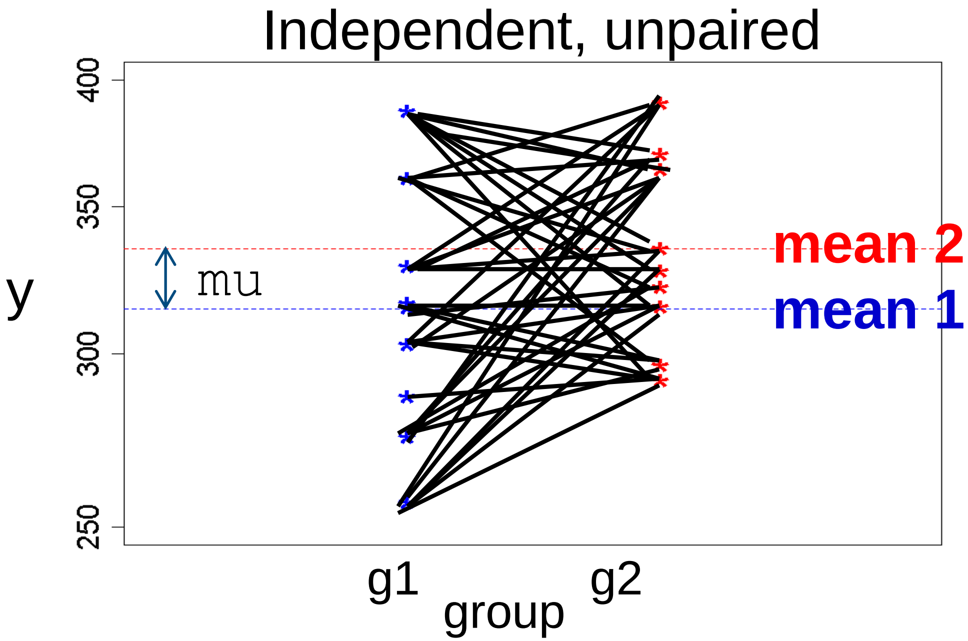
# Biostatistics
## Week 5

- Non parametric Wicoxon test on location
- sample size calculation / power analysis
- multiple testing

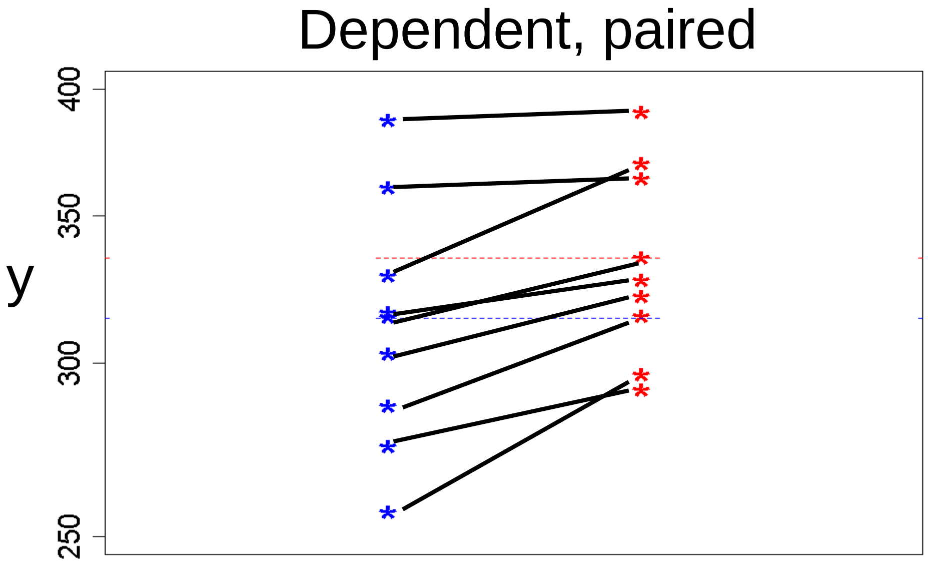
  - Bonferroni correction (for << 100 tests)

  - False discovery rate, p-value histogram (for >100 tests)

# Reminder: Unpaired and paired t-test on location



```
t.test(g1,g2, mu=0,
          var.equal=T, paired=FALSE)
```



```
t.test(g1,g2, mu=0,
          var.equal=T, paired=TRUE)
```

Breaking the match results in a valid group/treat effect but invalid p-values.

# Has caffeine intake influence on the reaction time?

- 10 “patients”
- We measure reaction times after treatment with coffee.
- Once coffee contains coffeine once not.

paired design

$H_0$: no difference with placebo or drug
population center is the same

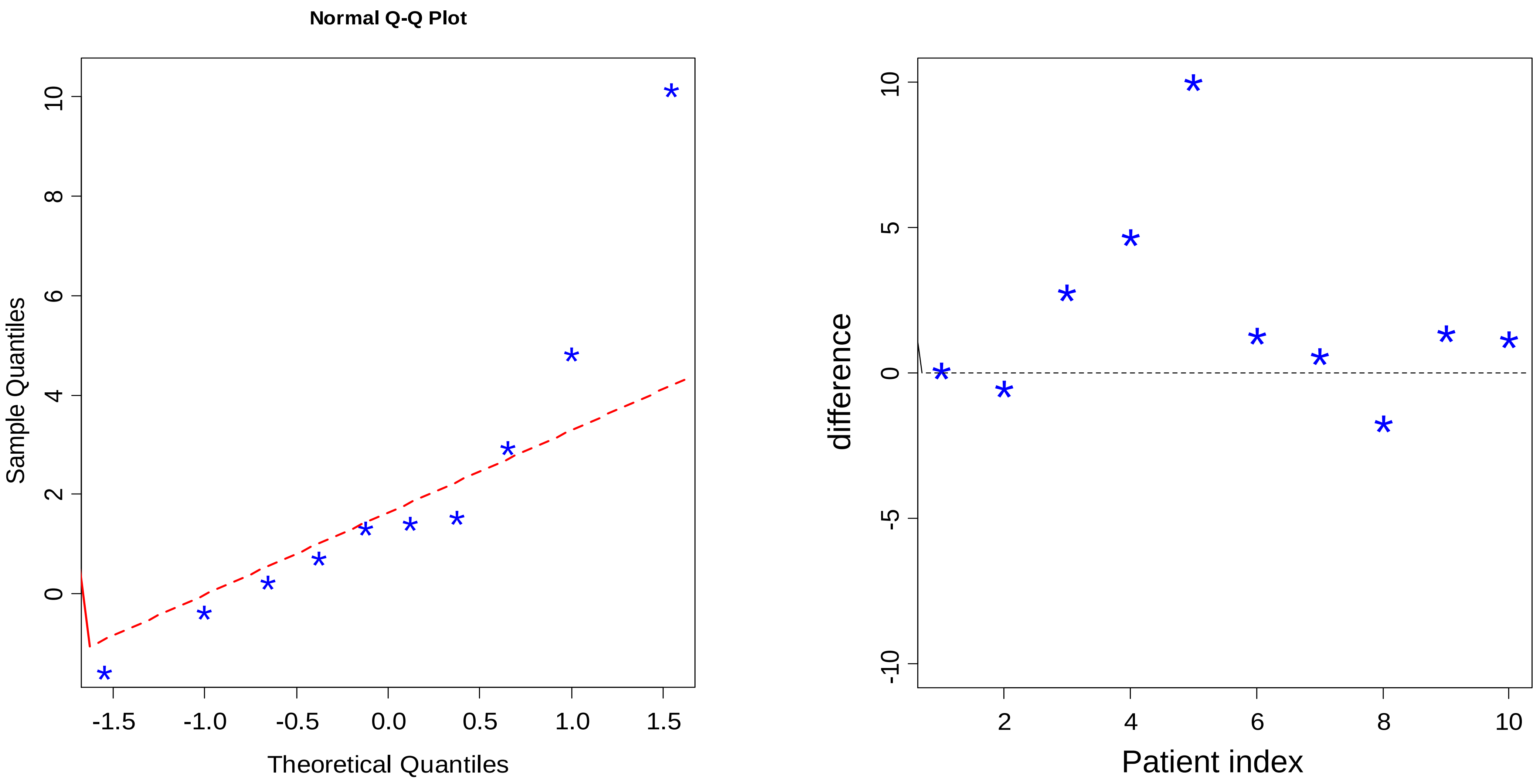
```
> t.test(exp$Differenz, mu=0, conf.level=0.95)

        One Sample t-test

data:  exp$Differenz
t = 2.1842, df = 9, p-value = 0.05678
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.08171953  4.66171953
sample estimates:
mean of x
     2.29
```

