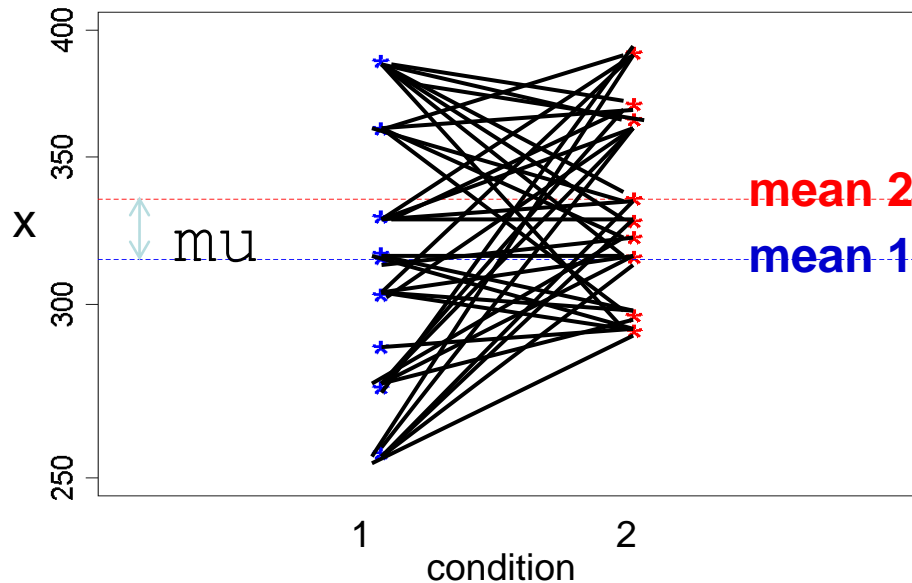


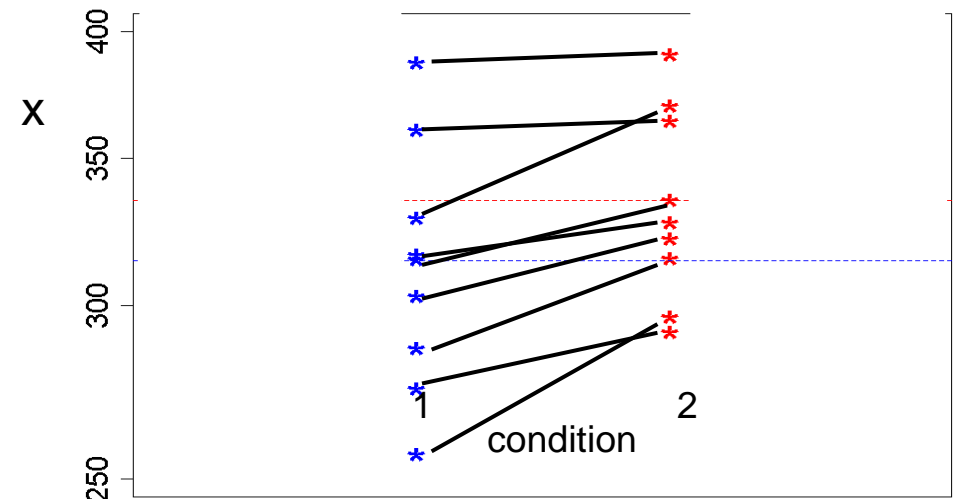
# Biostatistics Week 4

- Confidence interval continued
- Testing
  - t-test for means
  - 2 group comparison: paired and unpaired tests
- Outlook to non-parametric tests: Wilcoxon

Independent, unpaired



Dependent, paired



# Construction of an approximative 95% CI for population mean $\mu$ without assumptions on population distribution

$X_i$  i.i.d.  $i \in 1, \dots, n, n > 25, E(X) = \mu_x, \text{Var}(X) = \sigma_x^2$

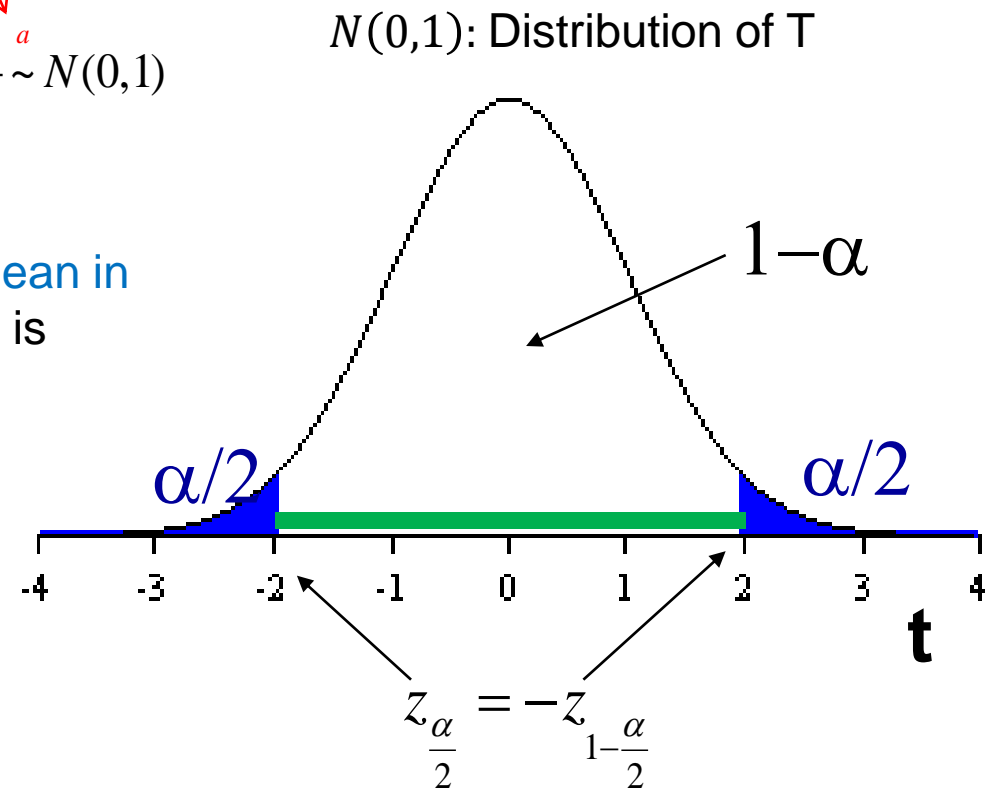
Test-Statistic or Pivot:

Central Limit Theorem  $\Rightarrow \bar{X} \overset{a}{\sim} N\left(\mu_x, \frac{s_x^2}{n}\right) \Rightarrow T = \frac{\bar{X} - \mu_x}{\frac{s_x}{\sqrt{n}}} \overset{a}{\sim} N(0,1)$

The test statistic  $T$  measures the distance to the mean in units of standard errors. The construction of the CI is based on the distribution of  $T$  under  $H_0 : \mu = \mu_x$

$$P\left(z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu_x}{\frac{s_x}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha$$

$$P\left(\bar{X} - \frac{s_x}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} \leq \mu_x \leq \bar{X} + \frac{s_x}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha$$



**approx. 95% CI for  $\mu_x$**   $\bar{X} \pm q_{0.975}^z \cdot \frac{s_x}{\sqrt{n}} \approx \bar{X} \pm 1.96 \cdot \frac{s_x}{\sqrt{n}}$

Quantile from  $N(0,1)$

standard error:  
 $se(\bar{X})$

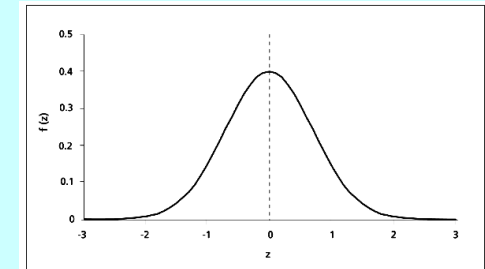
# Distribution of the test statistic T

assuming that values come from a Gauss centered at  $\mu$

$$X_1, X_2, \dots, X_n \sim N(\mu_x, \sigma_x^2) \text{ i.i.d.}$$

Variance  $\sigma_x^2$  is known.

$$T = \frac{\bar{X} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}} \sim N(0,1)$$



