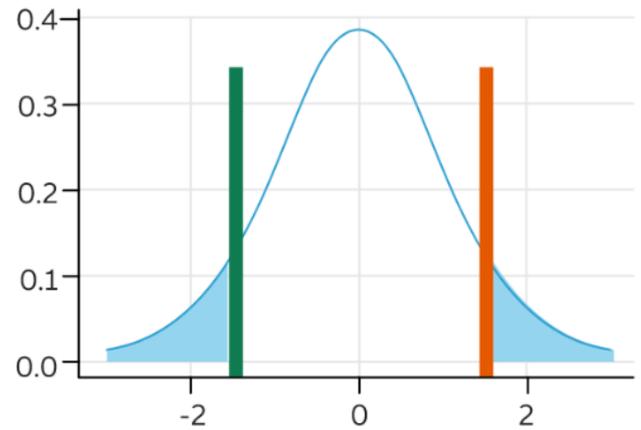
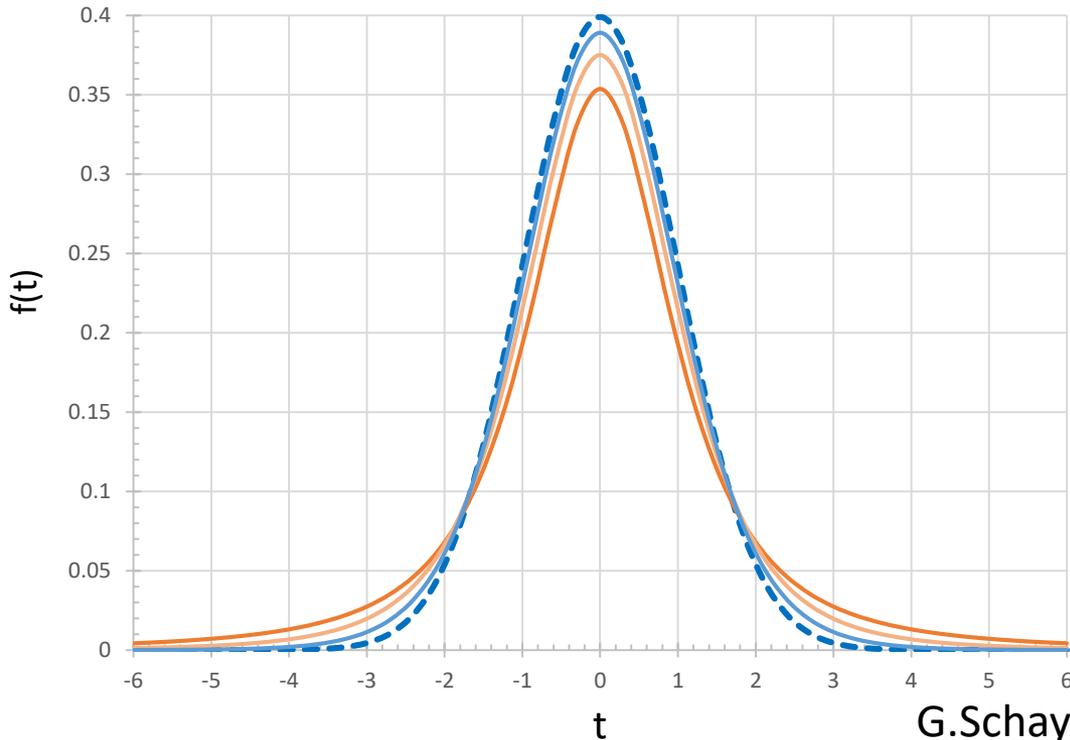
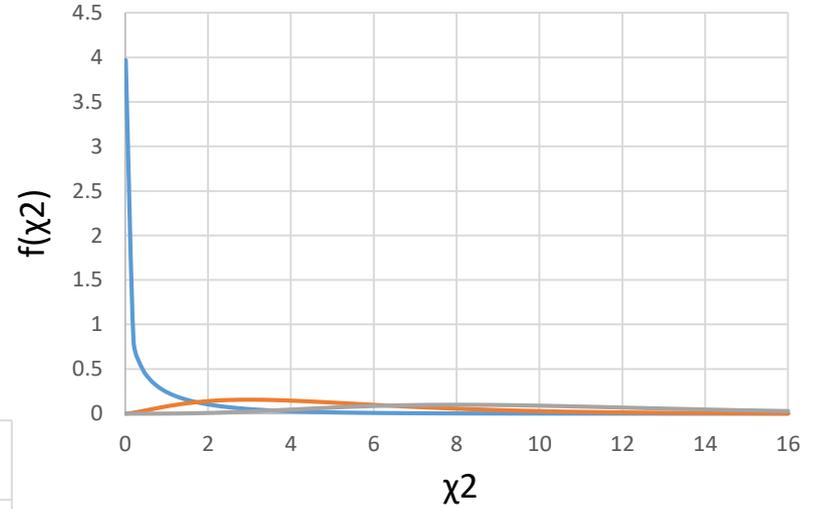
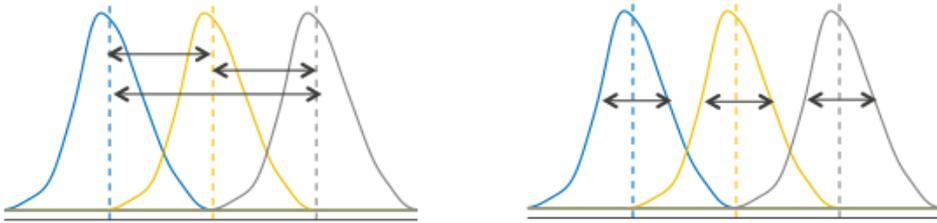