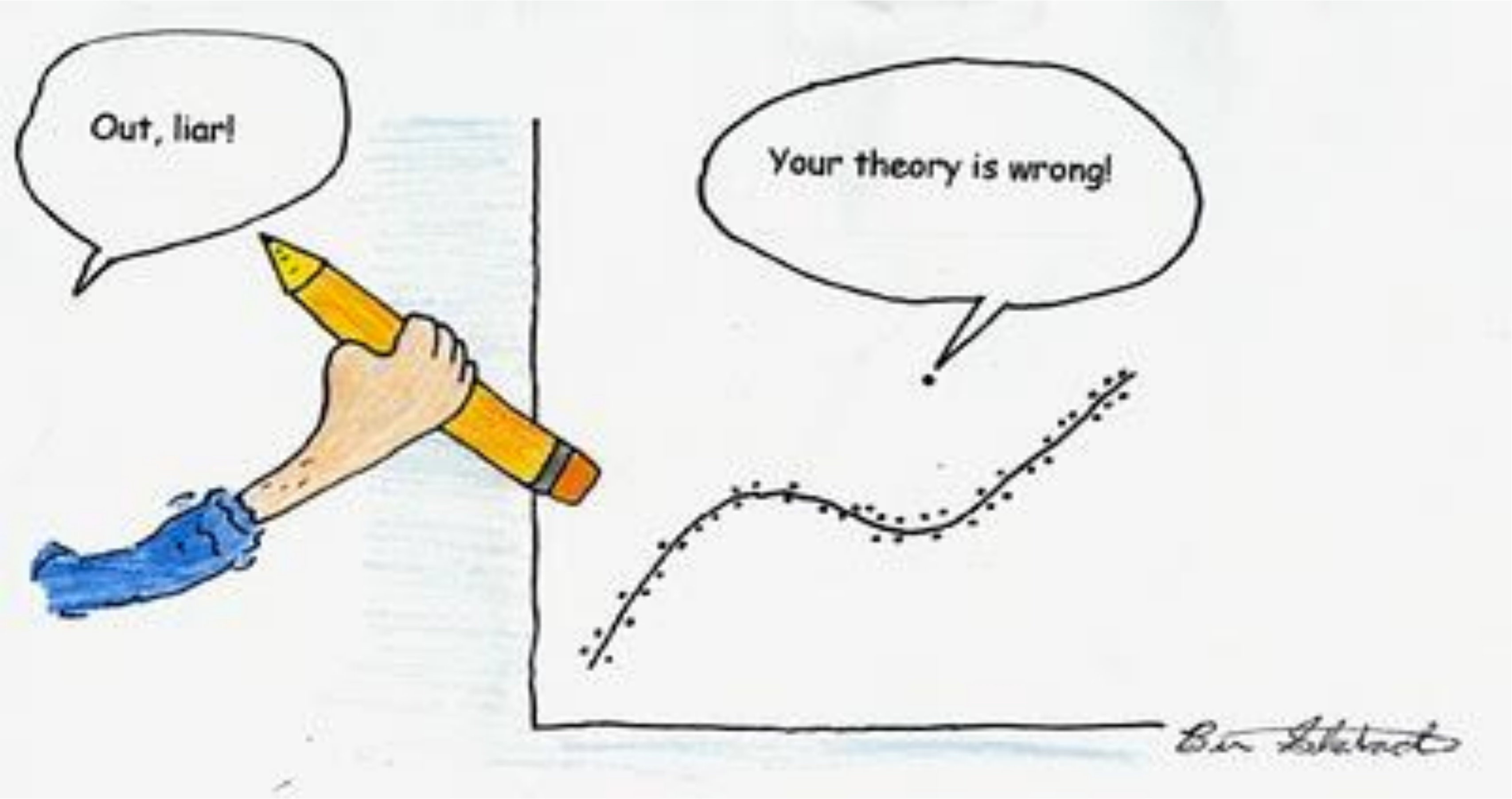
| Patient | Reaction time with coffeine | Reaction time with decof | diff |
|---|---|---|---|
| 1 | 44.5 | 44.9 | 0.4 |
| 2 | 55.0 | 54.8 | -0.2 |
| 3 | 52.5 | 55.6 | 3.1 |
| 4 | 50.2 | 55.2 | 5.0 |
| 5 | 45.3 | 55.6 | 10.3 |
| 6 | 46.1 | 47.7 | 1.6 |
| 7 | 52.1 | 53.0 | 0.9 |
| 8 | 50.5 | 49.1 | -1.4 |
| 9 | 50.6 | 52.3 | 1.7 |
| 10 | 49.2 | 50.7 | 1.5 |

# Visualization of the data



**There is a outlier! We must not perform a t-test!**

# How to handle outliers?



**Remove an outlier only, if you are sure that there was an error, e.g. the measurement went wrong.**
**Otherwise keep outlier an adapt your theory or use methods which can handle extreme values in an adequate way.**

# Look on ranks of the absolute differences

| index | abs(d)=│d│ | Rank(\|d\|) | sign(d) |
|---|---|---|---|
| 1 | 0.2 | 1 | - |
| 2 | 0.4 | 2 | + |
| 3 | 0.9 | 3 | + |
| 4 | 1.4 | 4 | - |
| 5 | 1.5 | 5 | + |
| 6 | 1.6 | 6 | + |
| 7 | 1.7 | 7 | + |
| 8 | 3.1 | 8 | + |
| 9 | 5.0 | 9 | + |
| 10 | 10.3 | 10 | + |

**Idea**: Look at sum of ranks of positiv and negative difference – they should be similar if the expected value of d is zero.