$n \rightarrow \text{big}(> 25) \quad t_{df=n-1} \rightarrow N(0,1)$

$$X_1, X_2, \dots, X_n \sim N(\mu_x, \sigma_x^2) \text{ i.i.d.}$$

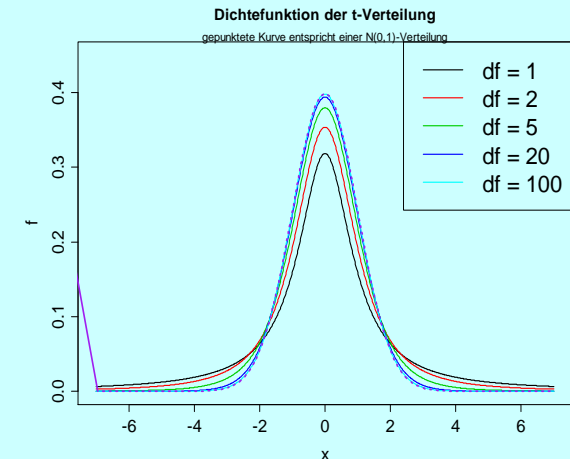
Variance  $\sigma_x^2$  is unknown and is estimated from the data

$$s_x^2 = \hat{\sigma}_x^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$T = \frac{\bar{X} - \mu_x}{\frac{s_x}{\sqrt{n}}} \sim t_{n-1}$$

$$se(\bar{x}) = \frac{sd(x)}{\sqrt{n}}$$

se: standard error of the mean  
variation of the estimator



Remark: Since beside the mean also the variance is derived from the random sample we have some additional variation when determining T and the distribution of T gets broader and is given by the  $t_{df=n-1}$

# The exact 95% CI for the expected value if values are normally distributed

$$\bar{x} \pm t_{n-1}^{97.5\%} \cdot \frac{sd(x)}{\sqrt{n}}$$

Quantile from  $t_{df=n-1}$

$se(\bar{x})$  : standard error of the mean

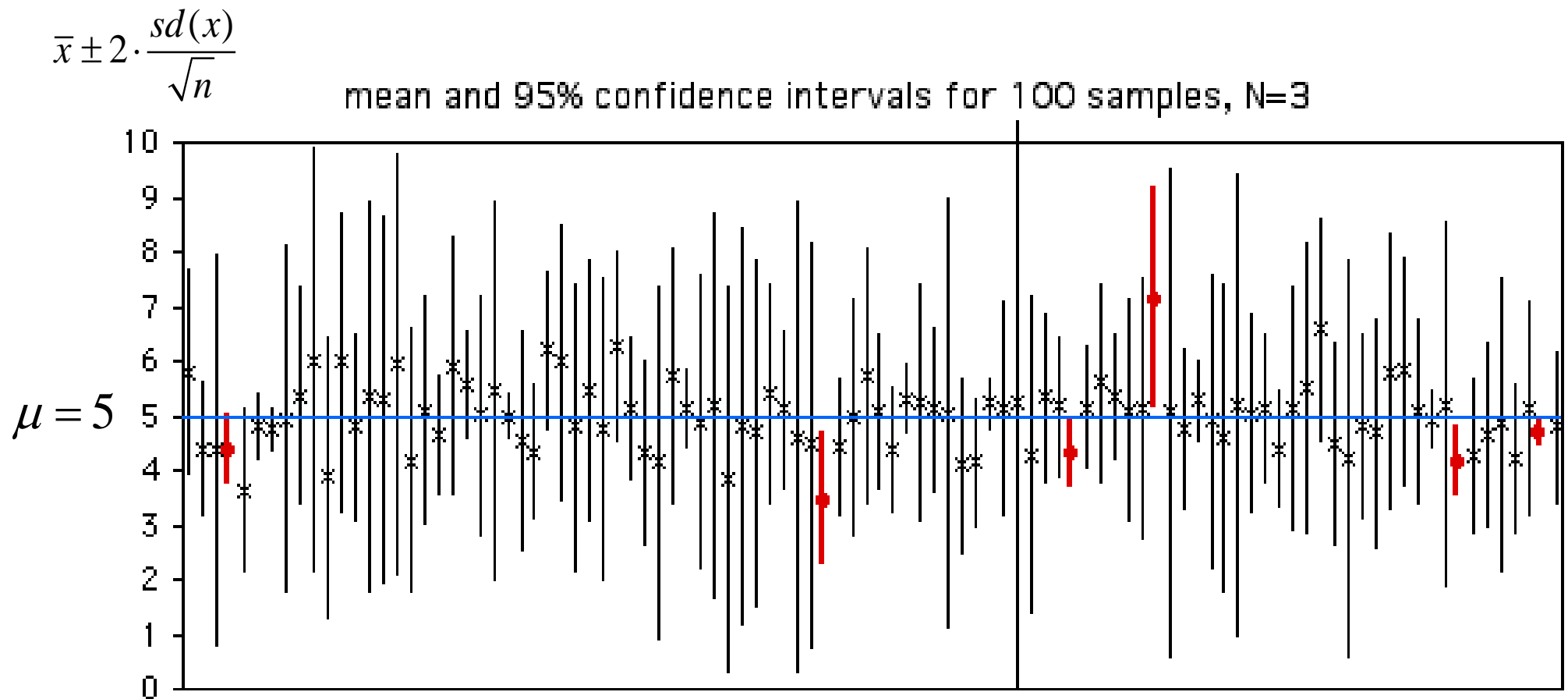
Please note that for the exact CI for the expected value, the quantiles of the t-distribution are used.

The t-distribution has a parameter df (degree of freedom), which must be set on n-1, where n is the number of observations in the sample.

Remark: If n gets large (>25) the quantiles of the t-distribution can be approximated by the quantiles of the N(0,1) distribution. In the large sample case (n>25) the assumption  $x \sim N(\mu, \sigma^2)$  is not essential!

The reason is the Central Limit Theorem that ensures that the mean is approximately normally distributed and therefor also the standardized mean.

# The CI is as random as the sample

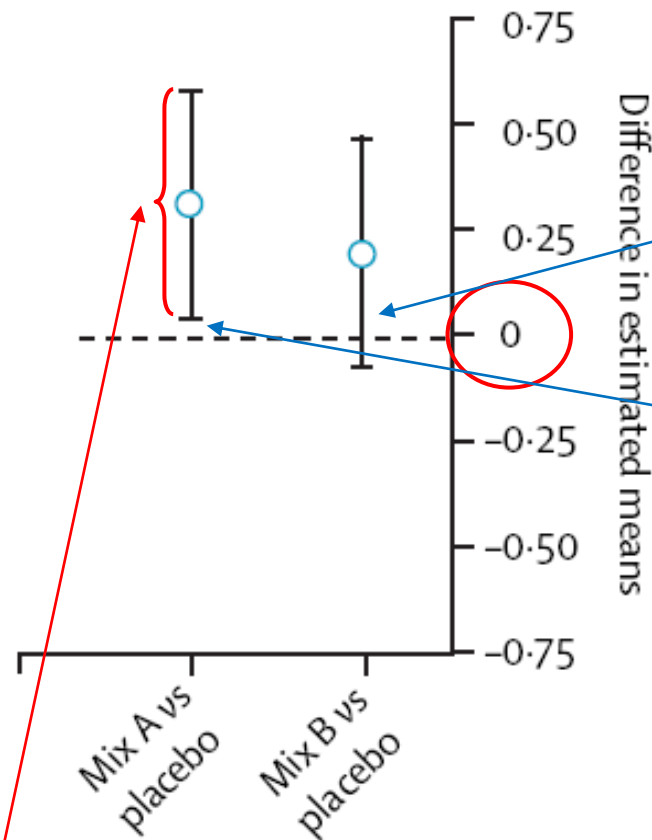


95 out of 100 95%-CI for  $\mu$  do cover the true population parameter  $\mu=5$  when simulating 100 random samples from a population following  $N(\mu=5, \sigma^2)$ .

With a 95%-CI we have a risk of 5% that our random sample was not typical for the population and the true population parameter is not contained by the CI.

