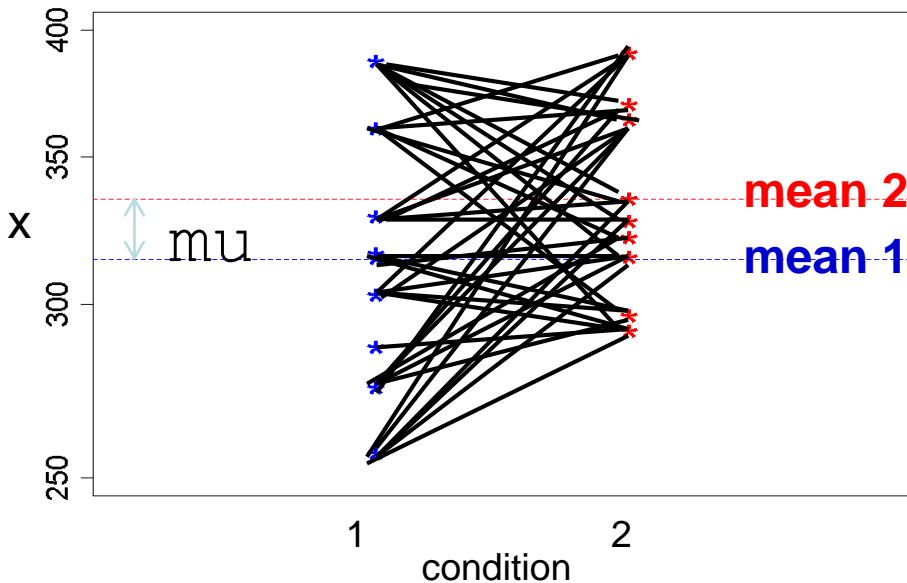


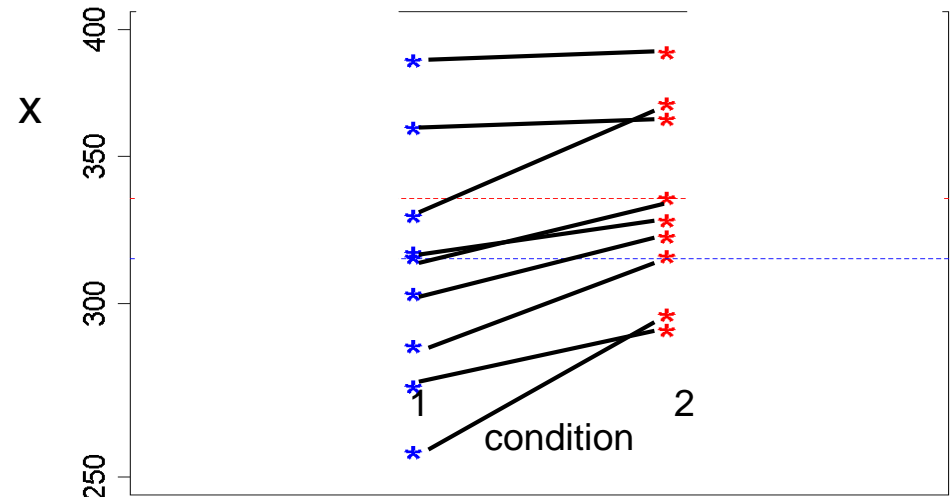
Biostatistics Week 4

- Confidence interval, significance (\neq relevance)
- Testing
 - t-test for means
 - 2 group comparison: paired and unpaired tests
- Outlook to non-parametric tests: Wilcoxon