$$U^+ = \sum R^+ \quad , \quad U^- = \sum R^-$$

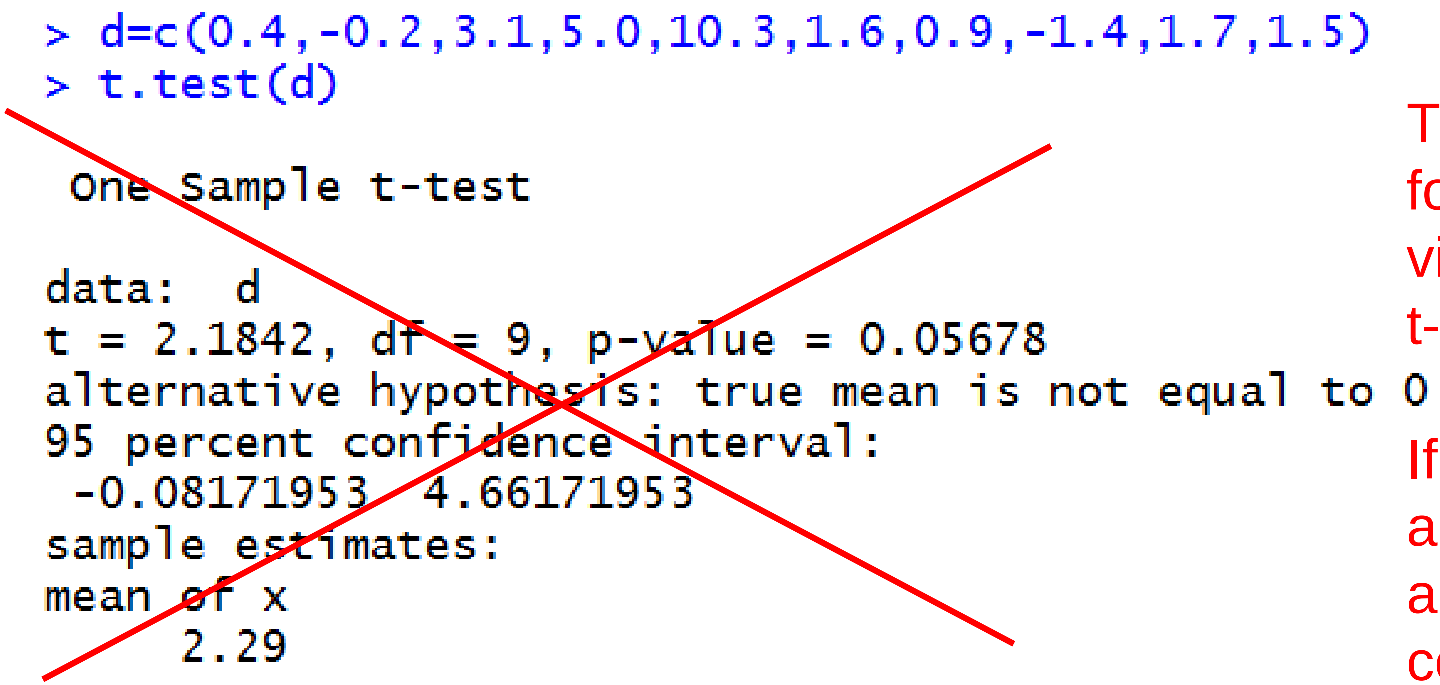
$$\mathit{Teststatistik:}\ \ U = \min(U^+, U^-)$$

Under $H_0$:

$$\sum R^+ \approx \sum R^- \approx \frac{1}{2}\sum_{k=1}^{n} k = \frac{1}{2}\cdot\frac{n}{2}\cdot(n+1)$$

$$\mathit{reject}\ H_0,\ \mathit{if}\quad U \ll \frac{1}{2}\cdot\frac{n}{2}\cdot(n+1)$$

# t-test or Wilcoxon-test?



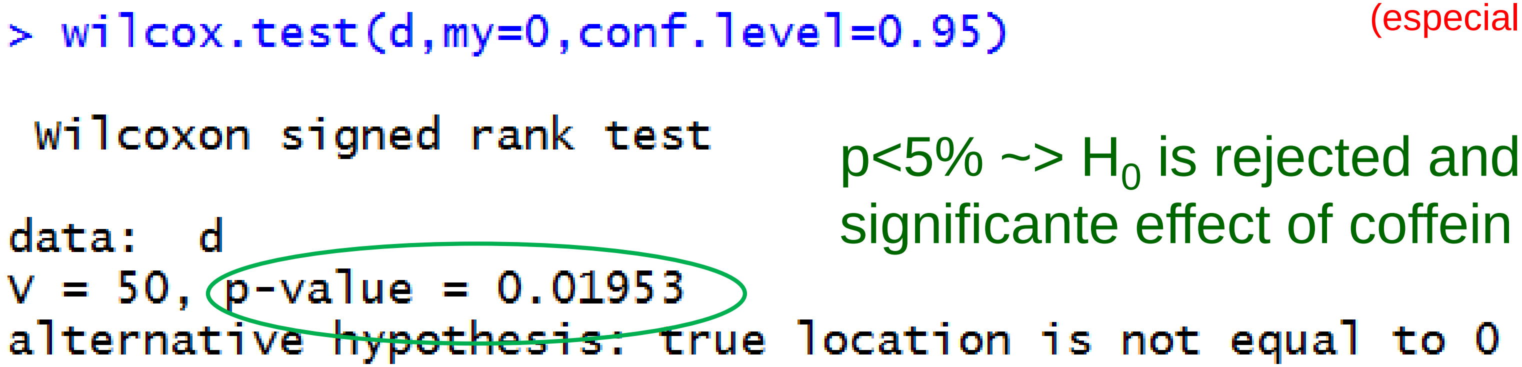
```
> d=c(0.4,-0.2,3.1,5.0,10.3,1.6,0.9,-1.4,1.7,1.5)
> t.test(d)

 One Sample t-test

data:  d
t = 2.1842, df = 9, p-value = 0.05678
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.08171953  4.66171953
sample estimates:
mean of x
     2.29
```

The normality assumption for the t-test is strongly violated, therefore the t-test must not be used.

If the t-test is performed anyway then the results are not reliable and can be completely wrong (especially with small sample sizes).



```
> wilcox.test(d,my=0,conf.level=0.95)

 Wilcoxon signed rank test

data:  d
V = 50, p-value = 0.01953
alternative hypothesis: true location is not equal to 0
```

p<5% ~> $H_0$ is rejected and we have shown a significante effect of coffein on the reaction time.

The 1-sample wilcoxon-test requires only a symmetric distribution, which is for difference from paired values always fulfilled.

# When to use non-parametric tests like the wilcoxon-tests?

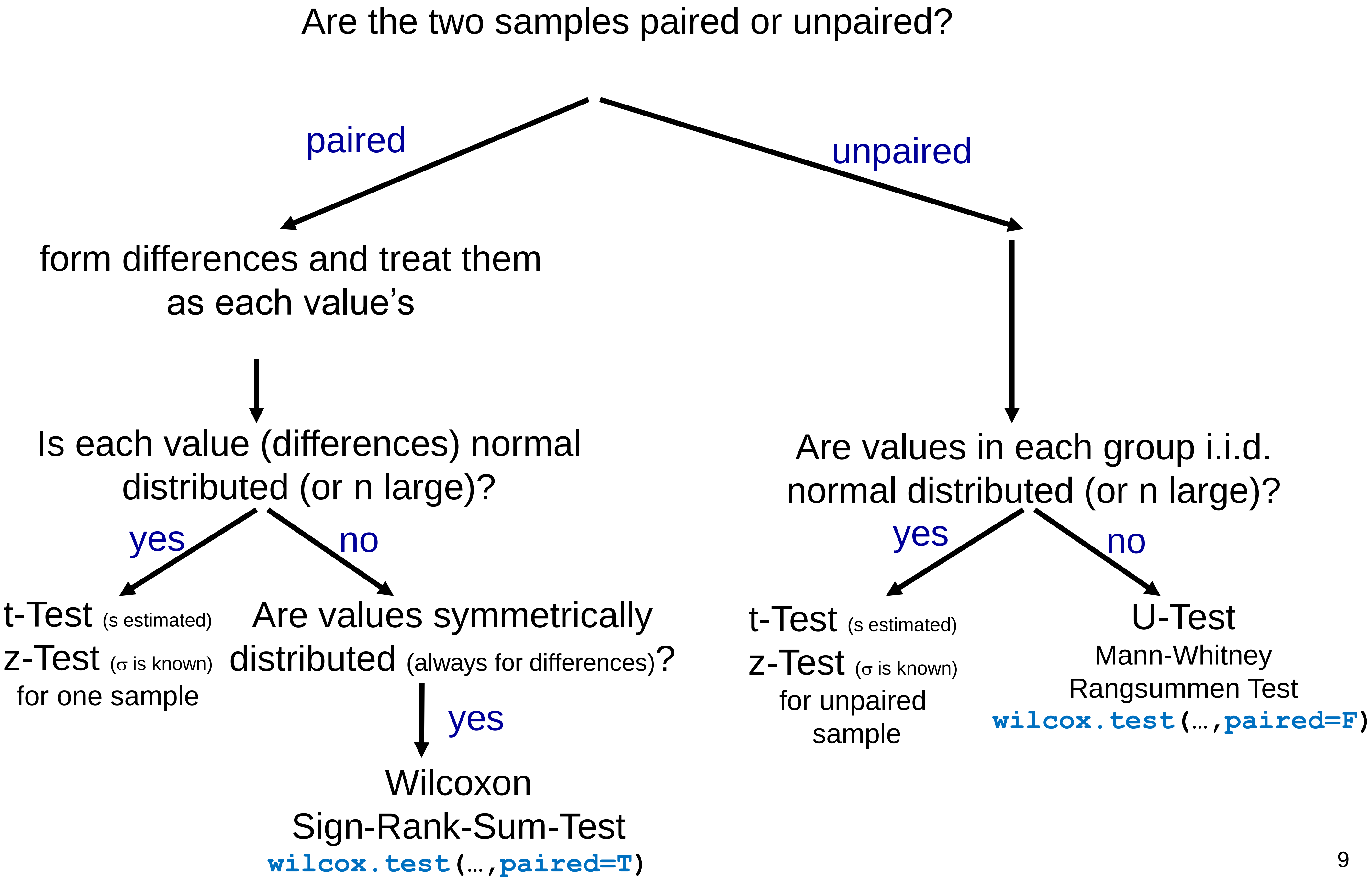
- If data do **not** follow a Normal-Distribution (and sample is not large)
- If there might be outliers
- If the sample size is very small (<≈10) and don't know if data come from $N(\mu,\sigma^2)$