# Interpretation of a confidence interval

## Example from paper on hyperactivity form McCann et al.

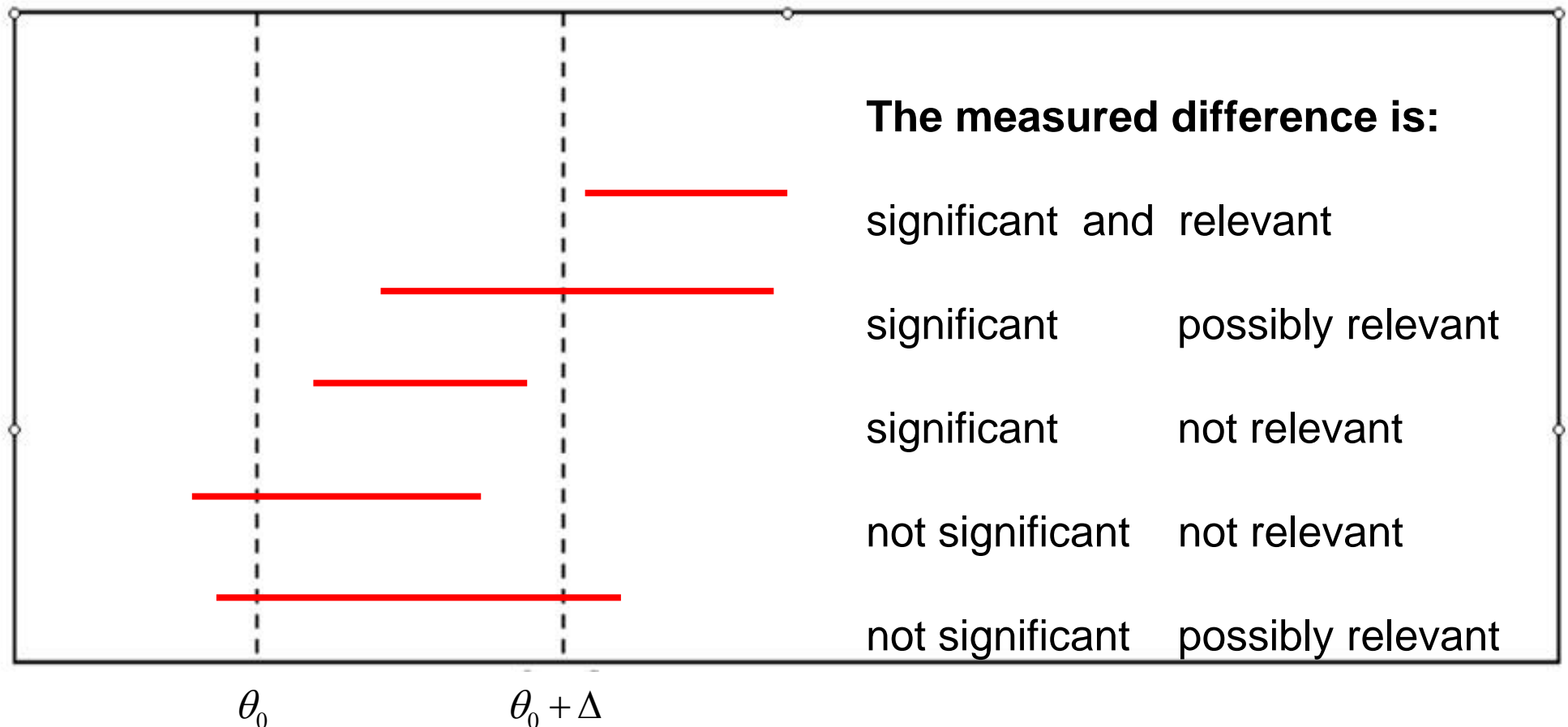


- The CI covers all plausible values for the true mean-difference – here the true treatment effect
- If 0 is covered by the CI it is plausible that the treatment effect is 0 – we have no evidence against  $H_0$ , that the treatment has no effect.
- If 0 is **not** covered by the CI, we say that the treatment effect is **significantly** different from 0.
- To have a reasonable chance (80%) to claim a relevant treatment effect to be significant we must plan the sample size to be large enough to be able to find a the effect to be significant if existing.

Here we see a 95% CI of the difference of the mean hyperactivity under placebo and under treatment with Mix A indicating a significant effect of Mix A.

With a confidence interval we can decide:  
 Is there a significant difference to a postulated value  $\theta_0$ ?  
 Is the difference relevant ( $>\Delta$ )?

Draw CIs that correspond to the description on the right



# Why to perform a statistical test?

## Typical research questions that trigger a statistical tests:

Does my drug work (does it lower the blood pressure)?

Which genes are differentially expressed between normal tissue samples and samples from various cancer stages?

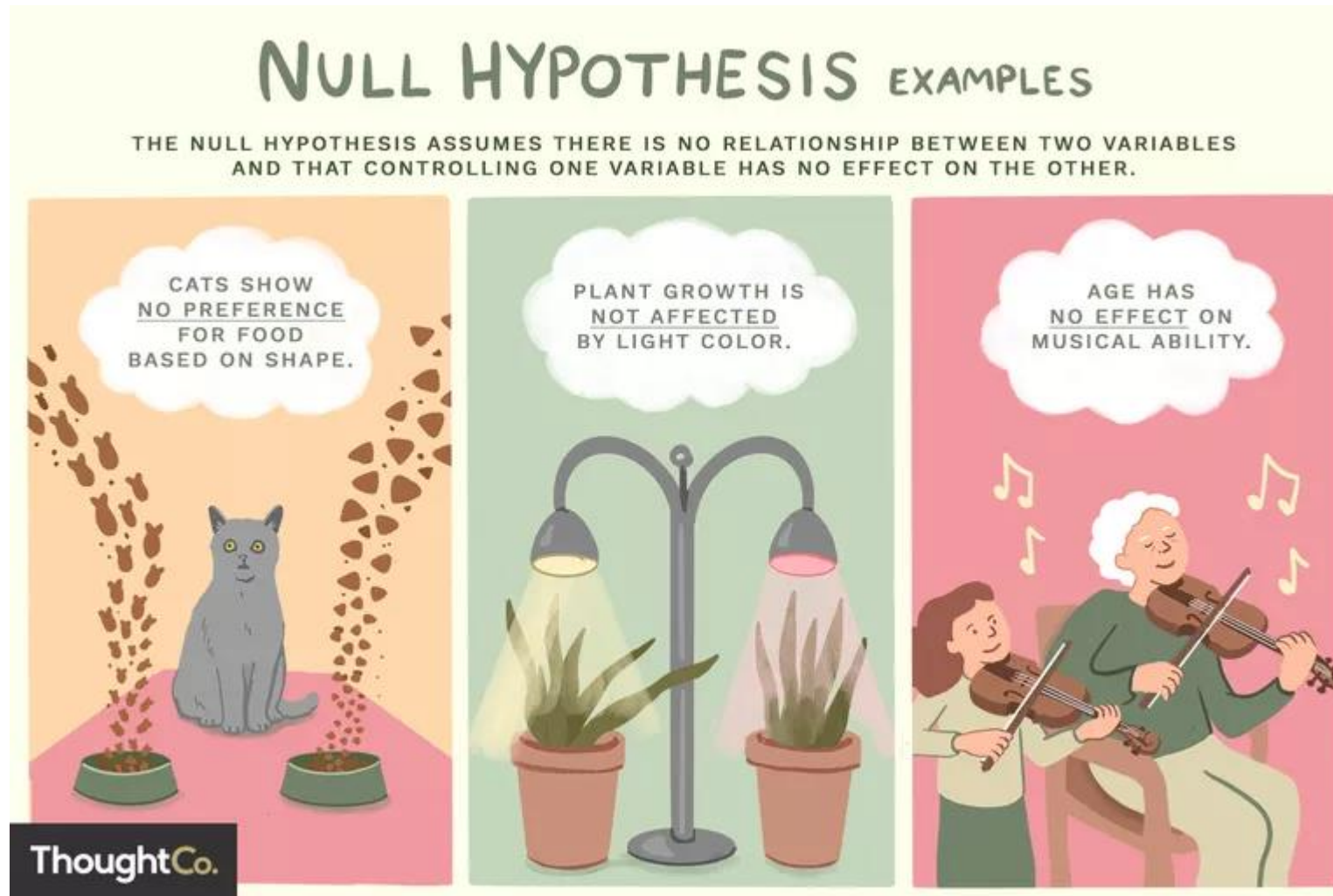
Is the plant grow affected by the color of the light

Has age an effect on musical ability

Has the shape of the cat food an impact on the preference of the cat?



# How to formulate a Null Hypothesis



Rule of thumb: Always use the boring stuff as Null Hypothesis

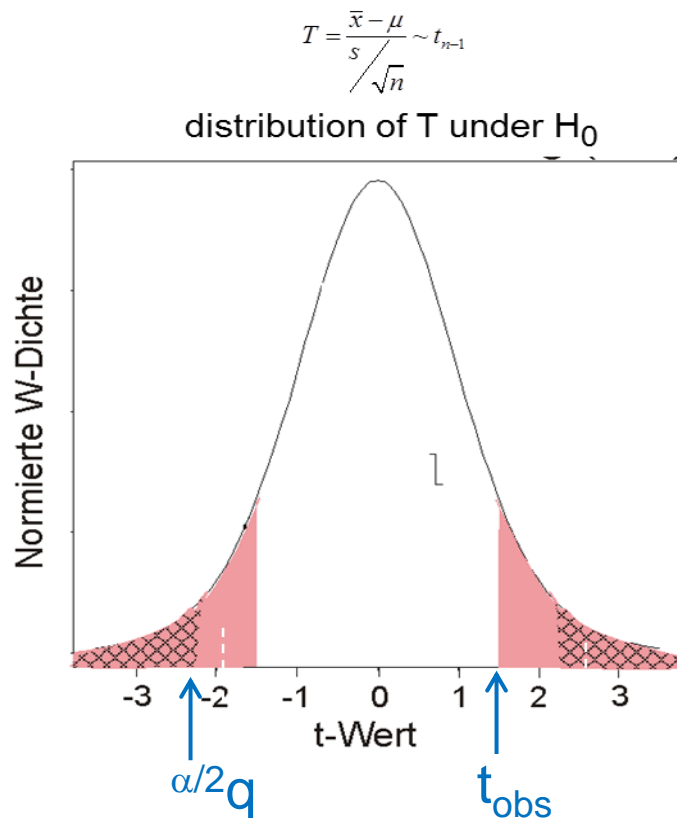
# Steps in a statistical test

1. Define your hypotheses (null  $H_0$ , alternative  $H_A$ ) and your significance level  $\alpha$  (the acceptable risk for an error of type 1)
2. Specify your test statistic  $T$  and its distribution under  $H_0$
3. Do an experiment, collect data, compute the value of  $T$ :  $t$
4. Use the position of  $t$  under the distribution of  $T$  under  $H_0$  to compute the p-value of what you observed (**p-value**: probability to get under  $H_0$  even more extreme  $t$ -values is than the observed one is called p-value)
5. Reject fail to reject (stay with)  $H_0$ . There is evidence that  $H_0$  can be rejected if  $p < \alpha$  or equivalently a CI that covers the Null parameter value.

