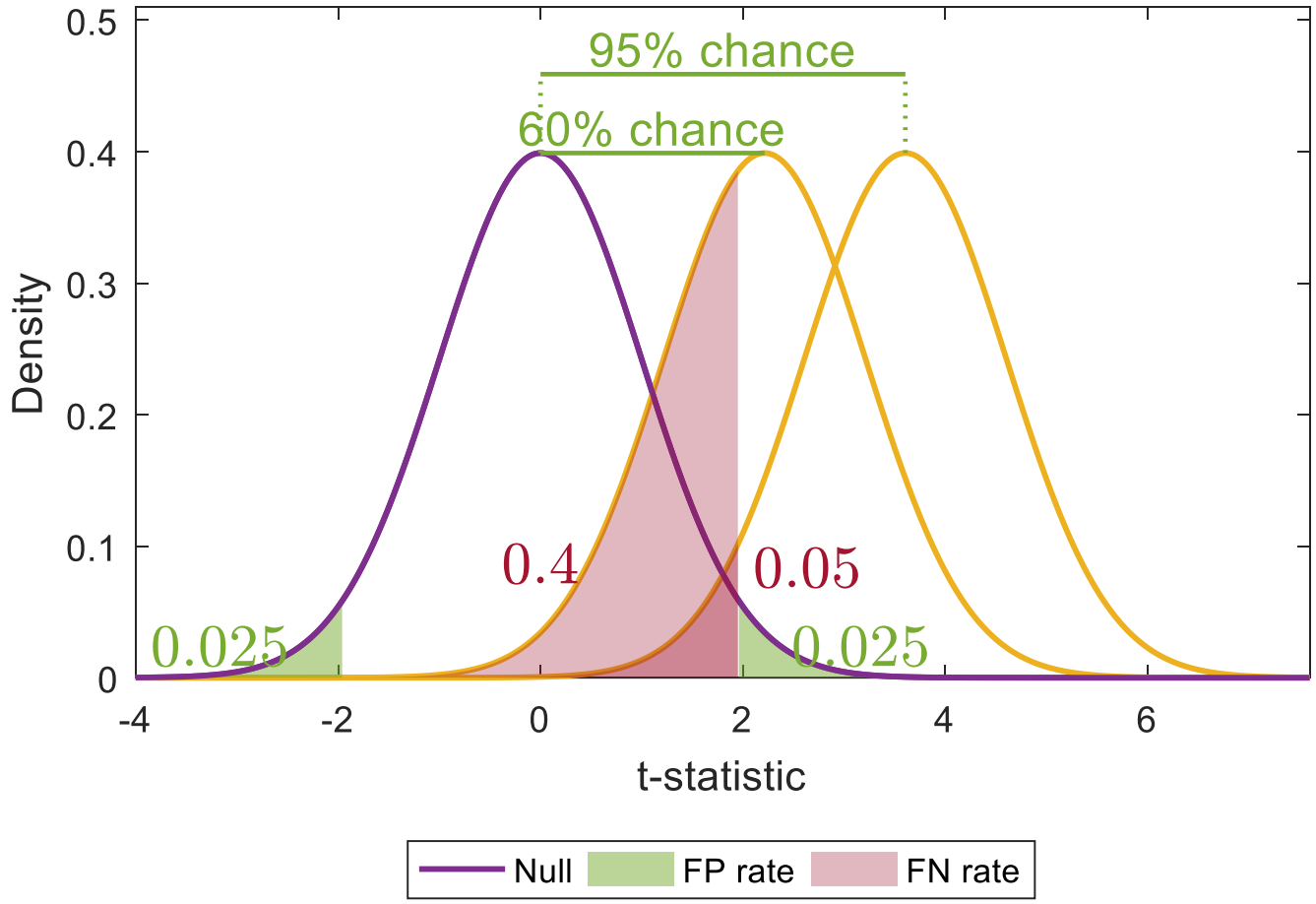
- Suppose you partition a leakage point according to a known intermediate bit *b*, and perform a *t*-test on the two groups (each of size *N/2*). Your *t*-statistic is below the threshold.
- Formally: “There is not enough evidence to reject, with confidence 1 - α, the null hypothesis that the mean power consumption does not depend on *b*.”
- This might mean:
  - The power consumption doesn’t depend on *b*.
  - The *mean* power consumption doesn’t depend on *b*, but the distributions differ in some other way.
  - The power consumption doesn’t depend on *b*, but it does depend on some other sensitive data value.
  - The power consumption doesn’t depend on *b*, but some other physically observable characteristic of the device does.
  - The difference in power consumption for *b = 0* and *b = 1* is too small for the test to detect.