**Remark 1**: in an unpaired situation there exists also a wilcoxon test, which is known as U-test or Mann-Whitney-test and which also uses a test statistic relying on the ranks of the data.

**Remark 2**: if the data (in each group) follow a Normal-Distribution, than the t-test has more power than the wilcoxon-test.

**Remark 3**: for small samples (<10) the normality of data can hardly be checked and the wilcoxon-test should be used if normality is questionable.

# Two-sample tests



Are the two samples paired or unpaired?

paired

form differences and treat them as each value's

Is each value (differences) normal distributed (or n large)?

yes

t-Test (s estimated)
z-Test (σ is known)
for one sample

no

Are values symmetrically distributed (always for differences)?

yes

Wilcoxon Sign-Rank-Sum-Test
`wilcox.test(…,paired=T)`

unpaired

Are values in each group i.i.d. normal distributed (or n large)?

yes

t-Test (s estimated)
z-Test (σ is known)
for unpaired sample

no

U-Test
Mann-Whitney Rangsummen Test
`wilcox.test(…,paired=F)`

# Decision errors revisited

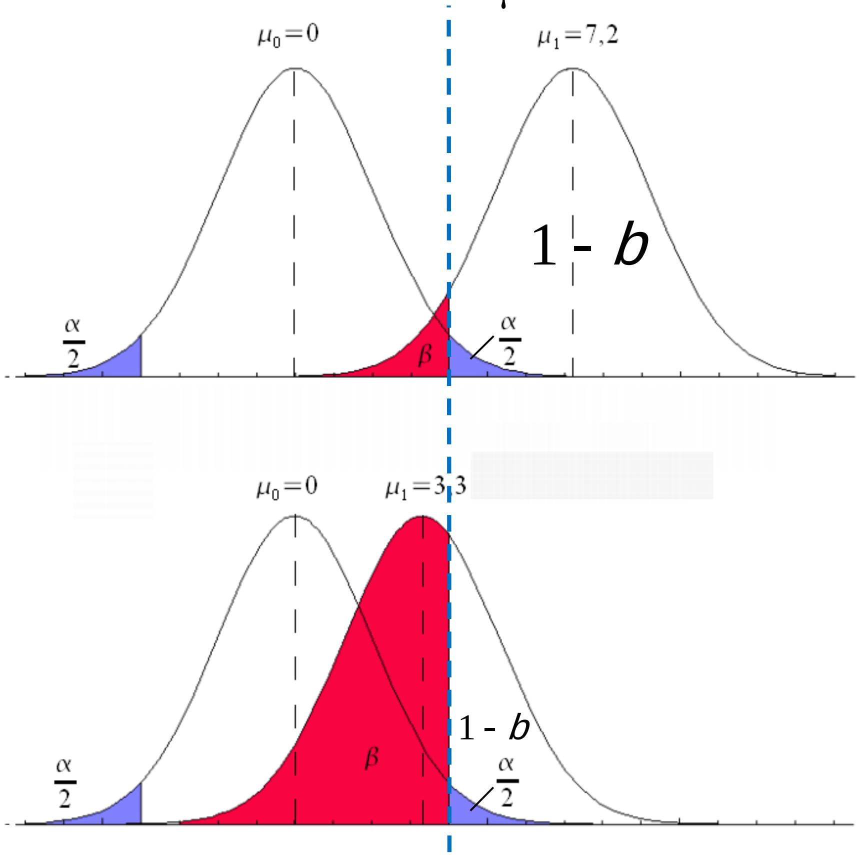
| | negative test accepting $H_0$ | positive test rejecting $H_0$ |
|---|---|---|
| $H_0$ is true | **True Negative**<br>(the probability for this correct test decision is (1-α) ) | **False Positive**<br>(the probability for a type-I error is α) |
| $H_0$ is false | **False Negative**<br>(the probability for a type-II error is β) | **True Positive**<br>(the probability for this correct test decision is (1-β)) |

Known old reality

$H_0$: $\mu_0=0$

$H_A$: $\mu_A \neq 0$

Unknown new reality: $\mu=7.2$



$H_0$: $\mu_0=0$

$H_A$: $\mu_A \neq 0$

Reality

$\mu=7.2$

$P(reject\ H_0 \mid H_0\ true) = \alpha$ probability for type I error

$P(\text{accept}\ H_0 \mid H_0\ false) = \beta$ probability for type II error

power = $1\text{-}\beta$

Effect size

=

difference between $H_0$ and unknown new reality

# What is the power of a test

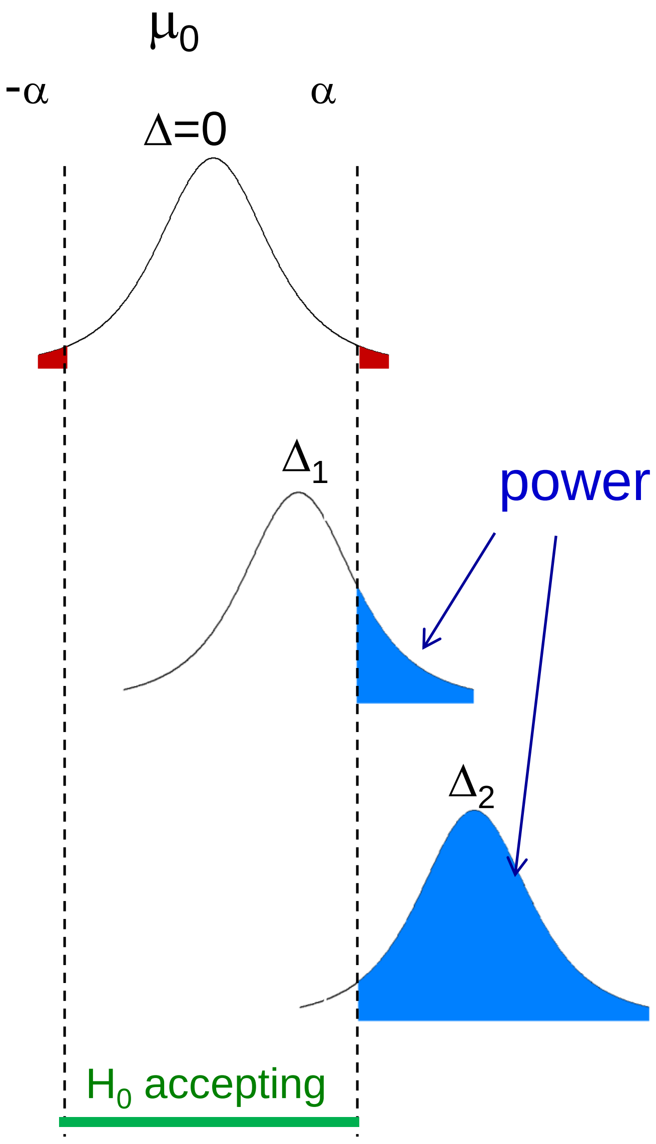
The power $(1-\beta)$, of a test is the probability to reject $H_0$, if $H_A$ is true.