# Interpretation of the p-value

The p-value corresponds to the probability to get an at least such extreme result as the seen one if we assume that the Null-Hypothesis is valid.  
(Therefore we reject  $H_0$  if this probability is small)

Graphically: the p-value corresponds to the area in the extreme tails (from the observed t-value outwards) under the density of the test-statistic distribution which is taken for a true  $H_0$ .



$$p = P(|t| \geq |t_c| \mid H_0 \text{ is true})$$
$$= P(p_{new} \leq p \mid H_0 \text{ is true})$$

p-value > 0.1 : no evidence for  $H_A$

p-value < 0.1 : weak evidence for  $H_A$

p-value < 0.05 : evidence for  $H_A$

p-value < 0.01 : clear evidence for  $H_A$

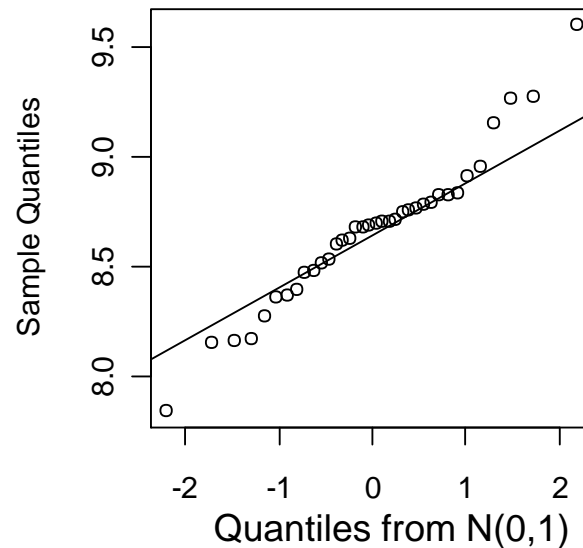
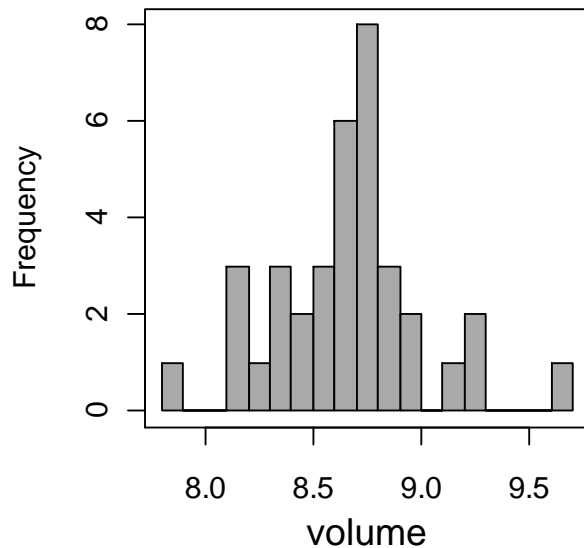
p-value < 0.001 : strong evidence for  $H_A$

# Type I and type II errors in a statistical test

Your Statistical Decision	True state of null hypothesis	
	$H_0$ True (example: the drug doesn't work)	$H_0$ False (example: the drug works)
Reject $H_0$ (ex: you conclude that the drug works)	<b><i>Type I error (<math>\alpha</math>)</i></b>	<b><i>Correct</i></b>
Do not reject $H_0$ (ex: you conclude that there is insufficient evidence that the drug works)	<b><i>Correct</i></b>	<b><i>Type II Error (<math>\beta</math>)</i></b>

## Example for a test problem

A new stem of bacteria was designed to produce a certain enzyme. A tube of bacteria can produce within 1 day in average a certain volume  $X$ . The vendor of these bacteria kit claims a volume of 8.2 ml per day. A purchaser wants to check this claim and measures for  $n=36$  tubes the produced volume within a day. He gets to the following results:



From data visualization roughly estimated:

$$\hat{\mu} = \bar{x} = 8.7$$

$$\hat{\sigma} = sd = 0.2$$

$$X_i \sim ? , \quad H_0 : ? , \quad H_A : ? , \quad T = ? , \quad T \sim ? , \quad {}^{95\%}VI = ?$$

# Example for a test problem

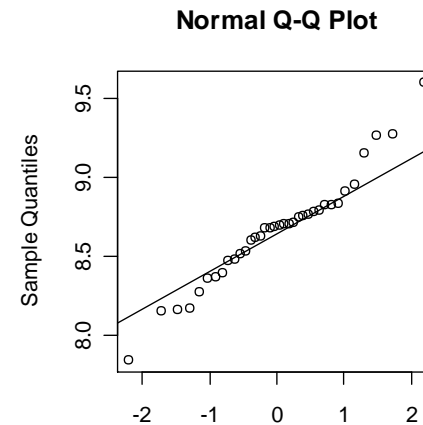
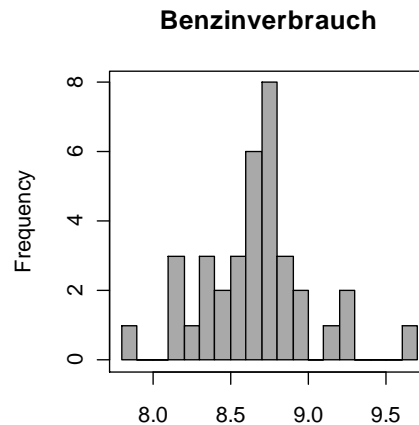
Postulated expected value  $\mu$  of 8.2 should be tested.

sample: daily-production of  $n=36$  tubes of bacteria

model for individual values:

$$X_i \text{ i.i.d. } X_i \sim N(\mu, \sigma^2)$$

model verification:



Null-hypothesis  $H_0: \mu = \mu_0 = 8.2$

Alternative-hypothesis  $H_A: \mu \neq \mu_0$

$$T = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \sim t_{n-1}$$

$${}^T VI^{1-a} = [-t_{35} q_{1-\frac{a}{2}}, t_{35} q_{1-\frac{a}{2}}] \Leftrightarrow$$

$$\begin{aligned} {}^m VI^{1-a} &= [\bar{x} - \frac{s}{\sqrt{n}} \cdot t_{35} q_{1-\frac{a}{2}}, \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{35} q_{1-\frac{a}{2}}] \\ &= [8.6, 8.8] \end{aligned}$$

$$8.2 \notin [8.6, 8.8]$$

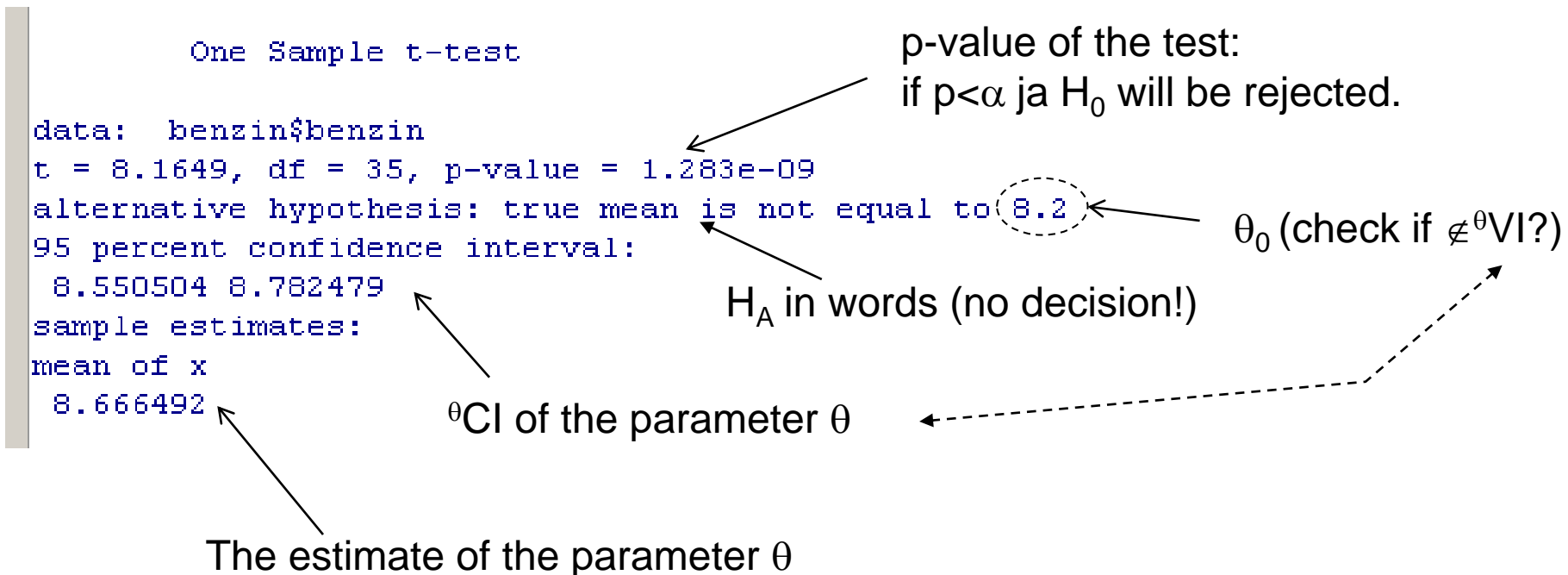
$$m_0 \notin {}^m VI^{1-a} \Rightarrow H_0 \text{ rejected}$$