Keeping sight of the intuition

- The *observed* difference is highly unlikely to be exactly zero, even if the true difference is, because of estimation error.
- The smaller the *true* difference, the more observations you need in order to be confident that a non-zero *observed* difference is more than just imprecision.
- **All we're trying to do is decide whether the observed difference is "non-zero enough" to suggest a true difference.**
- The criteria to decide this can feel cumbersome and opaque, but they're necessary if we want to make fair and consistent judgements from one evaluation to another.

Deriving the minimum effect size

One option: find out how big the effect size *would have needed to be* for the same test to detect it with reasonable probability.



Deriving the minimum effect size

There is a formula for this:

$$\zeta_{min} = \left(z_\beta + z_{\alpha/2}\right) \cdot \sqrt{\frac{2 \cdot (\sigma^2_{b=0} + \sigma^2_{b=1})}{N}}$$

Where $N/_2$ is the size of each group, z_θ is the critical value for the standard normal such that the probability of observing anything larger is θ, and $\sigma^2_{b=0}$, $\sigma^2_{b=1}$ are the variances of the two groups. In practice, $\sigma^2_{b=0}$, $\sigma^2_{b=1}$ are unknown and substituted with estimates $\hat{\sigma}^2_{b=0}$, $\hat{\sigma}^2_{b=1}$.

An example scenario

- Suppose we have 5,000 trace measurements for which bit *b* = 0, and 5,000 for which *b* = 1.
- We want no more than a 0.05 risk of a false positive.
- We know (e.g. from prior work) that the variance in each subset is 3.
- We compute the test statistic and it is smaller in magnitude than the threshold of ± 1.96.
- The effect size we could have detected with an equal probability of false positives is $\zeta_{min} = 0.125$.
- If there *is* a non-zero difference depending on *b*, chances are it is smaller than this value.

Re-running the experiment

- If values smaller than ζ_{min} are large enough to be interesting (e.g. to represent a security risk) then we will probably want to repeat the experiment with a larger sample size.
- We can work out how large the new sample needs to be by rearranging the formula for effect size and plugging in a chosen value ζ_{ch} which represents our best lower bound on 'interesting' effects.

$$N = 2 \cdot \frac{\left(Z_\beta + Z_{\alpha/2}\right)^2 \cdot \left(\sigma^2_{b=0} + \sigma^2_{b=1}\right)}{{\zeta_{ch}}^2}$$

- E.g., suppose we want to be able to detect (with equal probability of false positives and negatives, $\alpha = \beta = 0.05$) an effect one tenth of the size of ζ_{min}, i.e. $\zeta_{ch} = \frac{0.192}{10} = 0.0125$. Then we would need 1,000,000 traces total.

But what does it actually *mean*?!

- For effect sizes to be informative, we need to be able to interpret them practically. Otherwise, how do we know if an effect 'matters'?
- E.g. in other applications, an effect size might be translated into a financial cost saving, or a health outcome such as a recovery rate.
- Side-channel effect sizes are expressed in the units of the trace measurements, which depend on the set-up and any pre-processing.
- *Ideally* we'd want to translate this into a security cost, e.g.: "an effect of size ζ corresponds to a successful attack with M traces".
- Typically, we don't know how to do this, and we are obliged to look for 'expected effects' (as learnt from previous experiments) rather than 'practically relevant effects'.

Standardised effect sizes

- Granted that interpretation remains a challenge, it would at least be useful to be able to compare effect sizes from one side-channel scenario to another.
- This brings us back to Cohen's *standardised effect size,*

$$d = \frac{\zeta}{\sqrt{\frac{1}{2}\cdot\left(\sigma^2_{b=0}+\sigma^2_{b=1}\right)}}.$$

- (The denominator is the pooled population standard deviation under the assumption of equal samples).
- In our running example, the minimum detectable d is $\frac{0.125}{\sqrt{3}} = 0.0722.$