The power is given by the blue area.

The larger the difference $\Delta$ between $H_0$ and reality is, the larger gets the power. However, “reality” is not known -> it is hard to estimate the power.

For a given difference $\Delta$ between $H_0$ and reality the power gets bigger if the width of the distribution of the Test-Statistic, e.g. mean, gets smaller which can be achieved by increasing the sample size.

Since the reality can not be changed, in praxis the only way to increase the power is to increase the sample size.



# Sample size calculation

**Situation:**

Your group has developed a new drug for sleeping time elongation.

The new drug is only interesting if its sleeping time elongation surpasses the one from the «golden standard» by at leas 1h (relevant effect).

From a pilot study we know the standard deviation of the sleeping time in individual patients is: sd=1 (1h).

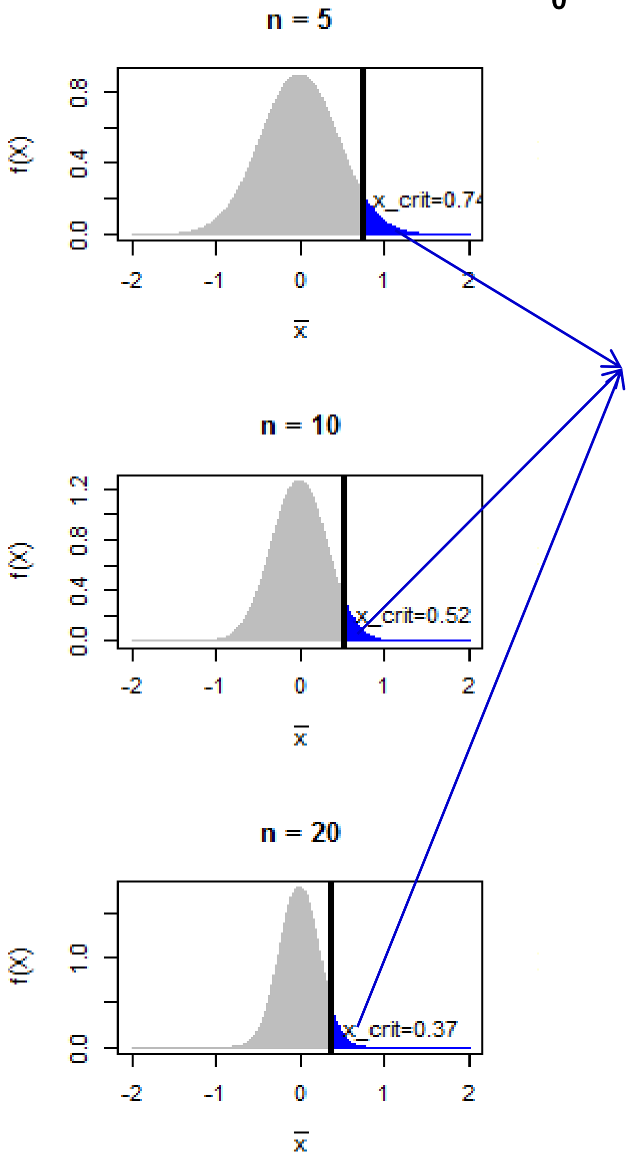
Given your drug surpasses the old drug by a mean sleeping time extension of 1h - how big should the sample size be chosen, so that you have a power of 80% and simultaneously an $\alpha$=5% that your test rejects the $H_0$ and proves the superiority of the new drug?