Teststatistik	$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$	Einstichproben- $t$ -Test
---------------	--	---------------------------

## The one-sample t-test in R

The name t-test comes from the use of t-distribution for test statistic T. The most important results are the CI for the parameter and the p-value.

```
>t.test(volume, alternative="two.sided", mu=8.2, conf.level=0.95 )
```



# Historical Excursion: Who has invented the t-test?



... the Guinness brewery in Dublin, Ireland.

The t-test was a by-product of Student / Gosset's efforts to select the best yielding varieties of barley.

S-PLUS - Commands

File Edit View Insert Data Statistics Graph Options Window Help

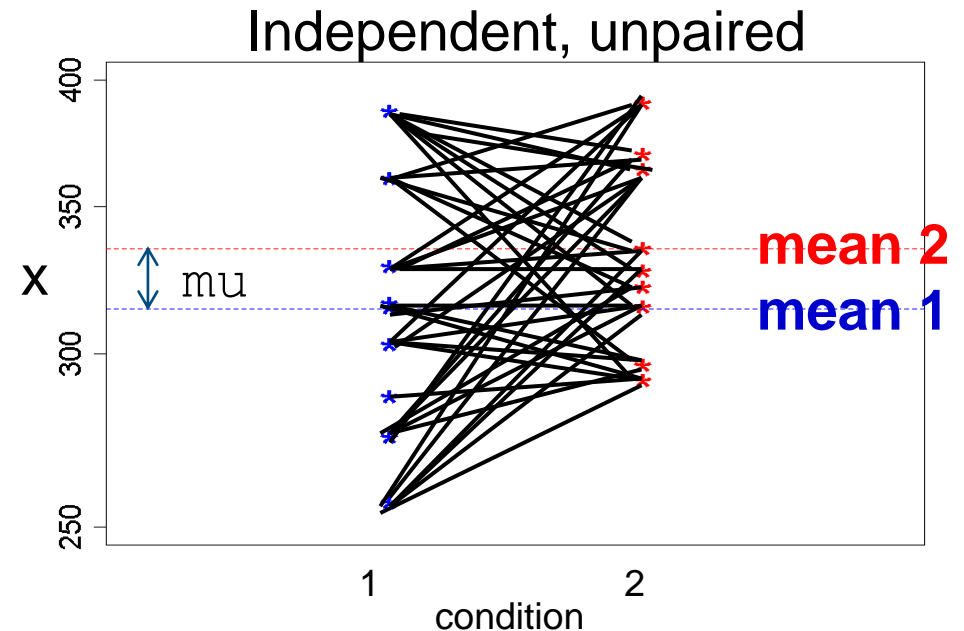
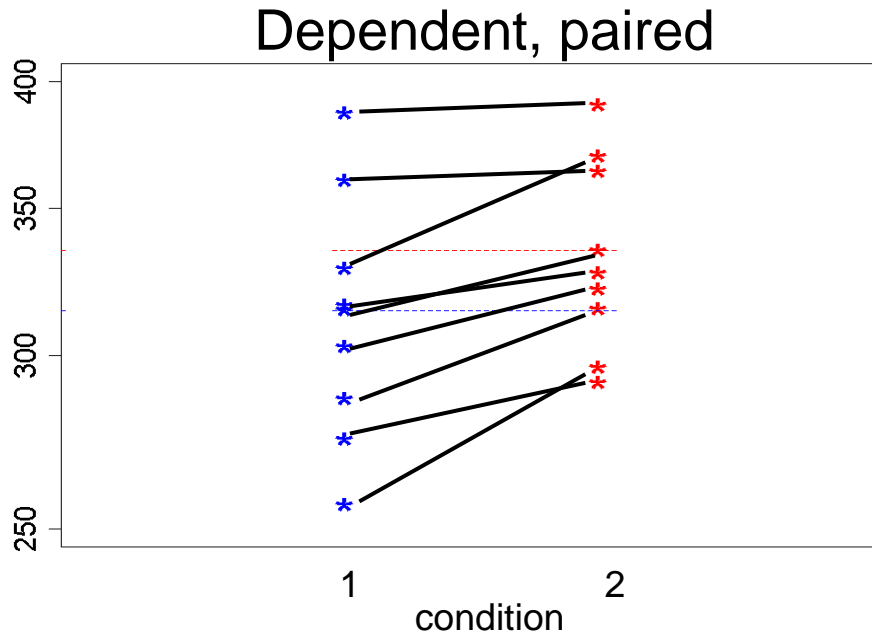
Commands

```
> barley
```

	yield	variety	year	site
1	27.00000	Manchuria	1931	University Farm
2	48.86667	Manchuria	1931	Waseca
3	27.43334	Manchuria	1931	Morris
4	39.93333	Manchuria	1931	Crookston
5	32.96667	Manchuria	1931	Grand Rapids
6	28.96667	Manchuria	1931	Duluth
7	43.06666	Glabron	1931	University Farm
8	55.20000	Glabron	1931	Waseca
9	28.76667	Glabron	1931	Morris
10	38.13333	Glabron	1931	Crookston
11	29.13333	Glabron	1931	Grand Rapids
12	29.66667	Glabron	1931	Duluth
13	35.13333	Svansota	1931	University Farm
14	47.33333	Svansota	1931	Waseca



# Is there a significant difference between 2 groups? What are paired/dependent samples?

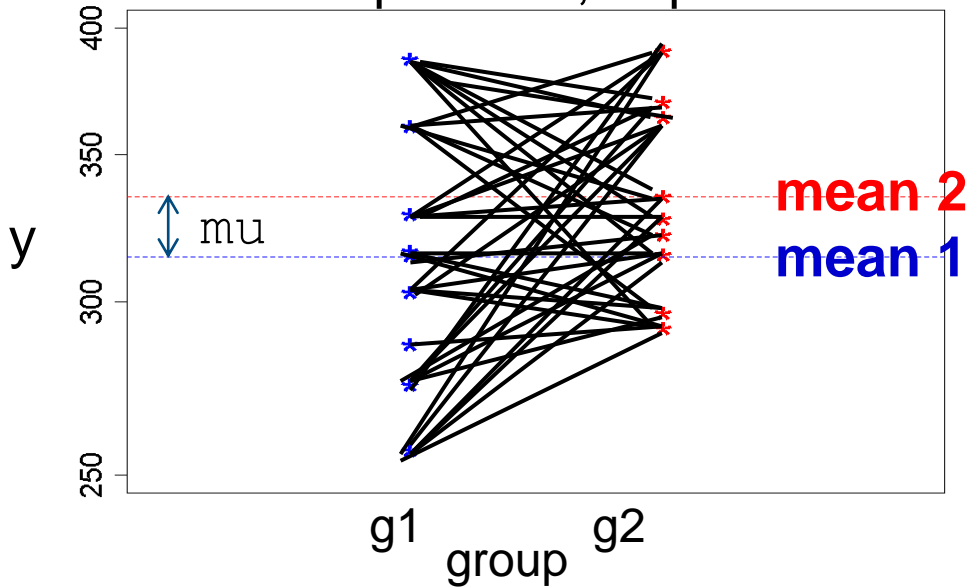


In a **paired design** always a pair of values from group1 and group2 correspond to each other (often 2 treatments were applied to each unit or person)  $\sim n_1 = n_2$   
In a paired design we test if the population **differences within pairs** are zero ( $\mu = 0$ ).