Hypothesis testing in practice

Student t-test, χ^2 -test, multiple comparisons

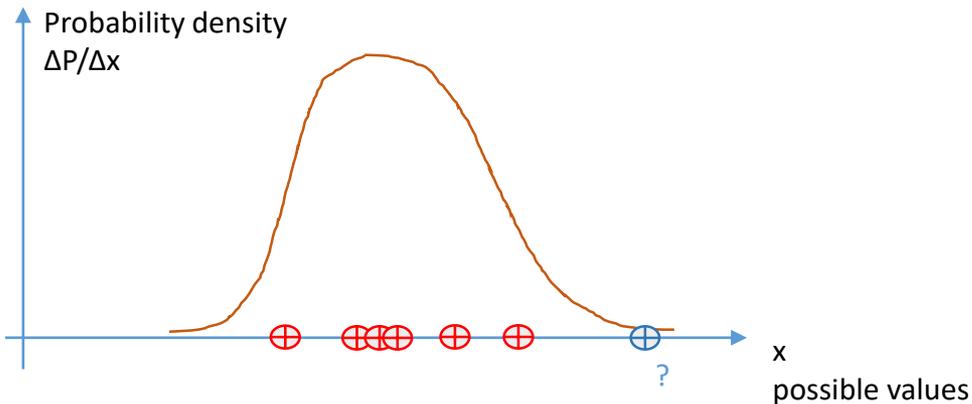


Dissect every (real) problem into dichotomy Yes/No questions

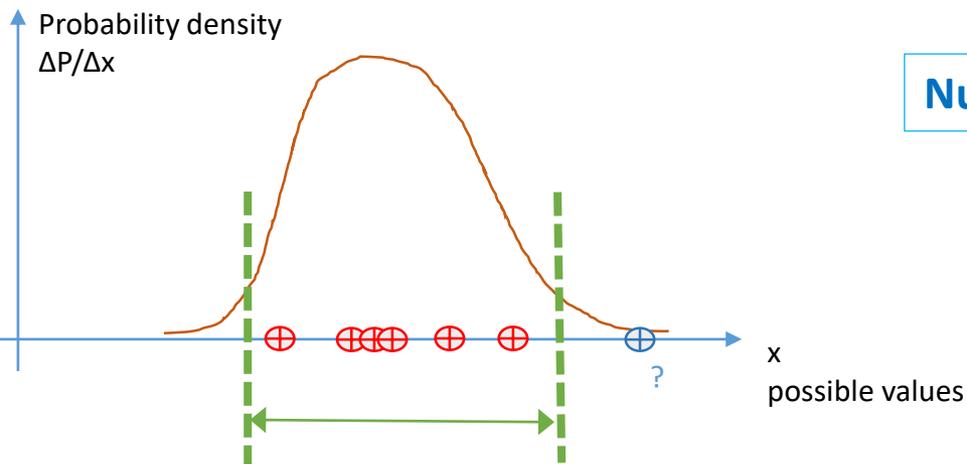
Compare the sample to **known, reference** data and distribution!

The question to answer is now: „Do the data what we got from the experiment (observation), **match the assumed distribution?**”

Null-hypothesis (starting assumption)



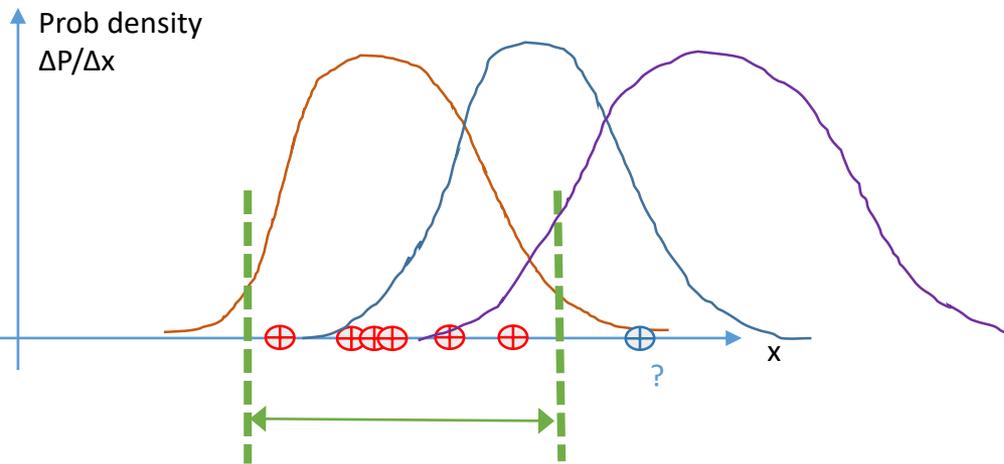
The question to answer is now: „Do the data what we got from the experiment (observation), **match the assumed distribution?**”



Null-hypothesis (starting assumption)

e.g.: „If we have measured as a mean blood glucose level of **6.2** (mmol/l) then does our patient have diabetes?”

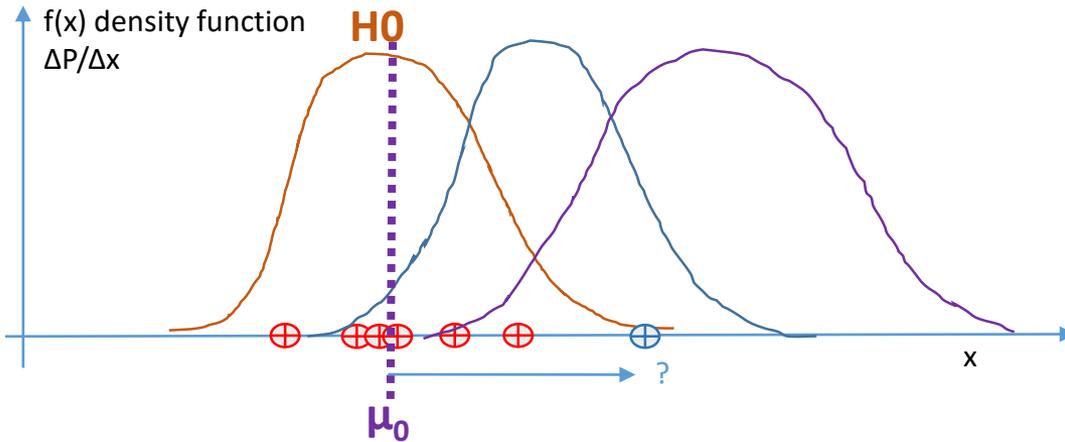
1. Y/N question: diabetes Y/N?
2. Starting assumption: the patient belongs to the well known reference, in this case the healthy normal range. (**4.0-5.9** mmol/L)



The answer is NOT unambiguous!

Our value may belong to multiple, even unknown distributions!