Classifying standardised effect sizes

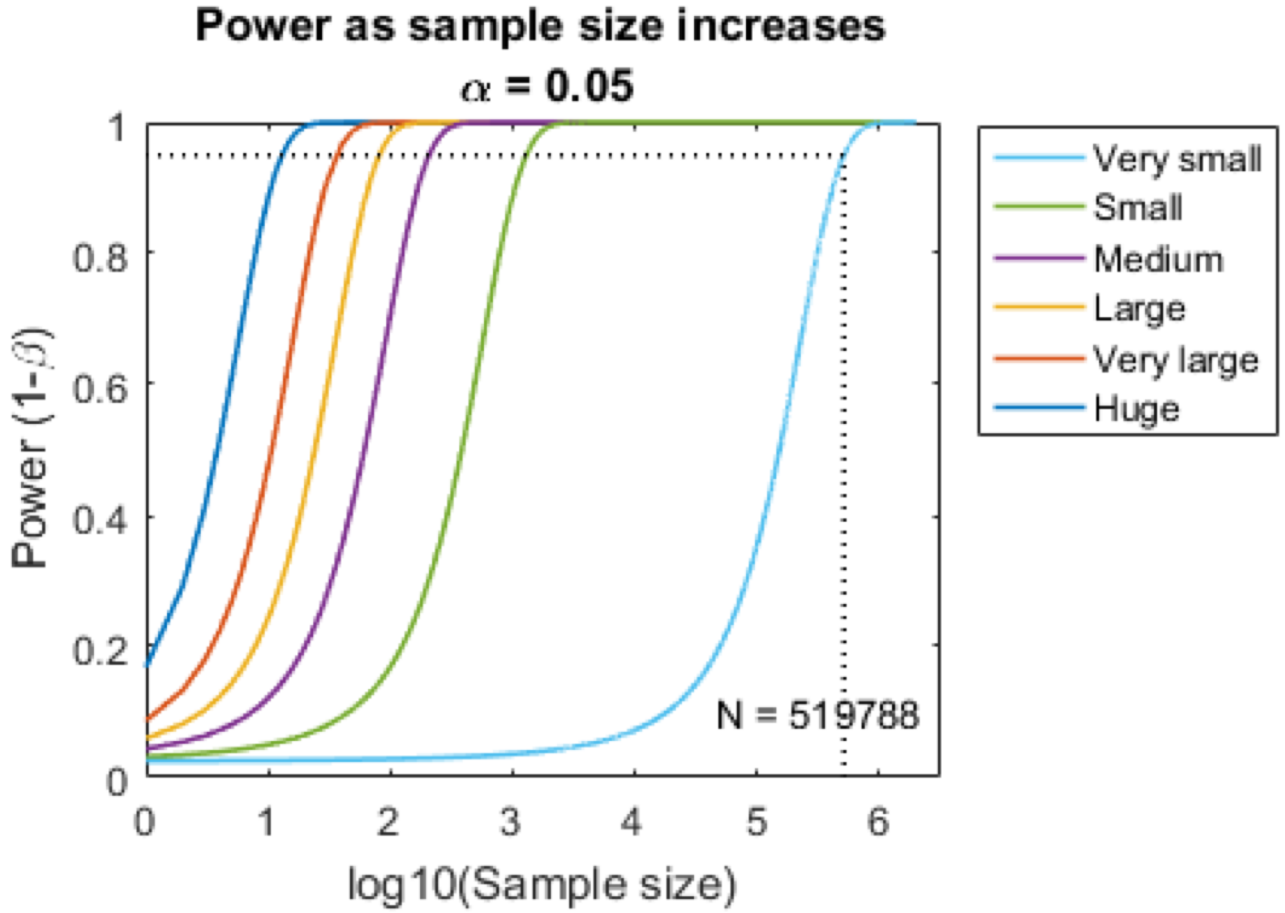
Effect size	d	Reference	Raw equivalent in our example
Very small	0.01	[Saw2009]	0.017
Small	0.2	[Coh1988]	0.346
Medium	0.5	[Coh1988]	0.866
Large	0.8	[Coh1988]	1.386
Very large	1.2	[Saw2009]	2.079
Huge	2.0	[Saw2009]	3.464

- In a side-channel context we typically encounter (and care about) 'very small' effects
- E.g. samples of size 10,000 from both an ARM board and an 8051 microcontroller produced estimated standardised effects of 0.04.
- Comparatively cheap and easy to acquire large samples of leakage traces (cf. applications requiring survey respondents, historic records, or experimental subjects, e.g.).

[Coh1988] J. Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 1988.
[Saw2009] S. S. Sawilowsky. *New effect size rules of thumb*. Journal of Modern Applied Statistical Methods, 8(2):597–599, 2009.