# Simulation with n=5,10,20

X: Sleep elongation compared the golden standard drug

$H_0$: $E(X)_0=\mu_{\Delta 0}=0$

**Distribution of $\bar{x}$ under $H_0$**



With a sample of size 5 we would reject $H_0$, if $\bar{x} > 0.74$

$\alpha = 5\%$ is fixed

With a sample of size 10 we would reject $H_0$, if $\bar{x} > 0.52$

With a sample of size 20 we would reject $H_0$, if $\bar{x} > 0.37$

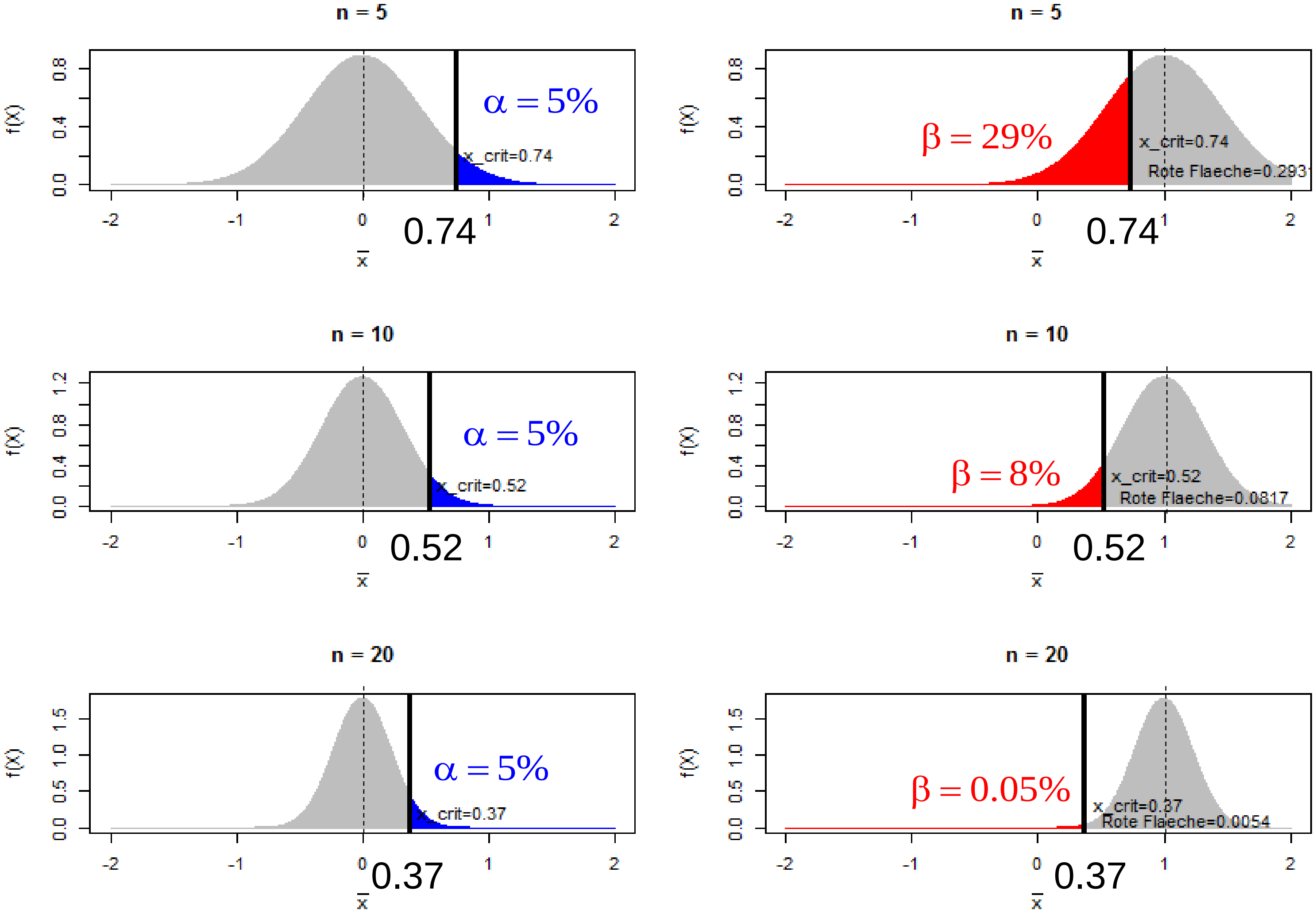
# Simulation with n=5,10,20

X: Sleep elongation compared the golden standard drug

$H_0$: $E(X)_0=\mu_{\Delta 0}=0$

$H_A$: $\mu_\Delta=1$

**Distribution of $\bar{x}$ under $H_0$** **Distribution of $\bar{x}$ given $\mu_\Delta=1$**



n=5: Reject $H_0$ in 69% of all simulation runs

**Power=1-β=69%**

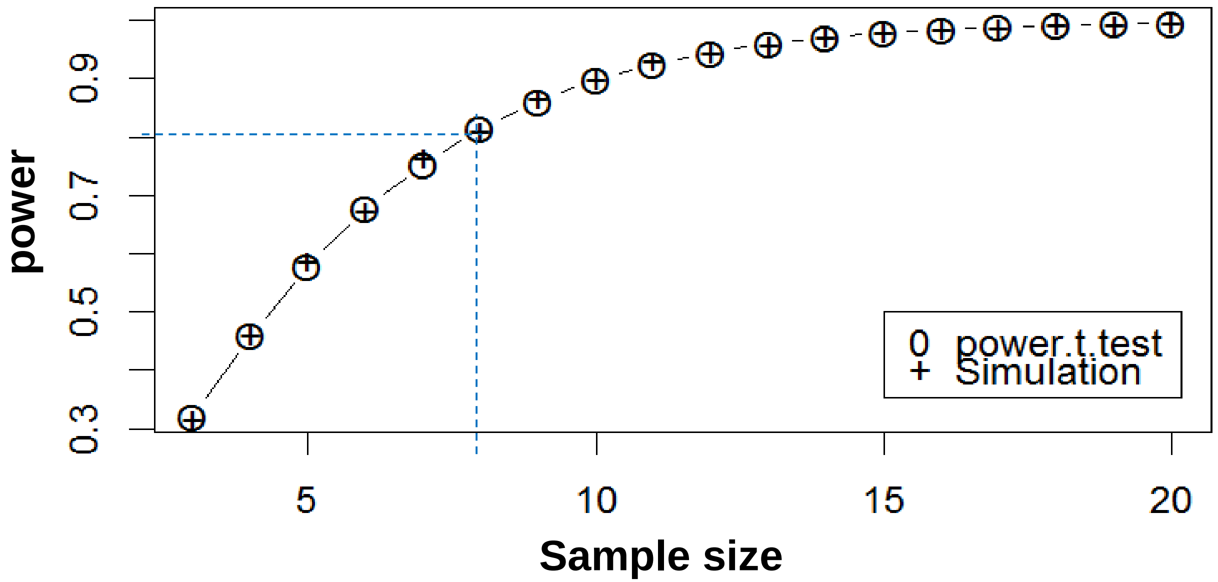
n=10: Reject $H_0$ in 92% of all simulation runs

**Power=1-β=92%**

n=20: Reject $H_0$ in 99.5% of all simulation runs

**Power=1-β=99.5%**

# Results



```
 n         power.simu
 3             0.3171
 4             0.4631
 5             0.5880
 6             0.6737
 7             0.7652
 8             0.8118
 9             0.8685
10             0.9001
11             0.9308
12             0.9452
13             0.9579
14             0.9712
15             0.9803
16             0.9862
17             0.9893
18             0.9926
19             0.9939
20             0.9960
```

```
> power.t.test(power=0.8, delta=1, sd=1,
              alternative="one.sided",type="one.sample")
```

```
     One-sample t test power calculation

              n = 7.727622
          delta = 1
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
```

# How to plan the size of a study?

- Choose the test you want to use in your analysis
- Determine/Estimate the variation of the observations
- fix significance level $\alpha$ (the accepted risk for a type-1-error, typically 5%)
- Fix relevant effect size (the minimal effect which is still relevant)
- Fix the power which gives the probability to detect an relevant effect (typically 80%)
  - Choose 1-$\beta$

Perform a sample-size calculation to derive the needed sample size at which the required power is given.

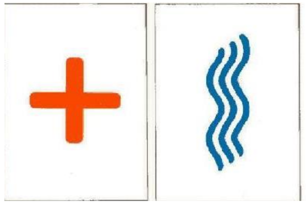
Good webpage for sample size calculations – menu based, but shows corresponding R-code:

http://powerandsamplesize.com/Calculators/

# Multiple testing: Rhine Paradox

- The parapsychologist **Joseph Rhine** hypothesized in the 1950's that some people had *Extra-Sensory Perception* (*ESP*).
- He tested for ESP by an experiment where people were asked to guess the color of 10 hidden cards:

red or blue.



- He discovered that almost 1 in 1000 had ESP – they were able to get all 10 right. **Surprised???**

$$P(10 \text{ correct answers} \mid \text{just guessing}) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \left(\frac{1}{2}\right)^{10} = \frac{1}{2^{10}} = \frac{1}{1024}$$

No, 1 in 1024 is what we would expect to get by chance if everybody is just guessing

# Multiple testing: Rhine Paradox

- He told these people they had ESP and called them in for another test of the same type.

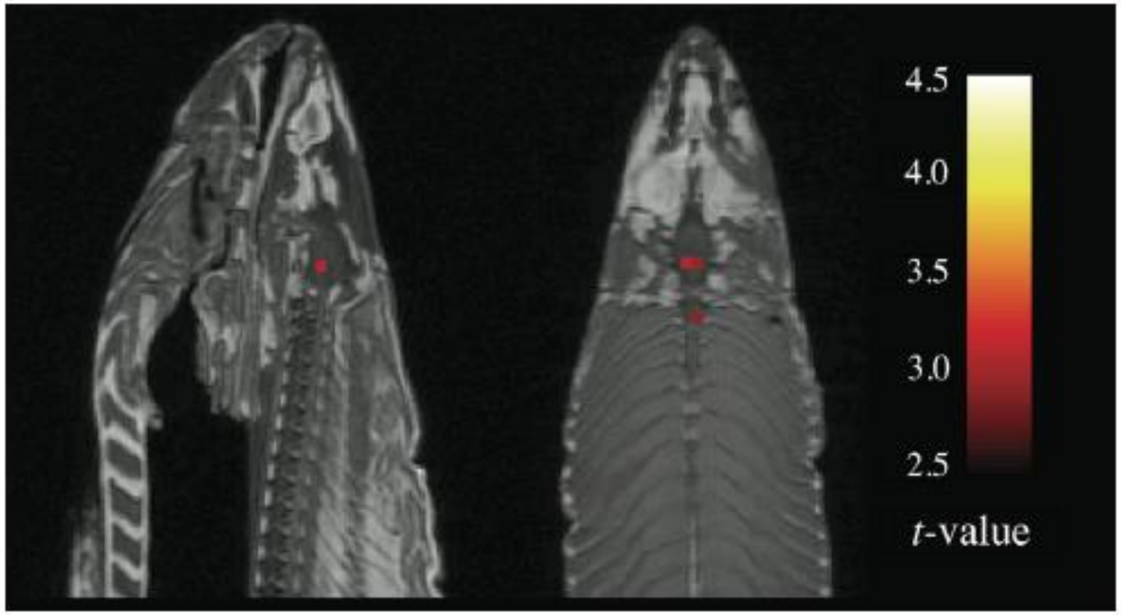
- He discovered that all of them had lost their ESP.
- **What did he conclude???**

# Multiple testing: Rhine Paradox

**He concluded that you shouldn’t tell people that they have ESP, because it causes them to loose it.**

☺

# fMRI revealed brain response to trans-species emotional stimuli in a dead salmon



A dead salmon was repeatedly confronted with 2 different human emotional stimuli.

Out of 8064 brain voxels in 16 voxels a significant different activity ($p \leq 0.001$!) was observed

This study received 2012 the IG nobel price (for *ignoble,* improbable research).