All we can say, is the probability of belonging to the known reference...

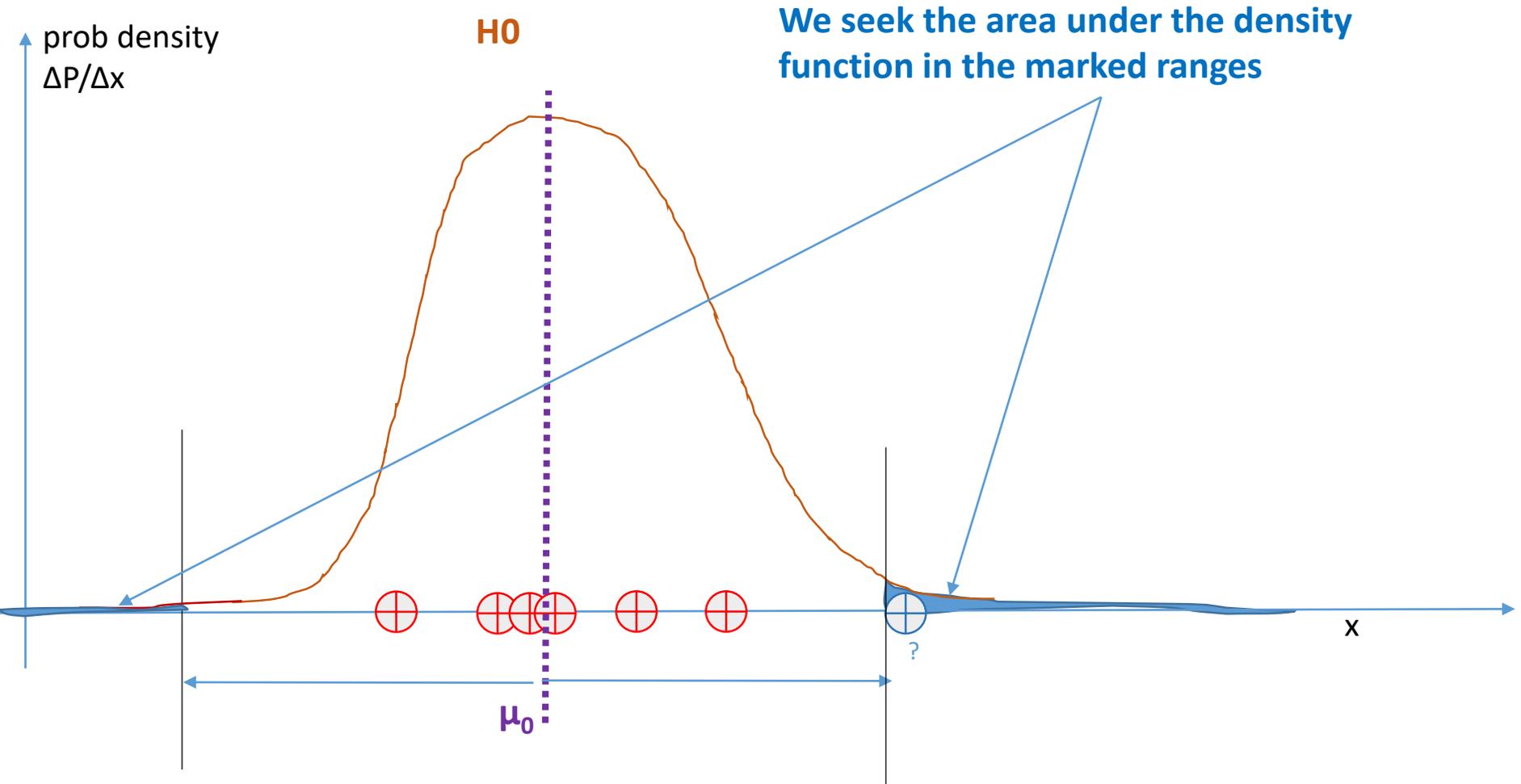


We specify the question in a calculatable way:

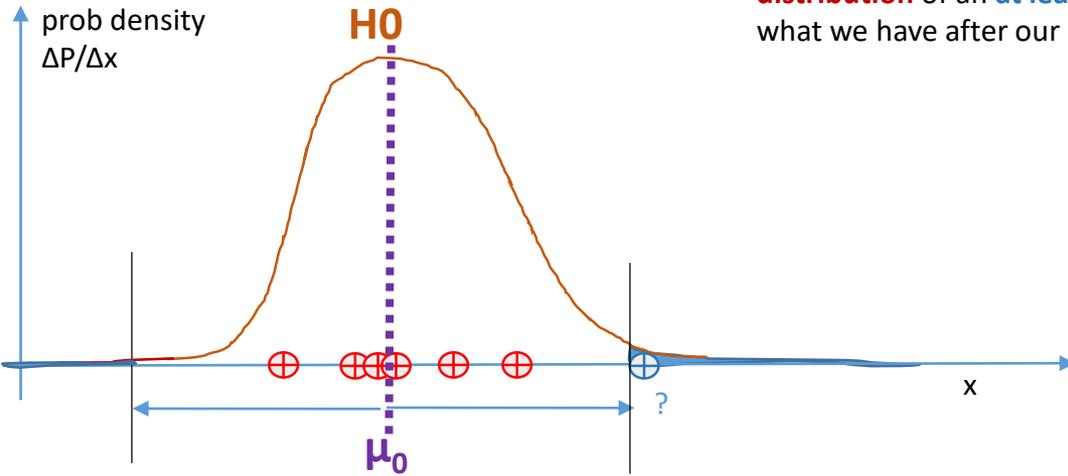
What is the probability **given that the random variable we got belongs to the H_0 distribution** of an **at least such sized deviation** from the expected value of H_0 (μ_0) as what we have after our observation series?

This (conditional) probability can be unambiguously calculated

What is the probability **given that the random variable we got belongs to the H0 distribution** of an **at least such sized deviation** from the expected value of H0 (μ_0) as what we have after our observation series?



What is the probability **given that the random variable we got belongs to the H0 distribution** of an **at least such sized deviation** from the expected value of H0 (μ_0) as what we have after our observation series?