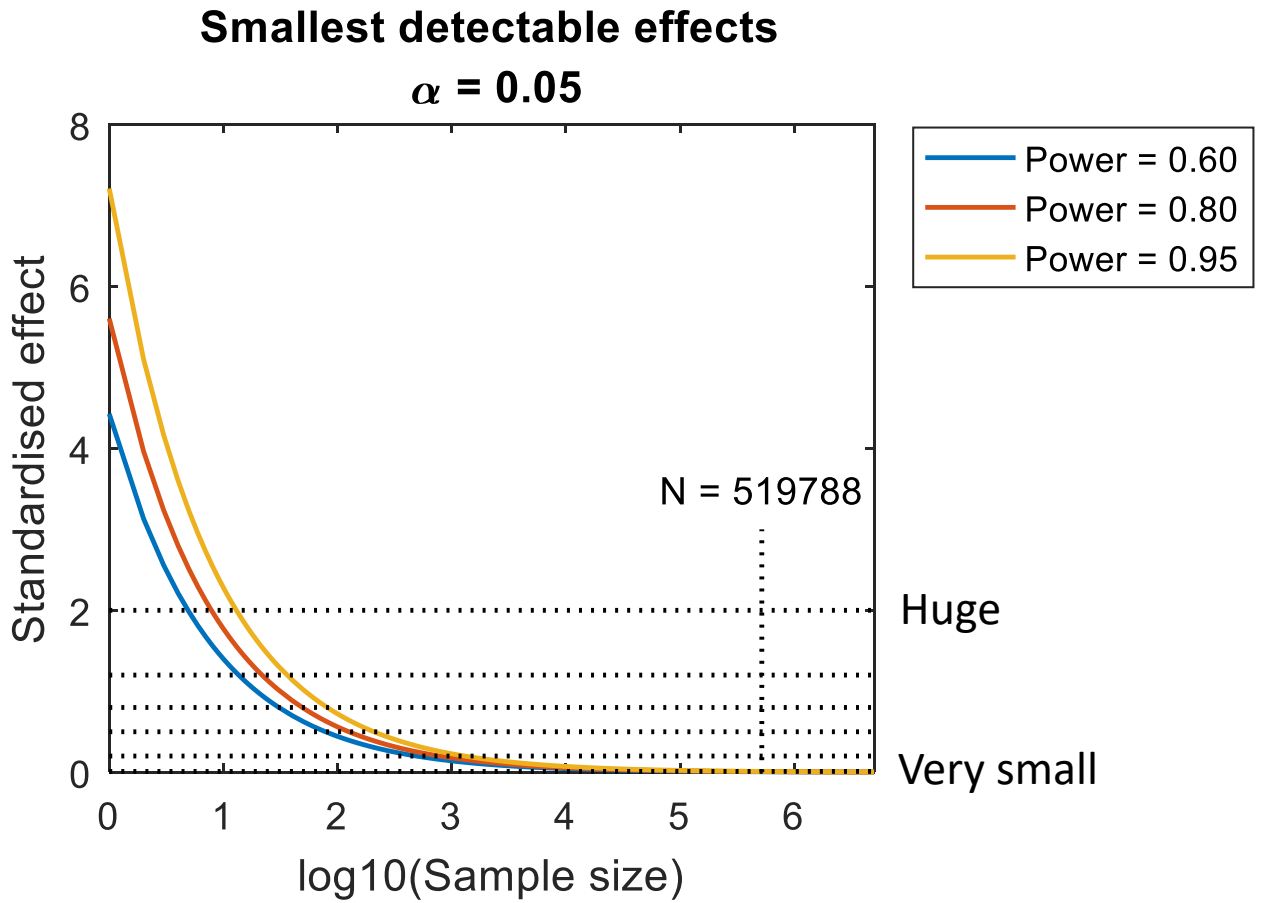
Effect size, sample size and error rates

- The formulae connecting effect size, sample size and error rates can be used to derive any one of these holding the others fixed.
- E.g. the sample size needed to detect a 'very small' effect with $\alpha = \beta = 0.05$ (i.e. power and confidence of 95%) is around 52k.



Effect size, sample size and error rates

- The relationship can alternatively be visualised by plotting the effect size as the sample size increases.
- Now the required sample size corresponds to where the 95% power test reaches the ‘very small’ threshold.