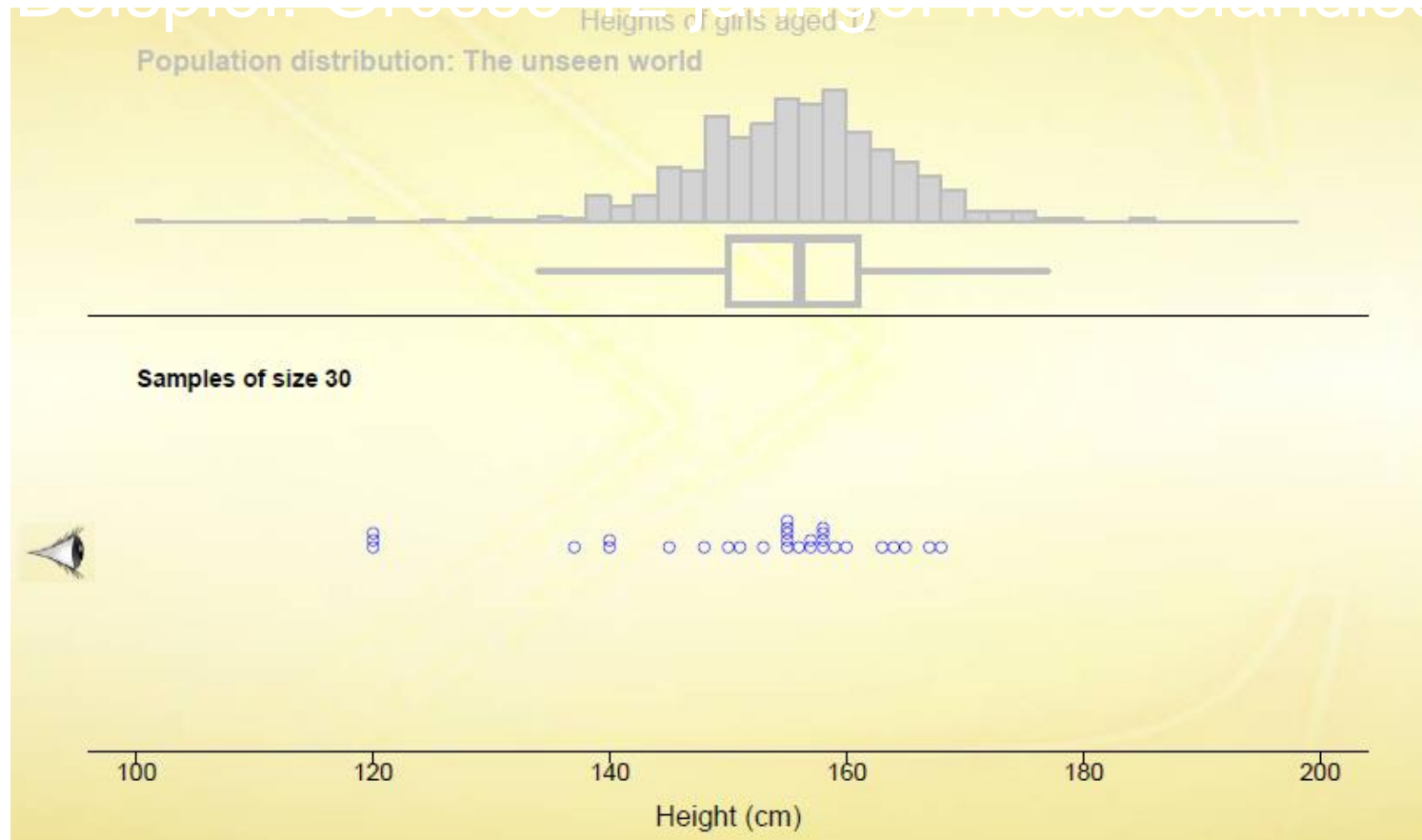
Independent, unpaired



Dependent, paired