In a **unpaired design** we test if the population means of **two independent samples**, e.g. corresponding to 2 treatment groups, are different ( $\mu \neq 0$ ). The 2 groups might have different sizes.

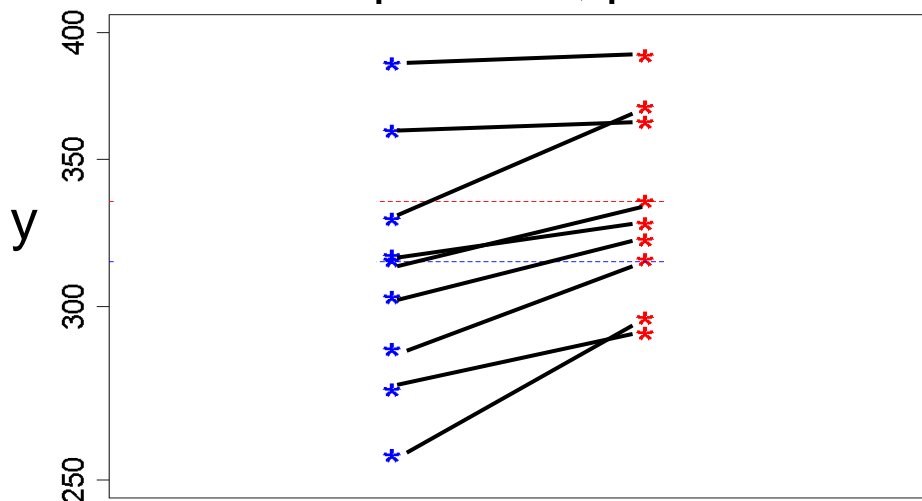
## Unpaired and paired data with continuous outcome

Independent, unpaired



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

Dependent, paired

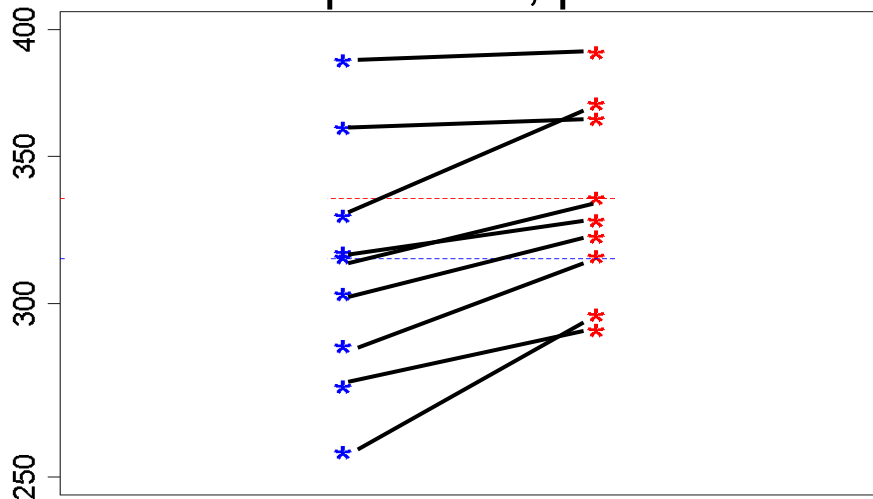


```
t.test(g1,g2, mu=0,  
       var.equal=T, paired=TRUE)
```

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

# Pros and Cons of a paired compared to unpaired design

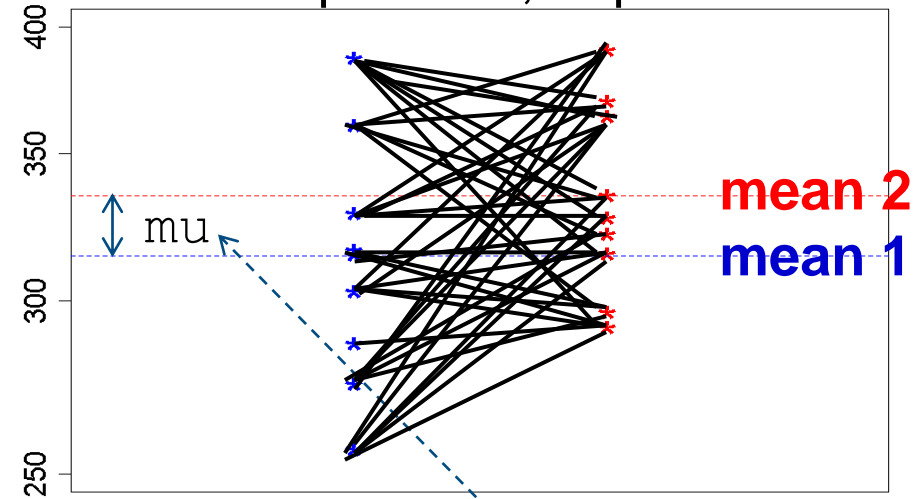
Dependent, paired



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

$$T = \frac{\bar{\Delta}_{pair}}{se(\bar{\Delta}_{pair})} \sim t_{(n_i-1)}$$

Independent, unpaired



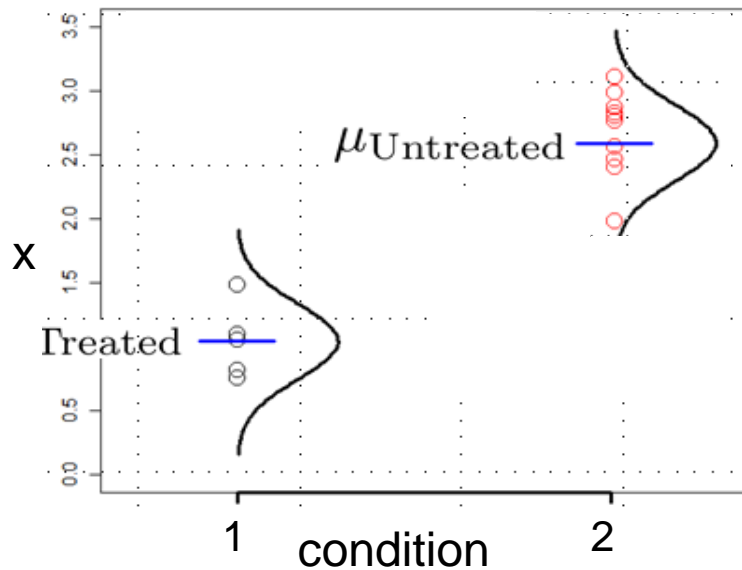
`t.test(g1, g2, mu=0, paired=FALSE)`

$$T = \frac{\bar{X}_1 - \bar{X}_2}{se_{pooled}} \sim t_{(n_1+n_2-2)}$$

- + In a paired design we can exclude the individual differences of the investigated persons or units and therefore **a paired design is preferable in cases where the individual differences are bigger than the treatment effect.**
- + We need less persons (observations units) to enroll for the same total size  $n=n_1+n_2$
- If the total number of observations  $n=n_1+n_2$  is the same and the effect size is much larger than the individual differences, then the standard error of the estimated group difference is larger in the paired design (compared to the unpaired design) since it relies on less comparisons.

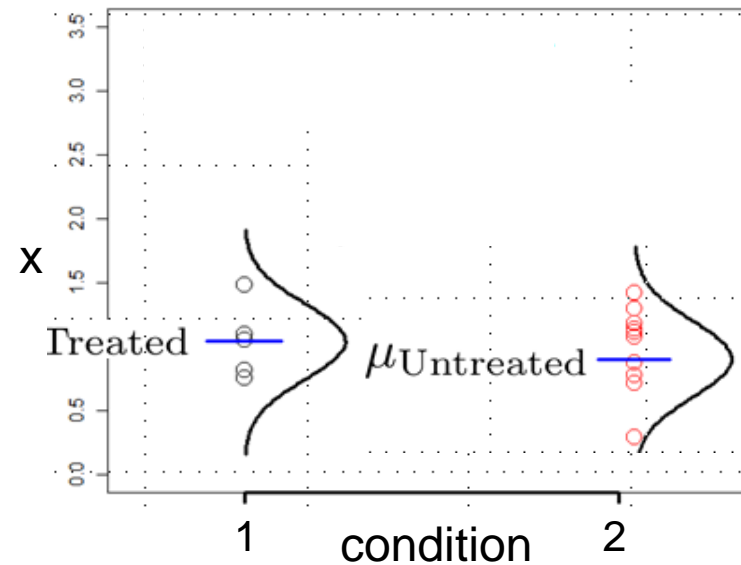
# Test for differences between treatments, unpaired design

Drug is effective



$p=0.0001$

drug fails



X: outcome of interest should be normally distributed →

- t-Test (comparison of 2-conditions)

$$T = \frac{\bar{X}_1 - \bar{X}_2}{se_{pooled}} \sim t_{(n_1+n_2-2)}$$

equal variance (t-test):  $se_{pooled} = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}$

different variance (Welch test):  $se_{pooled} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