Planning a test with the desired properties

- Ad-hoc approach:
  - Fix the rate of false positives α to be arbitrarily small.
  - Collect as large a sample as possible.
  - Hope for the best!
- Half-way approach:
  - If the above fails to detect leakage, compute the minimum detectable effect size for a reasonable false negative rate β and use it to reason about the outcome.
- Advisable approach:
  - Decide on the effect size of interest d (based on previous experiments or expected capabilities of the adversary).
  - Choose a tolerable rate of false positives α and false negatives β.
  - Compute the sample size needed to detect the effect size of interest with the desired error rates.
  - (If the sample size is infeasibly large, revise the scope of your evaluation).

How to choose the desired error rates?

- Tendency in the literature: consider only false positives and set a very small significance level, α.
- But: the smaller the α, the higher the burden of proof. This may not always reflect the priorities of the evaluation.
  - It might sometimes be just as bad (or worse) to miss a leak that *is* there as to find one that isn't.
  - Small but real differences can really matter in side-channel analysis: with the right attack they can lead to the same amount of information learned as a large difference.
  - It is important to clarify priorities beforehand and choose parameters with these trade-offs in mind.

Table: Sample sizes required for various different test specifications

	$\alpha = 0.05$			$\alpha = 0.00001$ (TVLA)		
Effect size	$\beta = 0.2$	$\beta = 0.05$	$\beta = 0.00001$	$\beta = 0.2$	$\beta = 0.05$	$\beta = 0.00001$
Very small	314,000	520,000	1,550,000	1,107,000	1,470,000	3,016,000
Small	800	1,300	3,900	2,800	3,700	7,600
Medium	130	210	620	450	590	1,210
Large	50	90	250	180	230	480
Very large	22	37	108	77	103	210
Huge	8	13	39	28	37	76

E.g. to balance the errors in the TVLA framework requires around 6 times as many data samples as a typical statistical application, whatever the effect size.

What does this look like in practice?

- Simulating traces allows us to test our analysis tools against a known (because *under our control*) leakage scenario.
- We can simulate (univariate) traces with a (standardised) effect size d via the following pseudocode:

```
simtrace = d*intvals[b] + randn(0,1)
```

- We are going to plan a test with a 'very small' effect in mind (d = 0.01) and see how it performs when the true standardised effect is 0.01 and when it is in fact zero.

Table: Sample sizes required for various different test specifications

	$\alpha = 0.05$			$\alpha = 0.00001$ (TVLA)		
Effect size	$\beta = 0.2$	$\beta = 0.05$	$\beta = 0.00001$	$\beta = 0.2$	$\beta = 0.05$	$\beta = 0.00001$
Very small	314,000	520,000	1,550,000	1,107,000	1,470,000	3,016,000
Small	800	1,300	3,900	2,800	3,700	7,600
Medium	130	210	620	450	590	1,210
Large	50	90	250	180	230	480
Very large	22	37	108	77	103	210
Huge	8	13	39	28	37	76

We expect to need 3,016,000 (simulated) measurements to detect the leakage under the TVLA criteria (requiring balanced errors), or 520,000 under the less demanding criteria typical of many other applications.