As a result **our decision will also be probabilistic!**

significance level

If the conditional probability is too small (below a **pre-defined threshold**) then we reject H0
If it is large enough then we keep the starting assumption (null hypothesis, H0)

The dichotomic decision tree

Question (from real life)



Transform into Y/N form (maybe dissect)

H0 setup

We use a well known situation as our assumption. (some reference)



We set up the threshold, i.e. how much we “stick to H0” -> this is called the **significance level.**

α or α_{\max}



Calculate the conditional probability $P(\text{at least such a deviation} \mid H_0)$



Decision

$P < \text{sign.}$

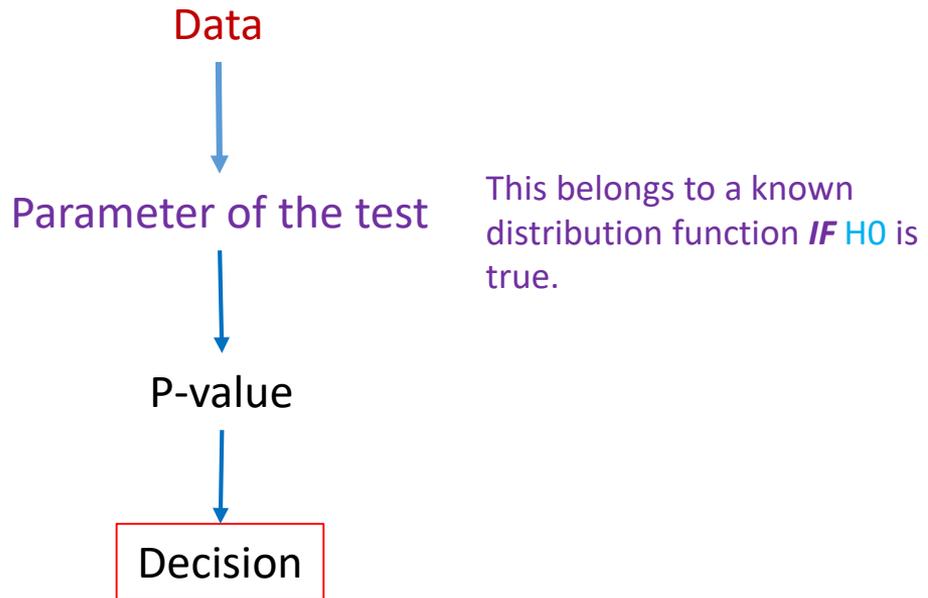
$P \geq \text{sign.}$

H0 is rejected

H0 is kept

Is not rejected

In most of the cases **we need to transform our data** to be able to compare to a known distribution. The actual mathematical transform depends on the nature of our raw data (nominal or numeric, what distribution it follows, etc.)



```
Frequency counts (test is for first level):
DIABETES
diabetes healthy
  9      91

1-sample proportions test with continuity correction

data: rbind(.Table), null probability 0.105
X-squared = 0.10641, df = 1, p-value = 0.7443
alternative hypothesis: true p is not equal to 0.105
95 percent confidence interval:
 0.04455927 0.16833464
sample estimates:
 p
0.09
```

We almost never calculate the parameter ourselves!

The software (like Excel,R,etc) does it 😊, and it yields the P-value for us to make the decision.

Is there a real difference between the mean values obtained in two series of experiments?

- A) We measure two times on the same objects (one-sample test), but in between there is something happening (e.g. therapy). This is the “before – after” type of comparison.
- B) We compare two different set of data (samples).

Mandatory conditions:

- i) The population should belong to a normal distribution.
N.B.: if we have a large sized sample then this is not mandatory, see later.
- ii) The standard deviations in the two groups are equal. if not then it is possible to correct for that: Welch-test.
- iii) The elements of the sample are independent (the distinct results do not influence each other in any way)

With i-iii being true, we can conduct a Student's **t-test**.

One sampled (paired)

Patient ID	Before treatment	After treatment
1	X1	Y1
2	X2	Y2
3	X3	Y3
...

H0: x and y data belong to the same distribution.

$$D1 = Y1 - X1$$

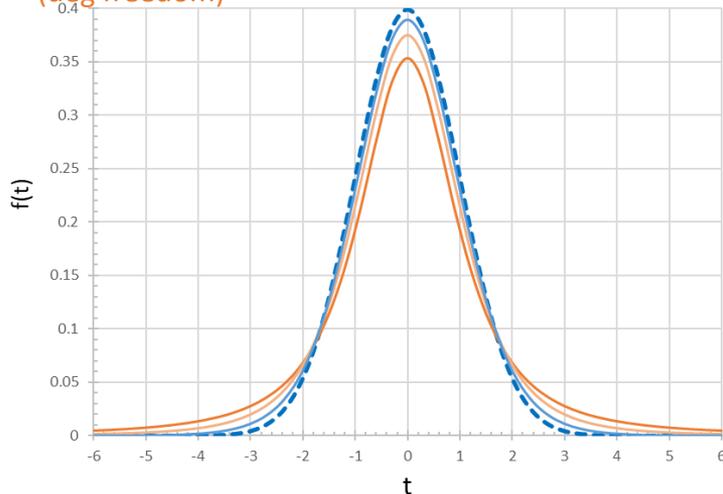
$$D2 = Y2 - X2 \quad \rightarrow$$

...

$$t_{n-1} = \frac{\bar{D}}{S_{\bar{D}}}$$

H0 → D=0 → by how many SD-s are we off?

If H0 is true then t=0, but a small deviation is acceptable. We have to use the N-1-th curve (deg freedom)



Two sampled

Group A	Group B
X1	Y1
X2	Y2
X3	Y3
...	...

H0: x and y data belong to the same distribution.