## CI interpretation in case of a unpaired group comparison

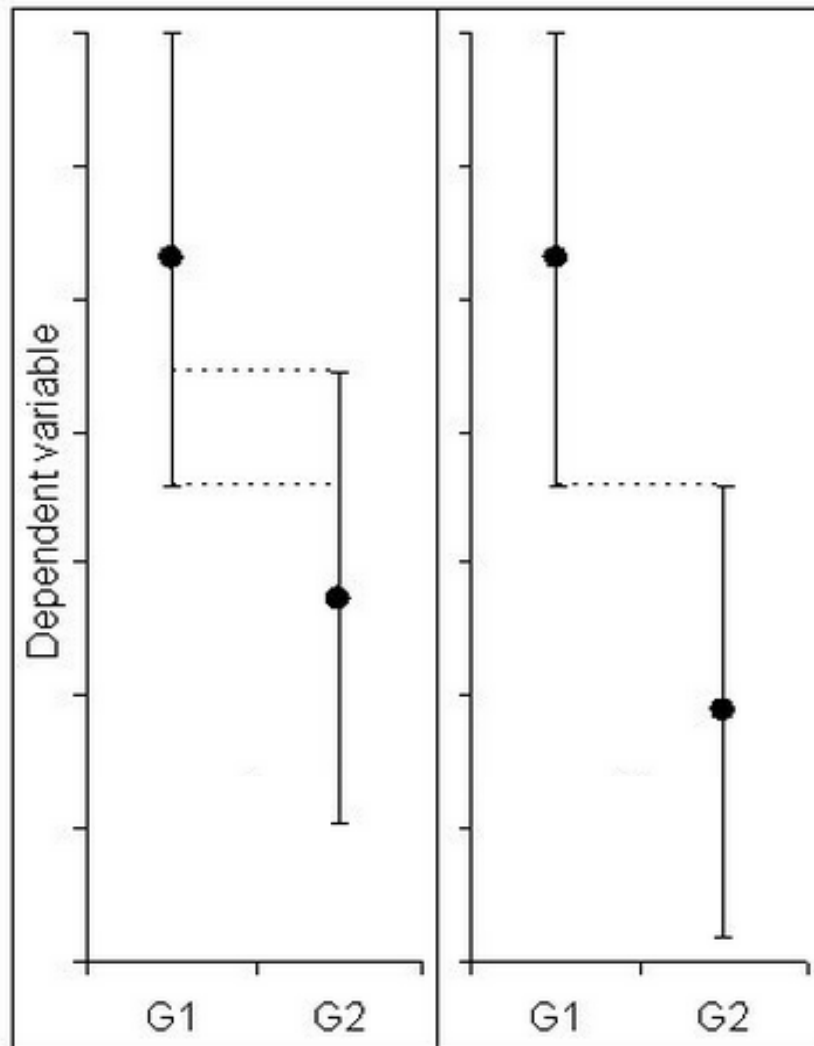
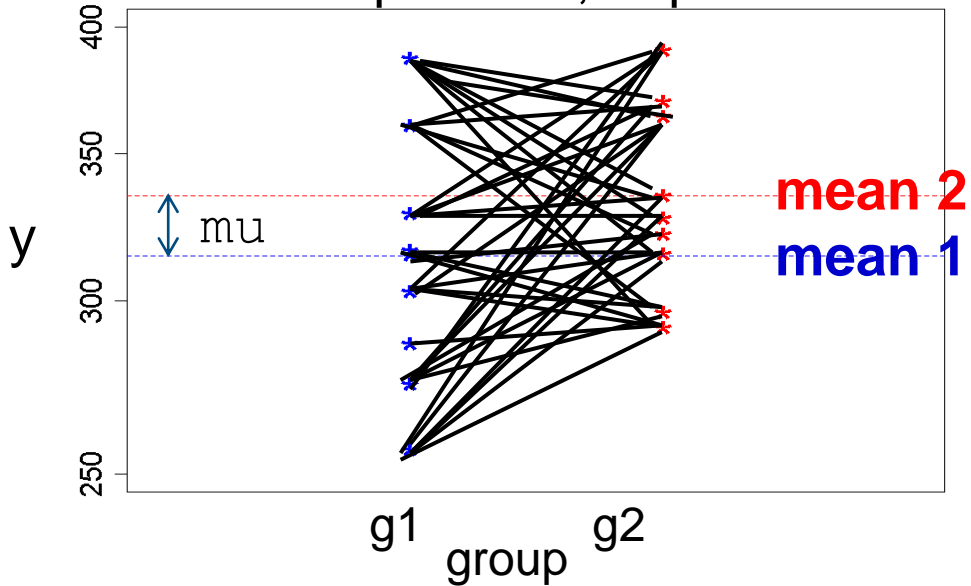


Figure 2: The CIs on the left overlap by about 1/4, half the average margin of error, which corresponds to a p value of  $\approx .05$ . The CIs on the right are just touching. This corresponds to a p value of  $\approx .01$  (Cumming and Finch, 2005).

If the 95%-CI of two populations means (derived from independent samples) are overlapping less than 25% then the difference is **significant, i.e. there is a high data based evidence for a real difference which is not due to sample variation.**

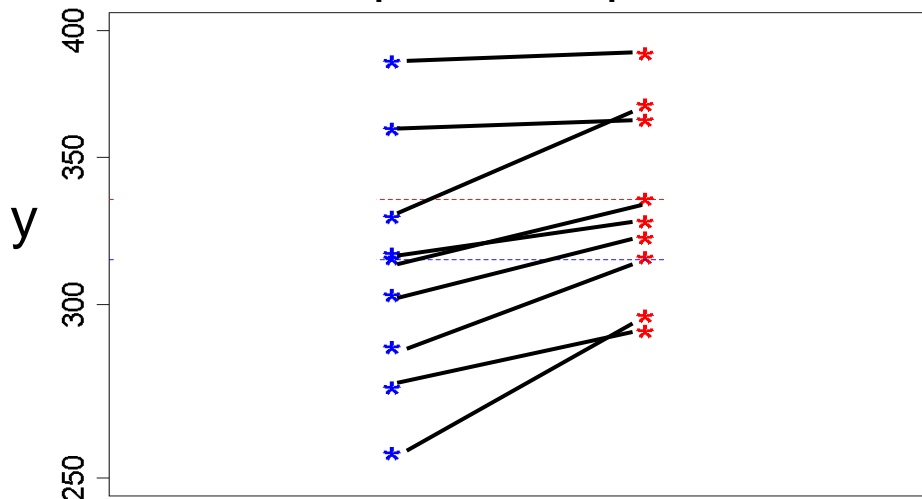
## Reminder: Unpaired and paired t-test on location

Independent, unpaired



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

Dependent, paired



```
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 caffeine 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
```