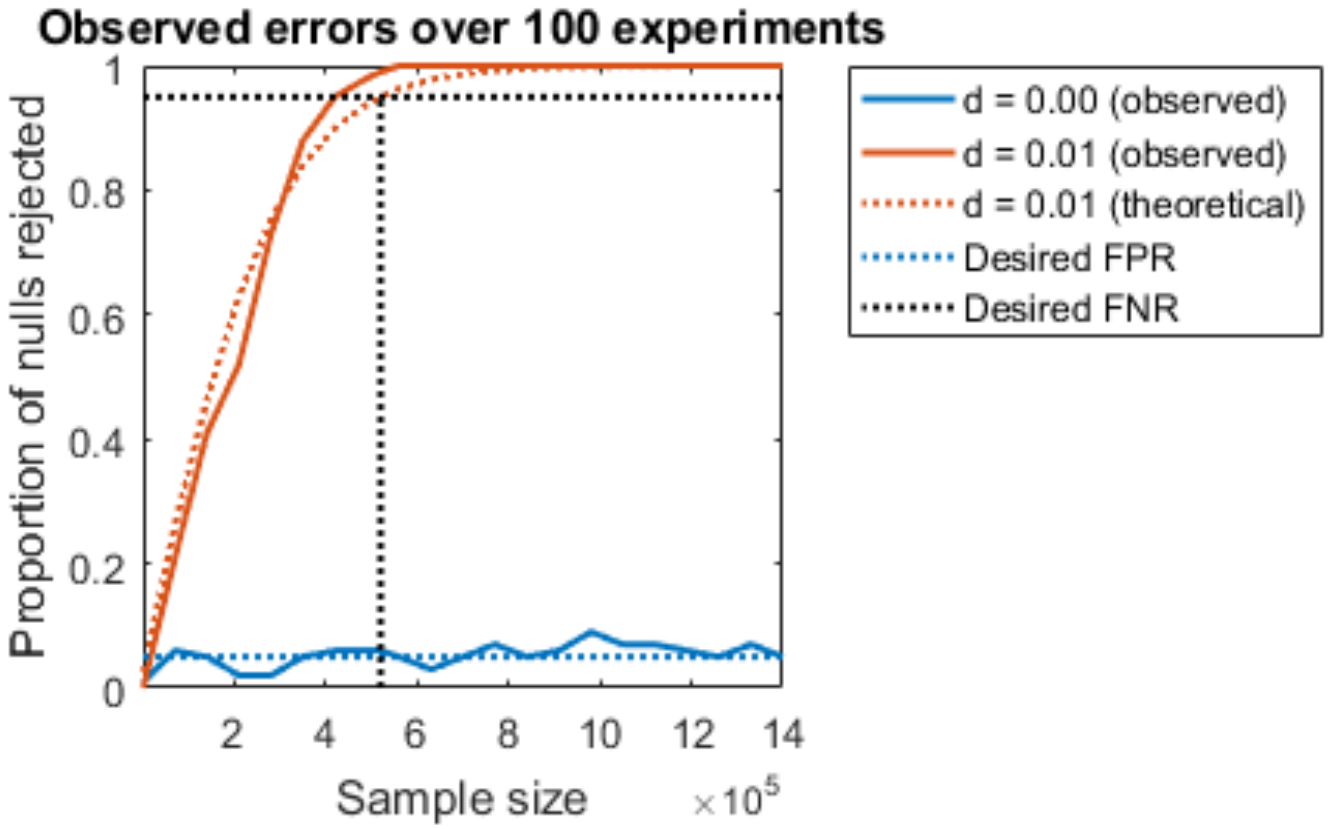
Results for a single experiment

Desired error rates	True effect	Sample size	t-statistic	Threshold	Reject the null?	Conclusion
$\alpha = \beta = 0.05$	Zero	1,040k	0.542	1.96	No	Test with 95% power does not find leakage
	Very small	520k	4.527	1.96	Yes	Leakage detected
$\alpha = \beta = 0.00001$	Zero	6,032k	0.562	4.42	No	Test with 99.99% power does not find leakage
	Very small	3,016k	9.188	4.42	Yes	Leakage detected

We get the outcomes that we expect in each case, but what we *don't* get from a single experiment is any sense of the error rates. We would have to repeat the experiments a large number of times to obtain this – which is something that we can do, if we want, in this simulated environment...

‘Verifying’ the error rates

We can see that the observed detection and false detection rates are well aligned with our theoretical expectations. However, this is not something we can or would do in a real evaluation scenario. *It is precisely because we can’t observe error rates in practice that we need reliable theory to help us control them.*



Multiple comparisons

- Hopefully, you are starting to get the hang of the idea that there are two types of errors, and that we need to design tests which reflect how strongly we care about each of them.
- We have shown how to do this for individual tests against selected points in a trace.
- Unfortunately, things are about to get a lot more complicated…
- Real trace acquisitions comprise large numbers (typically 1000s) of trace points; detection test are applied to each point separately to see if there is leakage *anywhere* and to locate it correctly.
- It turns out that this impacts on error rates in a very undesirable way.

Inflated rate of false positives

Under the simplifying assumption that the tests are independent, the overall probability of a false positive is:

$$\alpha_{overall} = 1 - \left(1 - \alpha_{per-test}\right)^{L}$$

Where $\alpha_{per-test}$ is the per-test significance level and L is the length of the trace (i.e. the number of tests).