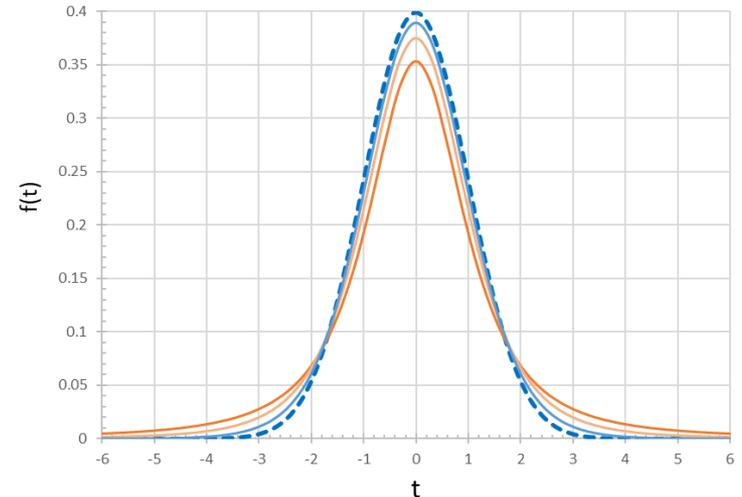
$$t \sim \frac{\bar{x} - \bar{y}}{S^*}$$

What is the relative difference of the means?