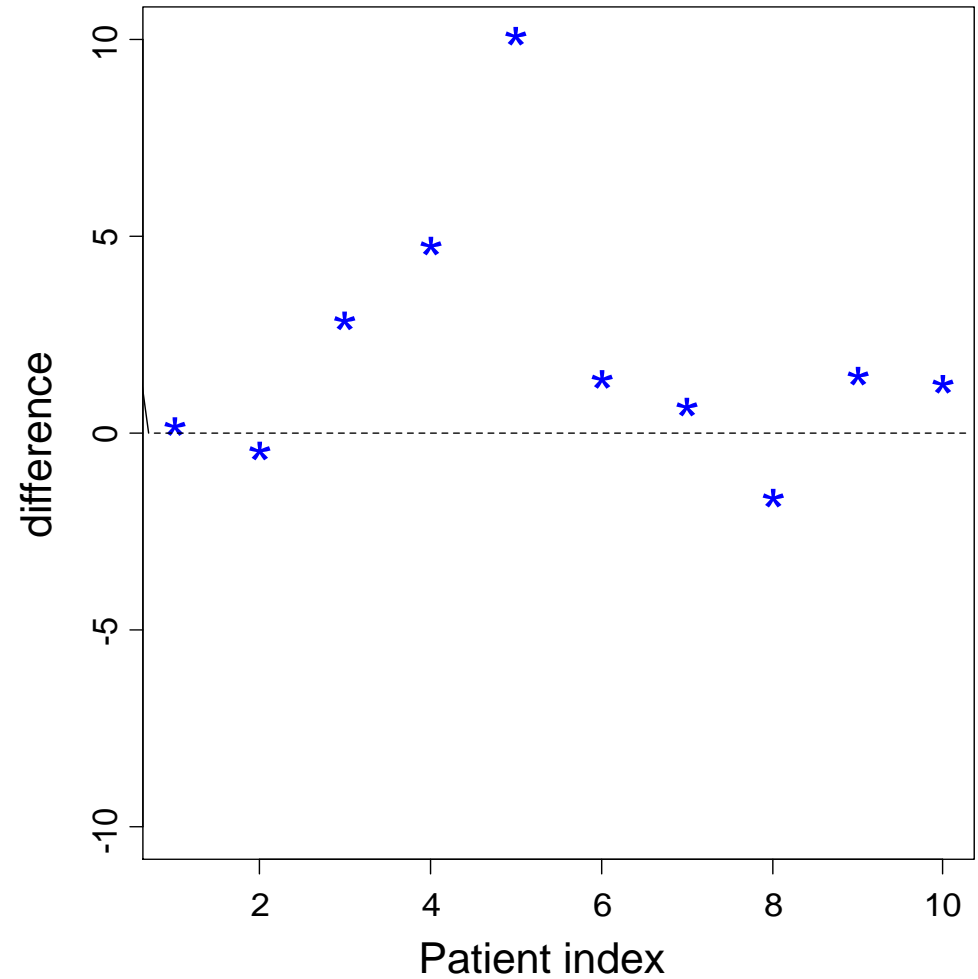
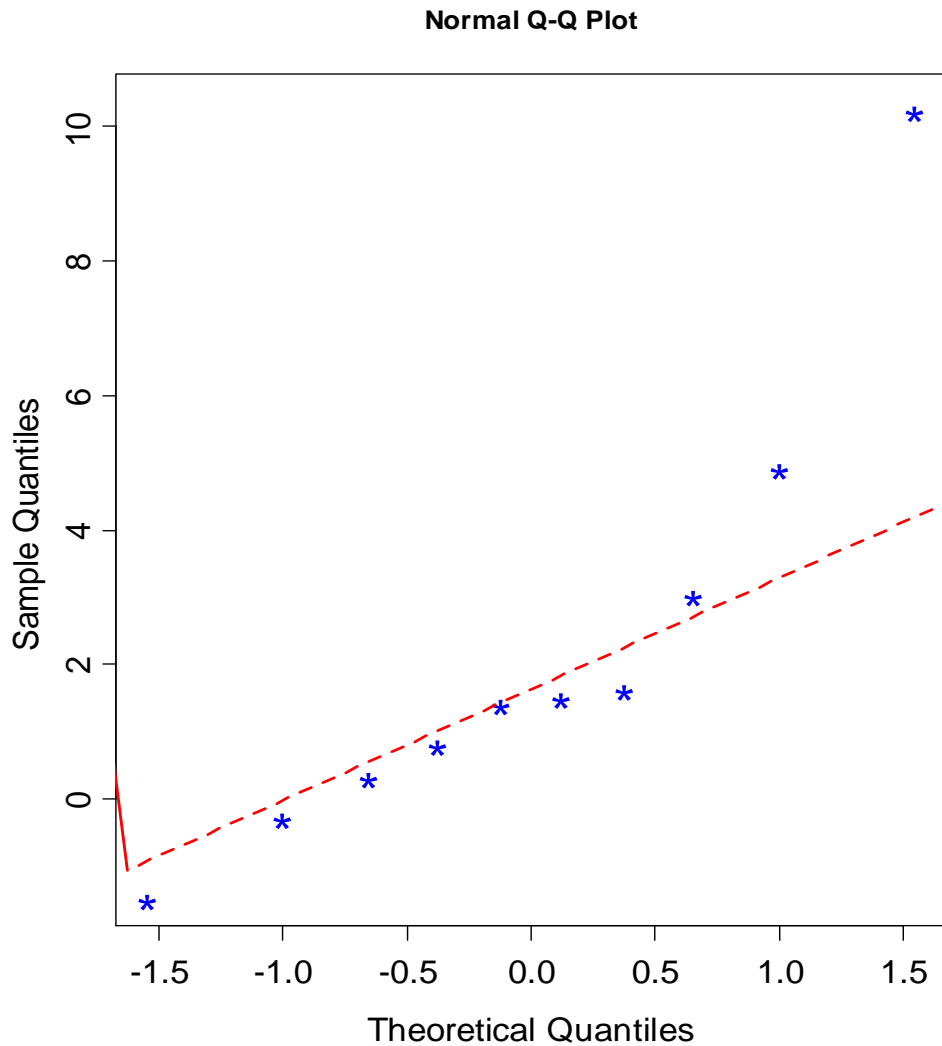
```
sample estimates:
```

```
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

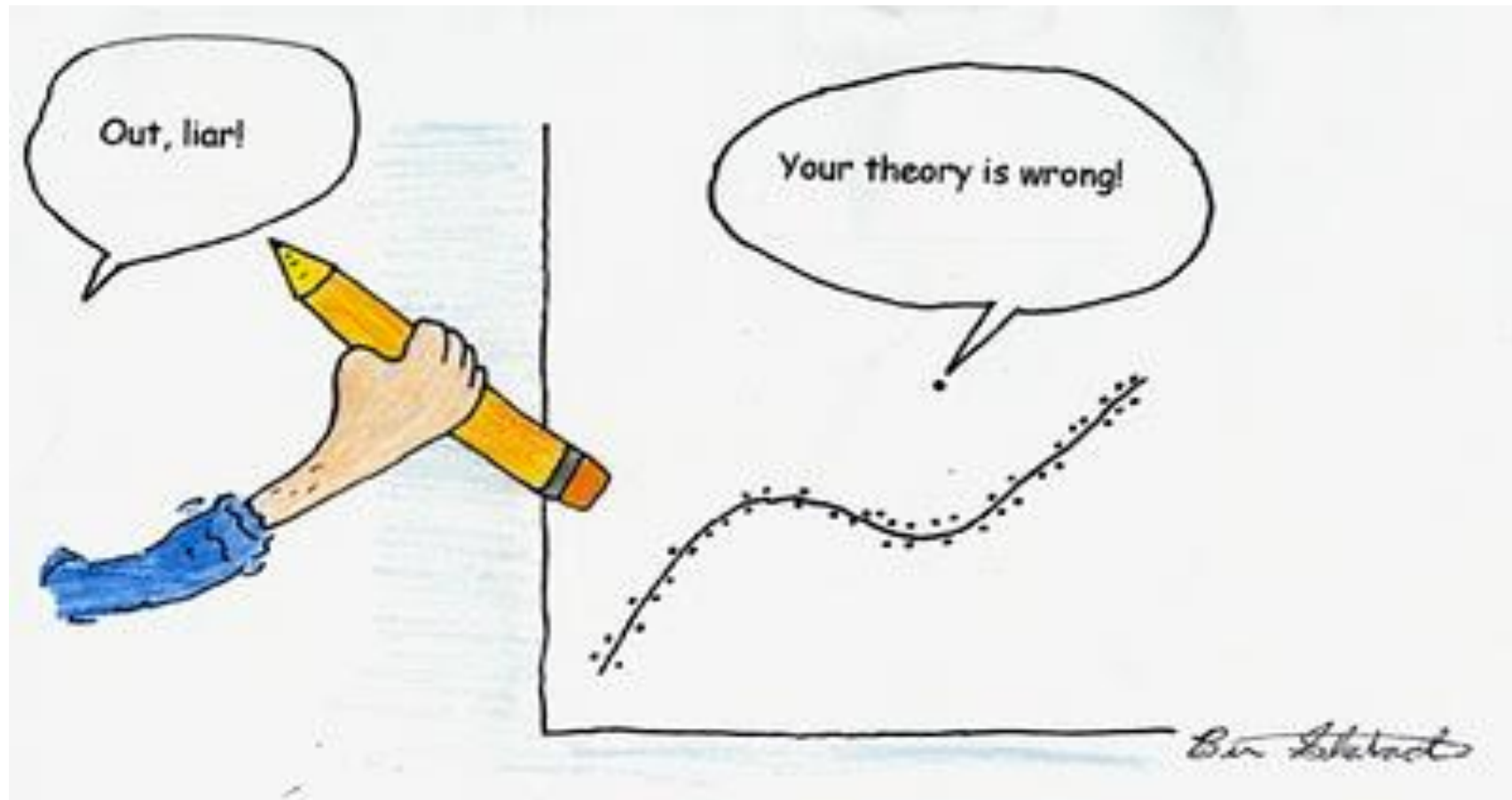
# Visualization of the data



**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 and 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.

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

*Teststatistik* :  $U = \min(U^+, U^-)$

Under  $H_0$ :

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

$$\text{reject } H_0, \text{ if } U << \frac{1}{2} \cdot \frac{n}{2} \cdot (n+1)$$

# 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
```

```
> 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
```

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).

$p < 5\% \rightarrow H_0$  is rejected and we have shown a significant effect of coffee 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** ( $< \approx 10$ ) 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.

# 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** ( $< \approx 10$ ) 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

unpaired

form differences and treat them  
as each value's

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

yes

no

t-Test ( $s$  estimated)  
z-Test ( $\sigma$  is known)  
for one sample

Are values symmetrically  
distributed (always for differences)?

yes

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

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

yes

no

t-Test ( $s$  estimated)  
z-Test ( $\sigma$  is known)  
for unpaired  
sample

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