Source: http://www.guardian.co.uk/science/2012/sep/21/ig-nobel-awards-dead-salmon

# The probability to get by chance a significant test result

The risk to get in **one test** an false positive result (that is p<$\alpha$ under H0) is only

$$P(reject\ H_0\ |\ H_0\ true)\ =\ \alpha$$

$$P(accept\ H_0\ |\ H_0\ true)\ =\ 1-\alpha$$

Assume n independent test‘s (with n independent samples) where the null-hypothesis $H_0$ is always valid (no effect nowhere)

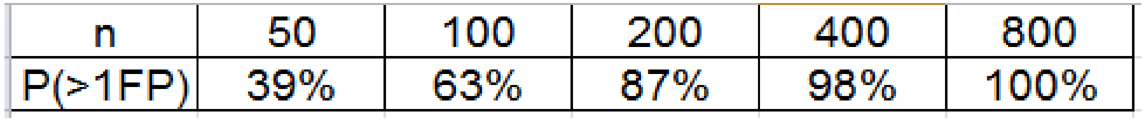
– the probability draw always the right test decision is:

$$P(accept\ n-times\ H_0\ |\ H_0\ true)\ =\ (1-\alpha)^n$$

– the probability of coming up with at least 1 false positive effect is:

$$P(\geq 1\ rejecting\ H_0\ |\ H_0\ true)\ =\ 1-(1-\alpha)^n$$

the probability of making at least 1 type one error at **n trials**, when $\alpha$ = 0.01%



| n | 50 | 100 | 200 | 400 | 800 |
|---|---|---|---|---|---|
| P(>1FP) | 39% | 63% | 87% | 98% | 100% |

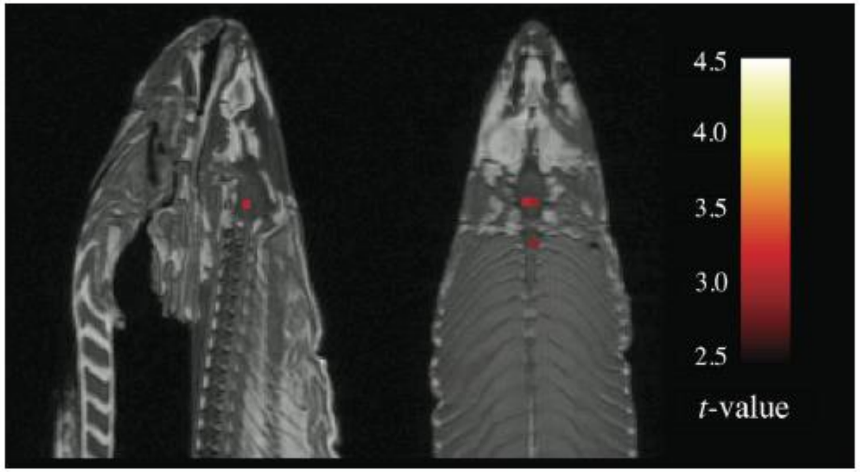
# Bonferroni correction for multiple testing and its effect on the dead salmons reaction

Bonferroni: when performing n independentl tests, conduct each test at significance level $\frac{\alpha}{n}$ !

When applying Bonferronis rule, we only have a risk of $\alpha$, to come up with ≥1 false positive effects (that is a significant test result although $H_0$ is true).

No brain region of the dead salmon showed a significant reaction to smiling people after Bonferroni correction.

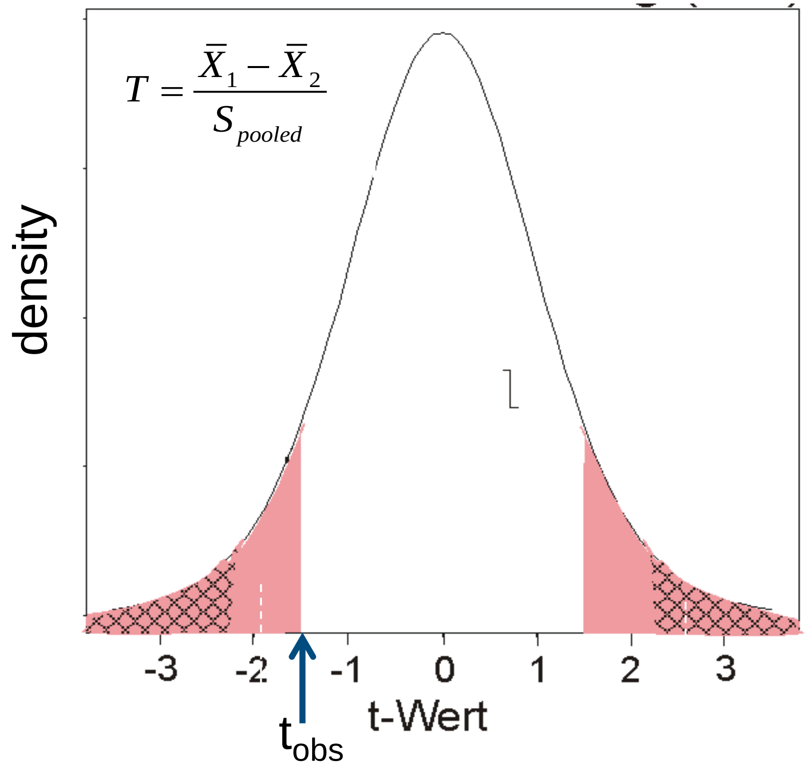
☺



# The p-value is uniformly distrubuted under $H_0$

The p-value corresponds to the probability to get an at least such extreme result as the observed result assuming that the Null-Hypothesis is valid
-> the p-value corresponds to the area in the extreme tails

distribution of T **under $H_0$**



$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_{pooled}}$$

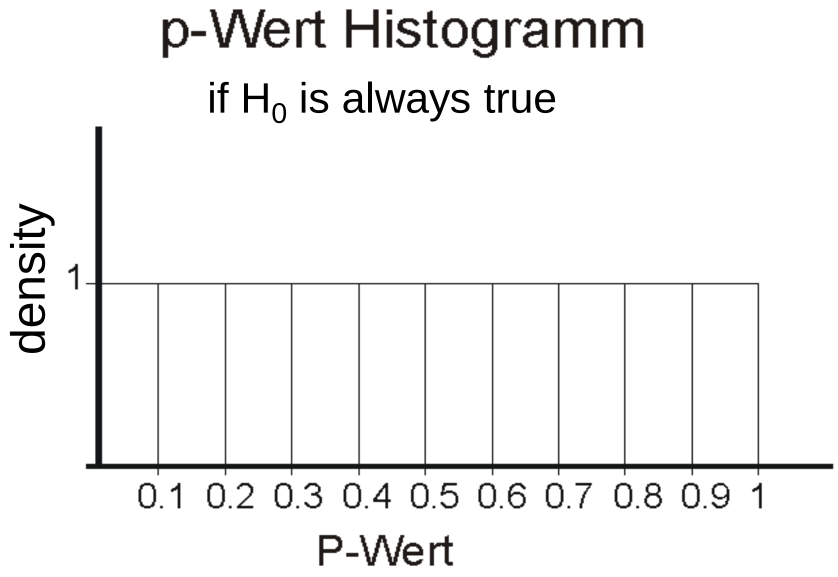
$$\text{p } = Prob(\,|t| > |t_{obs}|\ \mid H_0\ true\,)$$
$$= Prob(\,|p| \leq |p_{obs}|\ \mid H_0\ true\,)$$