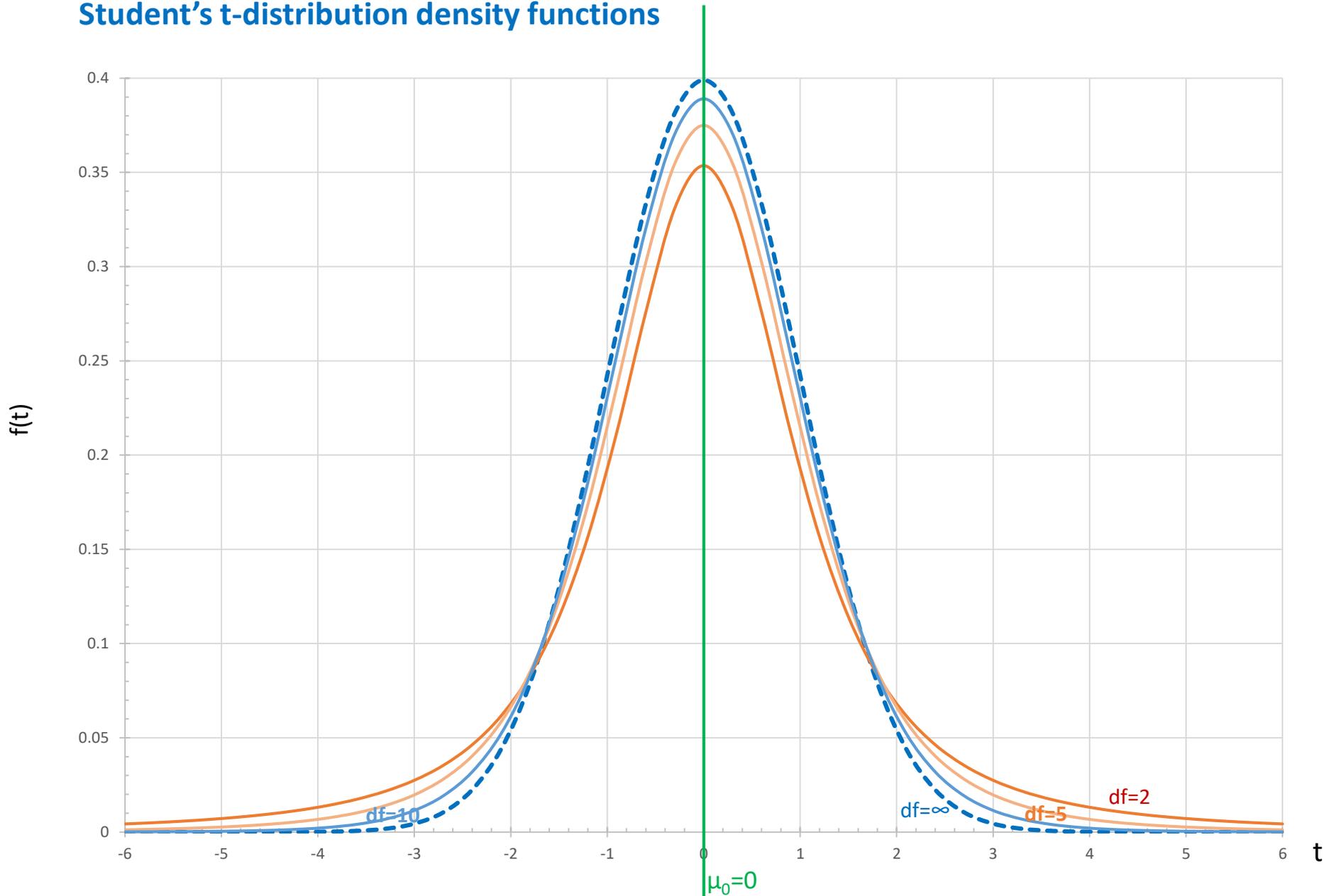
Parameter of the test

P-value

Decision



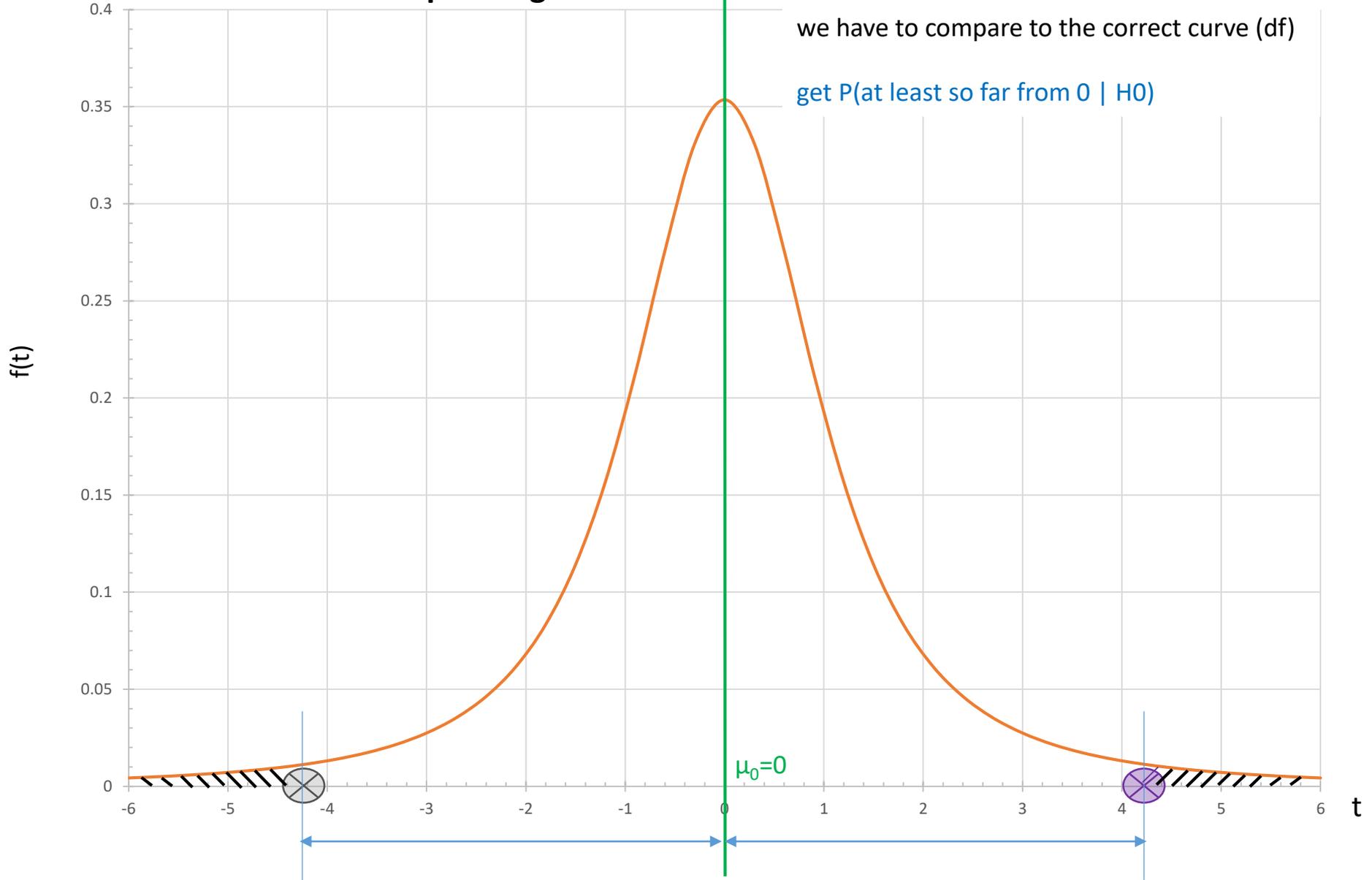
Student's t-distribution density functions



It is a family of functions, identified by the “degrees of freedom” as a number (>0)

the t-value is calculated from the sample as the parameter of the test.

now we seek the corresponding area as the P-value



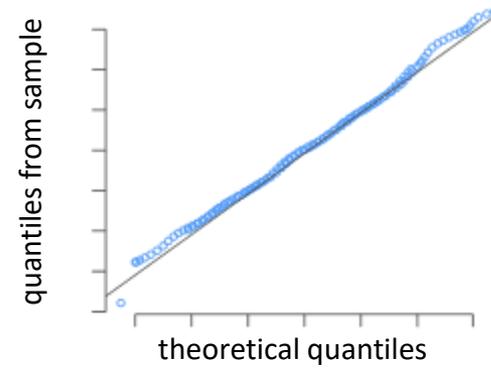
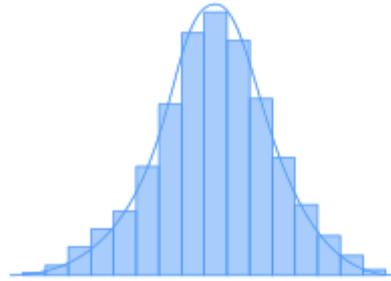
we have to compare to the correct curve (df)

get P(at least so far from 0 | H0)

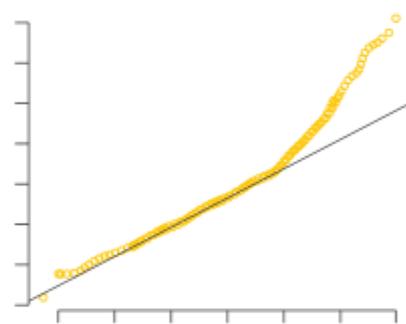
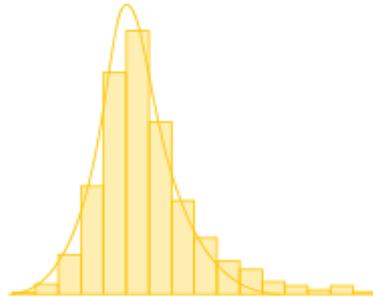
if P is too small (below threshold) we reject H0

Checking normality: QQ-plot

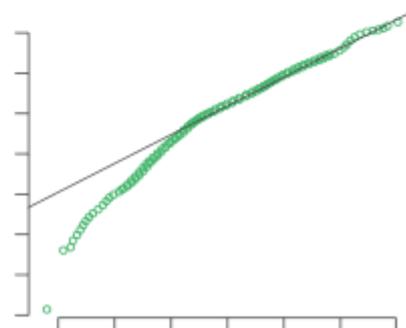
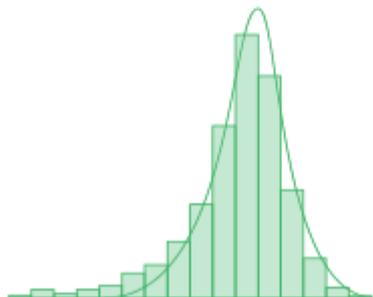
Normally distributed data



Right-skewed data



Left-skewed data



Note:

If our sample is large enough (more than 30...50 elements) then due to the central limit theorem, the mean will anyways follow a normal (Gaussian) distribution.

at least it will be close enough to normal, so the t-test can be used.

As can be seen in the formulae, the t-test actually only calculates the difference of the means, plus the standard deviation of the means is needed to get the t-value. So if the means are normal distributed, the formula is usable.

Since a large enough sample results in an approximately normally distributed mean, in this case the actual distribution loses its importance.

However, if we do know that our data are normally distributed, then the sample size can be any.

Independence:

we have to take great care during experimental design!
if in doubt, it needs a separate hypothesis test.

homogeneity of standard deviations (homoscedasticity):

there are specialized tests for this, like F-test, Cochrane-test, stb.
Ezek többnyire nagyon érzékenyek a normalitásra is.

it is best to prepare for heteroscedasticity, and correct for it (Welsch)

What to do with multiple comparisons?