$\alpha_{per-test}$	$L = 1$	$L = 10$	$L = 100$	$L = 1{,}000$	$L = 10{,}000$	$L = 100{,}000$
0.05	0.05	0.40	0.99	1.00	1.00	1.00
0.00001	0.00001	0.00010	0.00100	0.00995	0.09516	0.63212

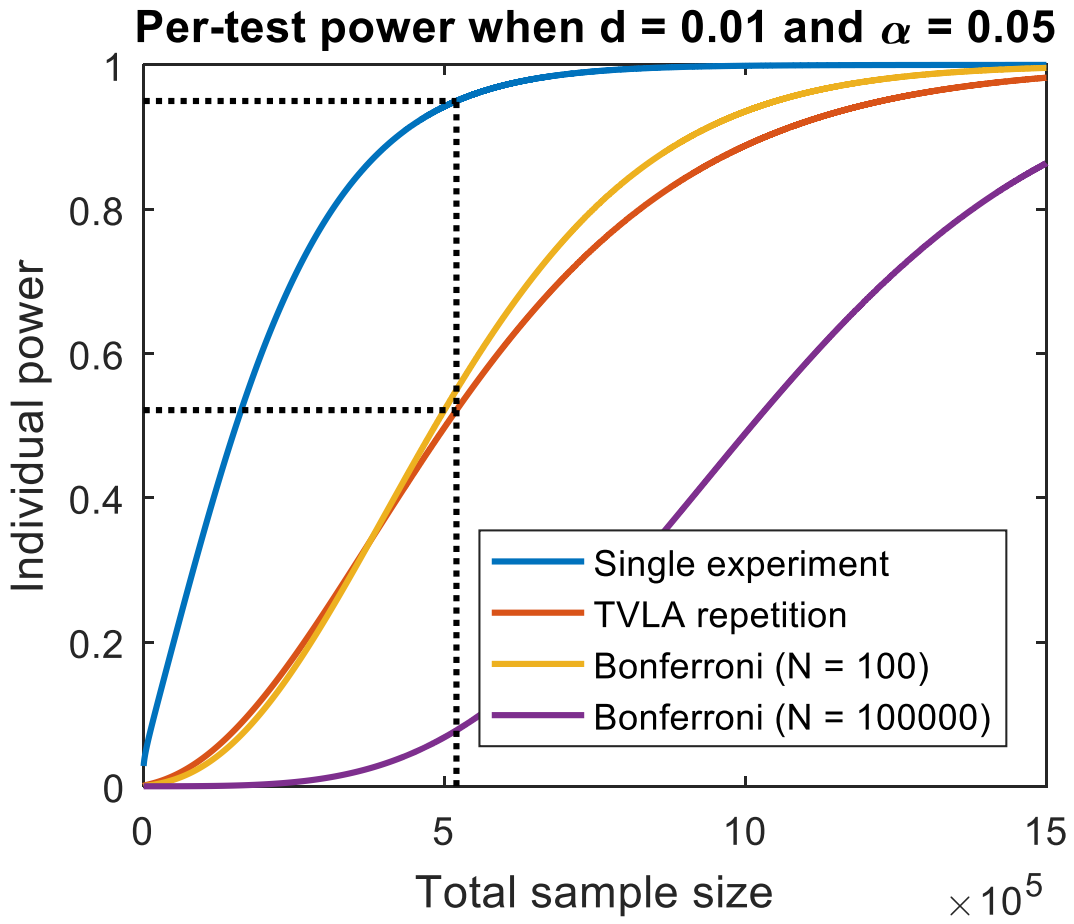
Choosing $\alpha_{per-test} = 0.05$, as is popular in many statistical applications, is disastrous for overall false positive rates in realistic leakage evaluation scenarios.

Some simple ‘fixes’

- Bonferroni correction: adjust the per-test criterion in such a way that the overall error rate is controlled as desired, $\alpha_{per-test} = {}^{\alpha_{overall}}/_{L}$.
- The very small α implied by the TVLA recommendations could be viewed as an *ad-hoc correction*, implicitly controlling $\alpha_{overall} < 0.05$ for trace lengths up to around 5,000 (without adjustment).
- TVLA requirement to *repeat* the detection procedure on a second independent acquisition is another implicit guard against false detections, effectively squaring the per-test rate so that the overall rate is $\alpha_{overall} = 1 - \left(1 - \alpha_{per-test}^2\right)^L$.