Given $H_0$ is true in all tests:

p=0.1: 10% of all tests get a p-value≤0.1 if $H_0$ is always true

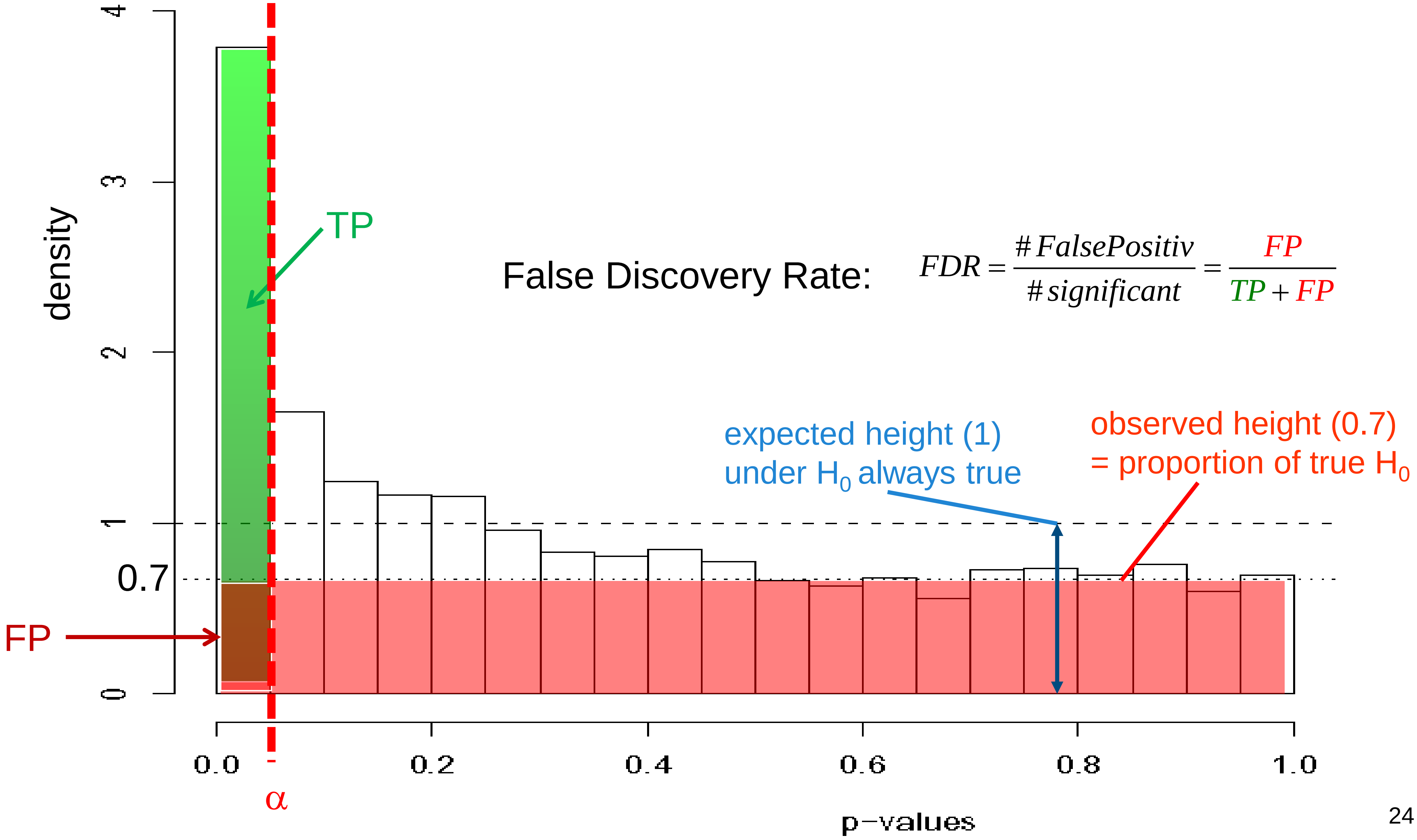
p=0.2: 20% of all tests get a p-value≤0.2 if $H_0$ is always true

…



if $H_0$ is always true

# How to estimate the ratio $p_0$ of truly null voxels?
# How to estimate the false discovery rate?



False Discovery Rate: $FDR = \frac{\# FalsePositiv}{\# significant} = \frac{FP}{TP + FP}$

expected height (1)
under $H_0$ always true

observed height (0.7)
= proportion of true $H_0$

# Judging a p-value histogram

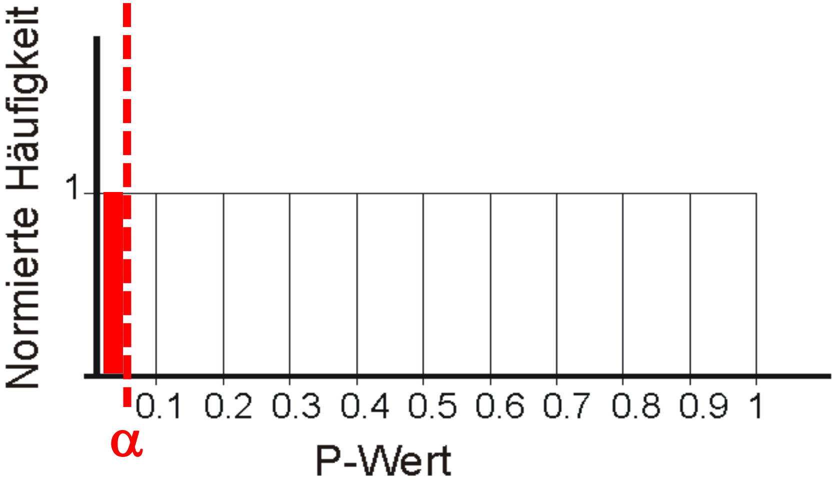
The p-value histogram helps to judge the results from many *independent* tests

Flat is Bad!

For all tests n independent $H_0$ is true

~ 100% of all significant findings are false positive.

(see proportion of red area in left bar).

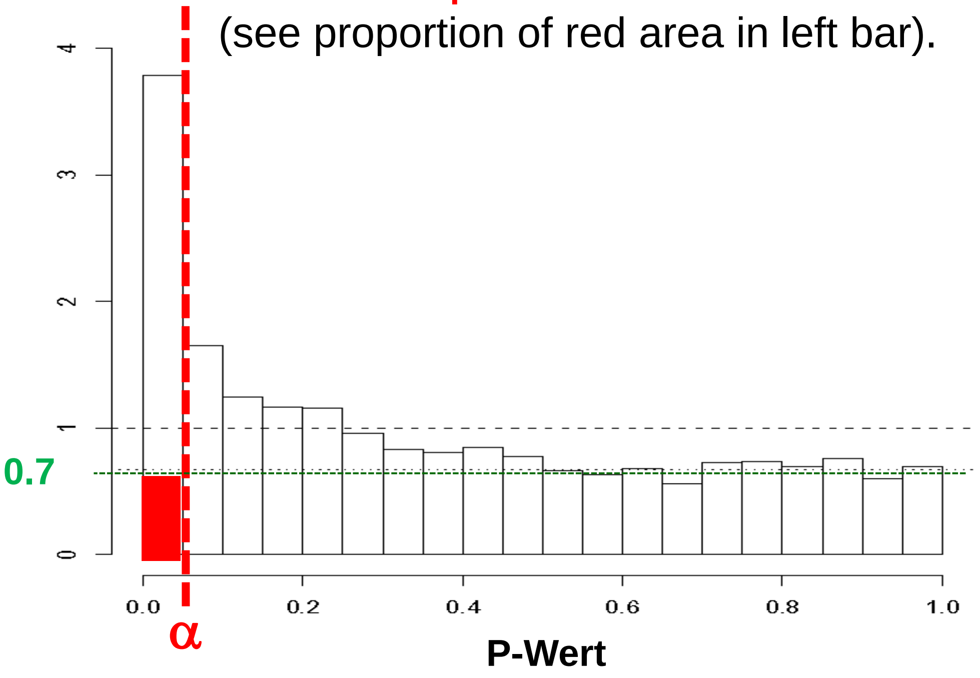


The peak we seak!

For 70% of all tests $H_0$ is true.

Only ~20% of all significant findings are false positive

(see proportion of red area in left bar).



This method was proposed from John Storey – see http://www.pnas.org/content/100/16/9440.full

# Take home message
# Multiple Testing

It is tempting, but not o.k. to forget about all non-significant tests and just publish the significant effects.

You **need to take account for the multiple testing**

- by p-value correction such as Bonferroni correction or
- using other measures like FDR or
- confirm the “found effect” in a new control experiment.