Naïve idea: compare by pairs. -> here we make multiple probabilistic decisions -> at the end we can have greatly increased probability of decision error.

Solution:

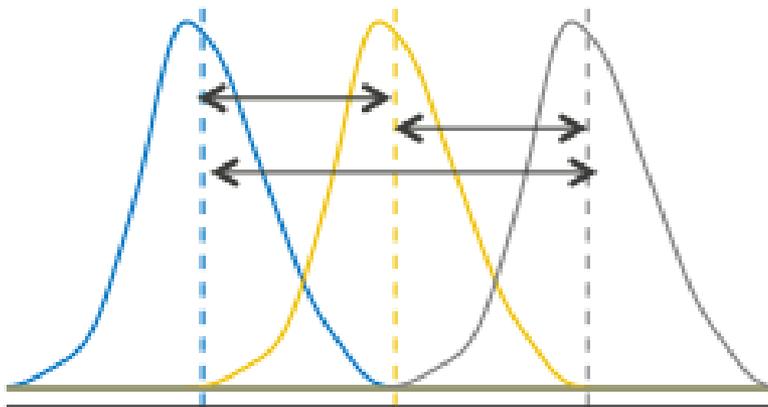
- A) Correct for the increased error probability, and effectively lower the threshold (significance) level.
- B) First decide if ALL groups belong to the same population in ONE step! if they turn out to be all the same we are done, if not then we look for the “odd one out” with *post-hoc* tests.

H0: all the same

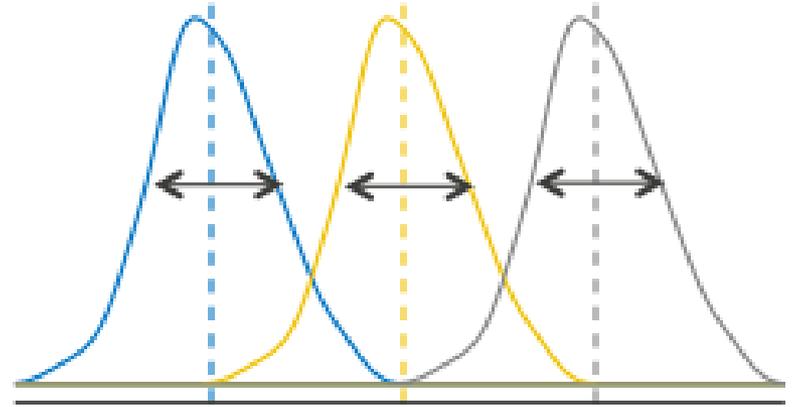
ANalysis Of VAriance (ANOVA)

idea: examine the **standard deviations!**

if the groups actually belong to the same population, we get identical SD-s in-group or cross-group scenarios.



cross-group SD-s



intra-group SD-s

Testing independence for *categorical* data

-> for continuous variables a regression is more appropriate

χ^2 - tests

Setting: we have observations, each yielding at least two different variables. We seek to test if these are independent or not.

e.g. does the usage of mobile phone affect the ability to keep focus on a subject? Phone usage patterns are categorical. Is there a correlation between lung cancer and smoking? Both variables are categorical.

	B1 smoking	B2 non-smoker
A1 has cancer	n1	n2
A2 does not have cancer	n3	n4

A and B are the two variables, they may have multiple possible values, not just Y/N, so the table here can be larger than 2x2.

n values are the observed frequencies

	B1 smoking	B2 non-smoker	
A1 has cancer	n1	n2	$P_{A1}=(n1+n2)/N$
A2 does not have cancer	n3	n4	$P_{A2}=(n3+n4)/N$
	$P_{B1}=(n1+n3)/N$	$P_{B2}=(n2+n4)/N$	

For the (marginal) probabilities we may use the independency rule (H0=indep.):

e.g. $P(A1 \text{ AND } B1) = P_{A1} * P_{B1}$, etc.

if the joint probabilities are calculated, then the **expected absolute frequency** is $P * N$. We will assume that $n_i = P * N$ for all cells. Small deviations are probably OK, as usual.

	B1 smoking	B2 non-smoker	
A1 has cancer	n1	n2	$P_{A1}=(n1+n2)/N$
A2 does not have cancer	n3	n4	$P_{A2}=(n3+n4)/N$
	$P_{B1}=(n1+n3)/N$	$P_{B2}=(n2+n4)/N$	$N=n1+n2+n3+n4$

observed frequencies
(contingency table)



	B1 smoking	B2 non-smoker	
A1 has cancer	$P(A1*B1)*N$	$P(A1*B2)*N$	
A2 does not have cancer	$P(A2*B1)*N$	$P(A2*B2)*N$	
			$N=n1+n2+n3+n4$

if H0 is true: expected table

we can calculate the deviation (squared, so + and - will not cancel each other, just like in the SD)