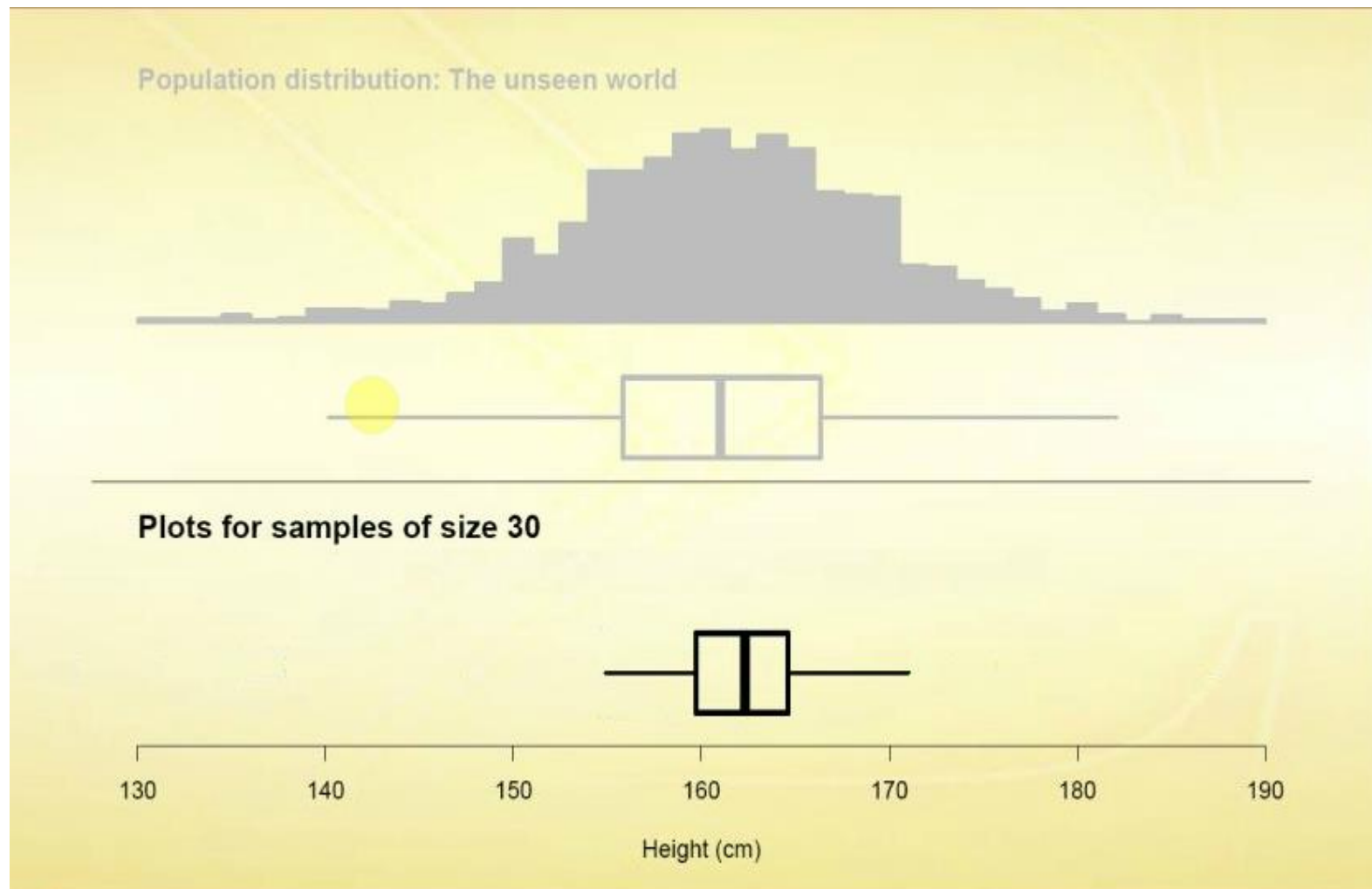
Where is the center of the population?