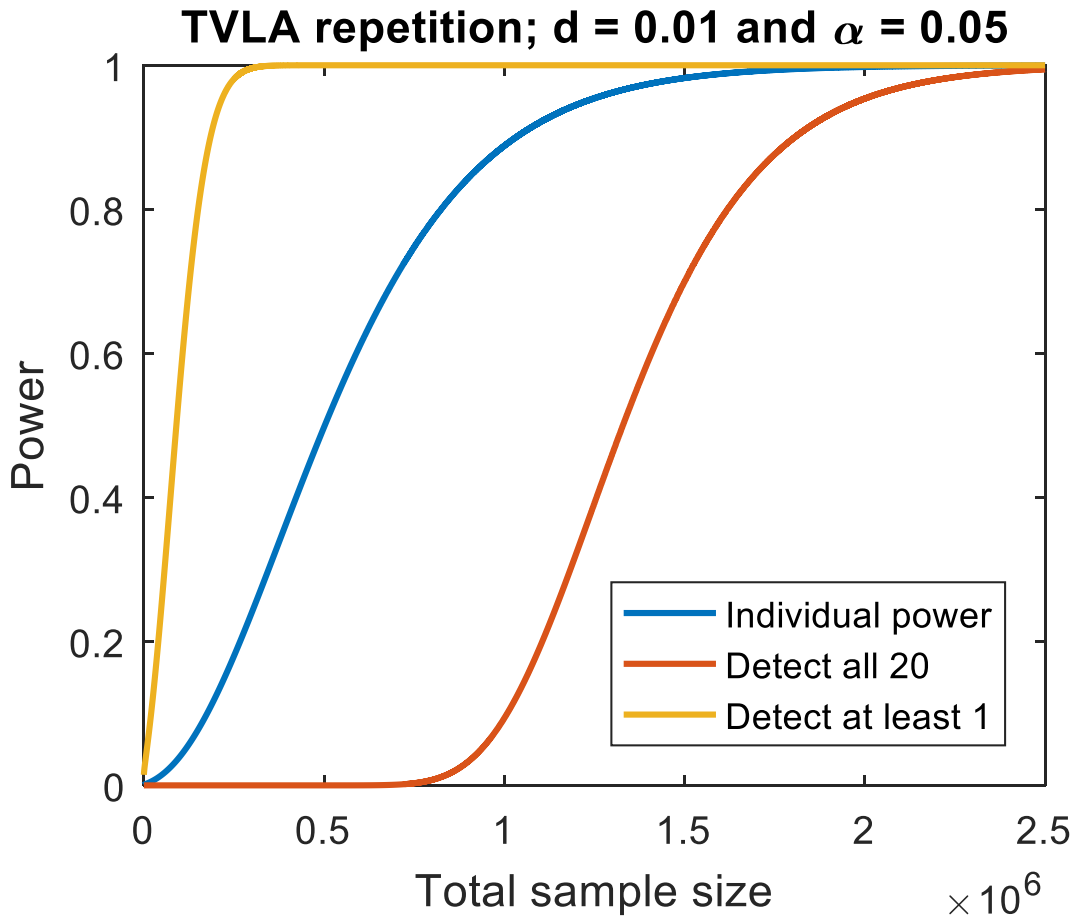
$\alpha_{per-test}$	$L = 1$	$L = 10$	$L = 100$	$L = 1{,}000$	$L = 10{,}000$	$L = 100{,}000$
0.05	0.0025	0.0247	0.2214	0.9182	1.0000	1.0000
0.00001	1.0E-10	1.0E-09	1.0E-08	1.0E-07	1.0E-06	1.0E-05

Impact on false negatives



- It should be clear from our earlier discussion that measures to correct for inflated false positives necessarily induce an increase in false negatives.
- The independence assumption leads to conservative significance criteria and unnecessarily penalised detection rates. Better understanding about the dependence structure could lead to better trade-offs (outside of scope for today).

Different notions of ‘overall’ detection rate



- The impact of multiple testing depends on the notion of ‘overall detection rate’ that we care about.
- Probability of detecting *all* true leaks will be even lower than the power of each individual test.
- Probability of detecting *at least one leak* grows as the number of true leaks increases, so that testing more points has the potential to help as well as harm the detection rate by that definition.

Concluding remarks

- Ignore false negatives at your peril: naïve applications of the TVLA framework risk prioritising the avoidance of false positives at severe cost to the test's ability ('power') to detect true leakages.
- Be realistic about the goals of your evaluation: it is possible to configure a test to be relatively confident of finding at least one leak, but TVLA cannot be used to 'find all leaks'.
- Don't claim more than your test really shows: failure to find a leak does not prove the absence of leaks. You can never show the strength of a device against an attack, you can only show (or fail to show) the weakness of it.