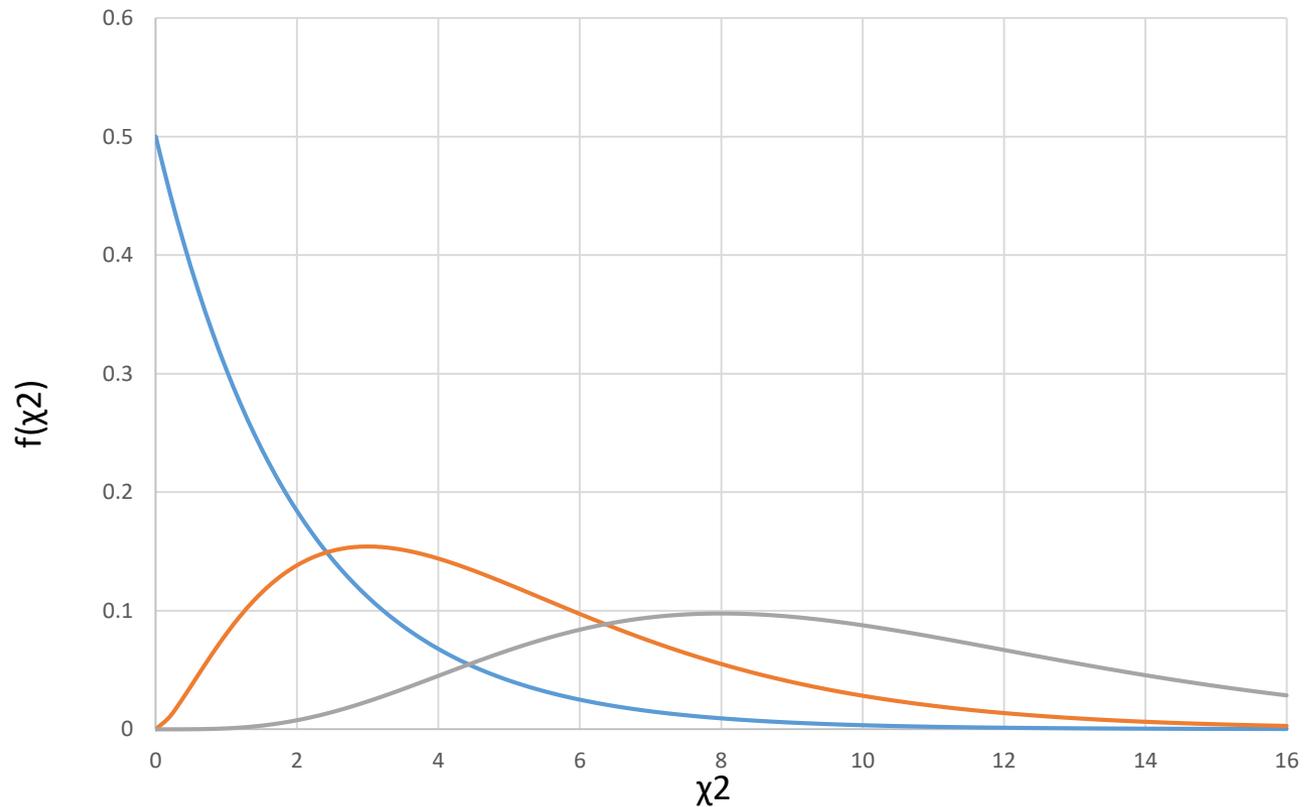
$$\chi^2 = \sum_{A,B} \frac{(Observed - Expected)^2}{Expected}$$

χ^2 is the parameter of the test.

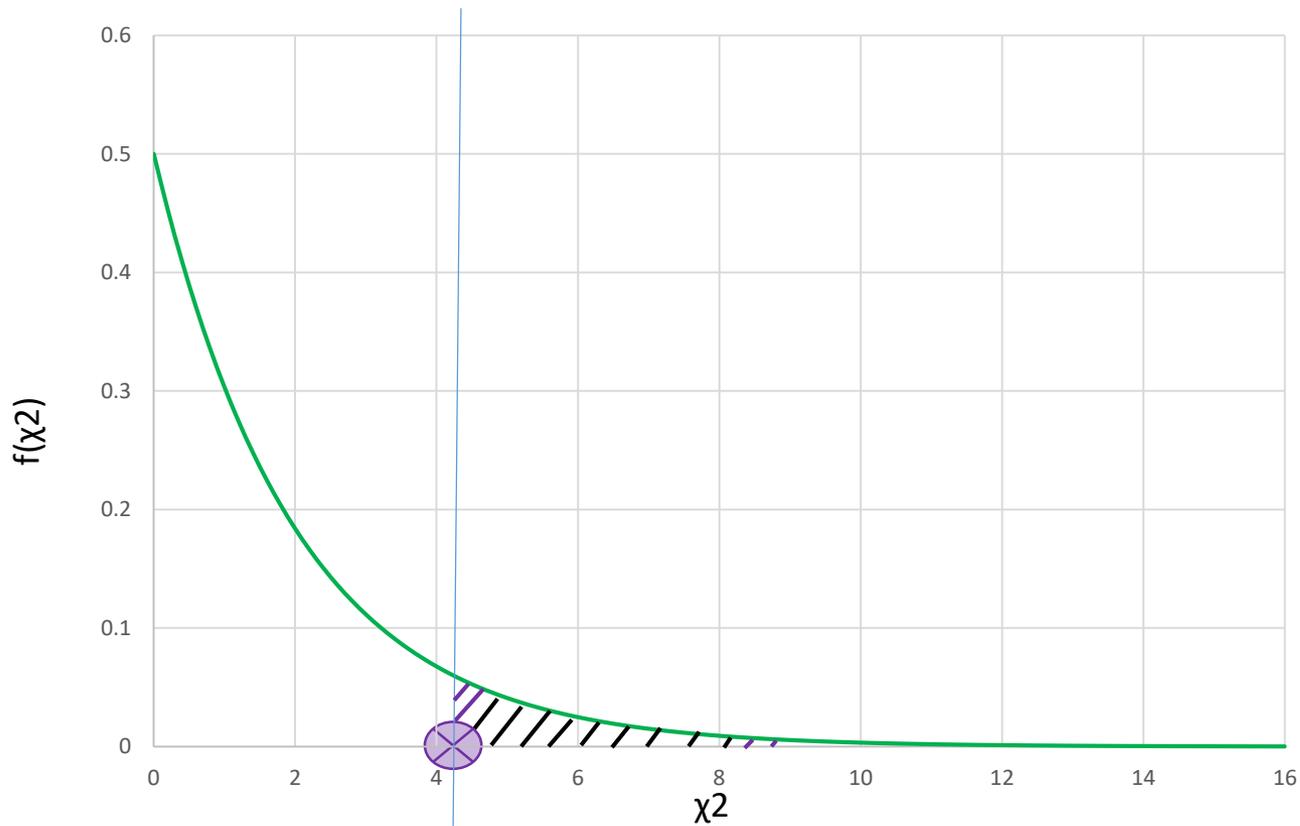
Then the **P-value** is the area under the curve having the correct df.

df = (number of rows -1) x (number of columns -1)

here we have $(2-1)*(2-1) = 1 * 1 = 1$, so we need the curve No 1 from the family.



Again we need the area under the curve (χ^2 is at least so large)



Decision is as usual, for too small P we reject.

-> A χ^2 test can also be used to check if the data fits any particular distribution, we just need the expected frequencies