$n=30$

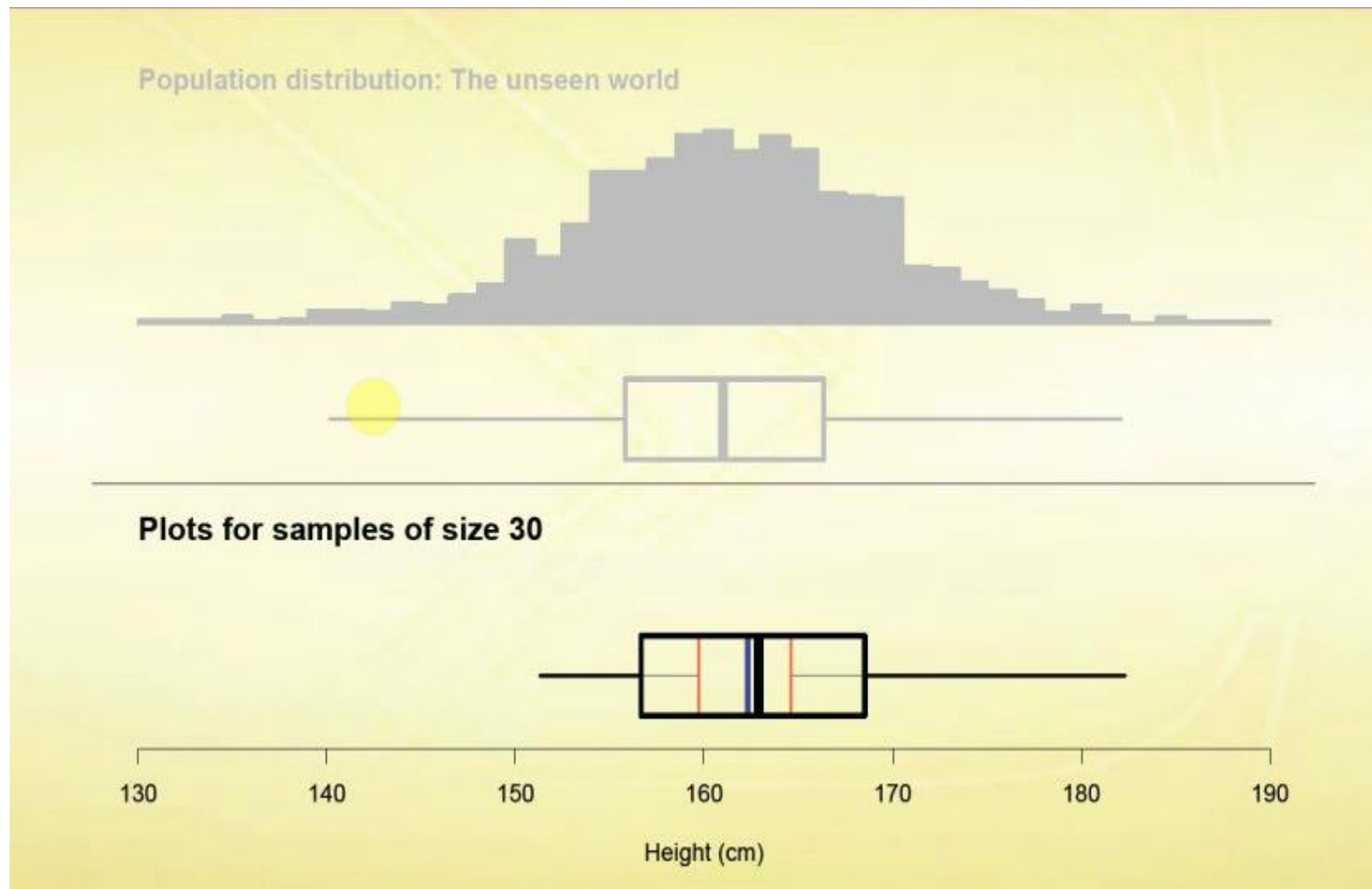
CtsVar_1samp_Dots30.pdf

Where is the center of the population?



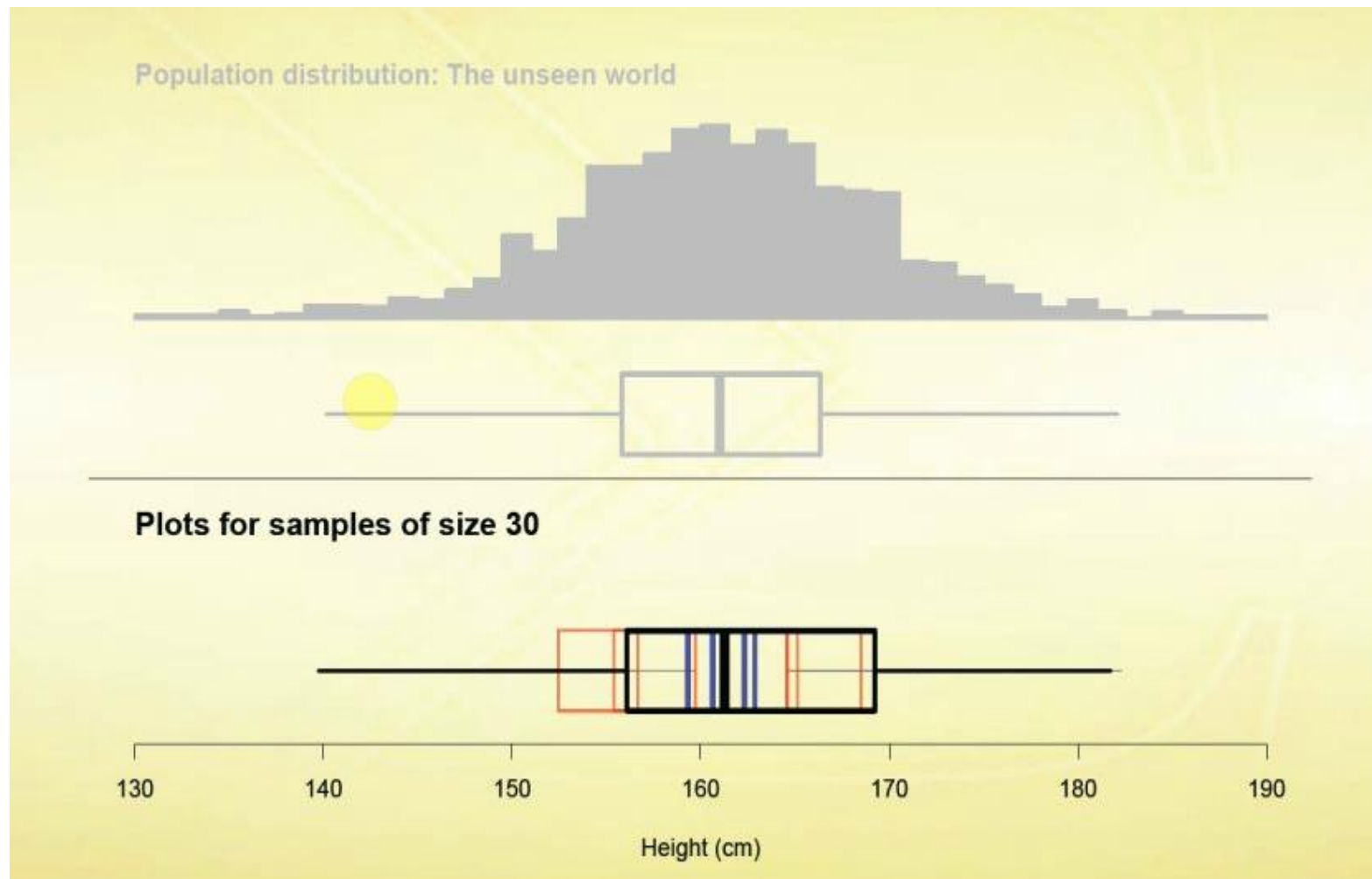
Visualize boxplot with memory

Where is the center of the population?



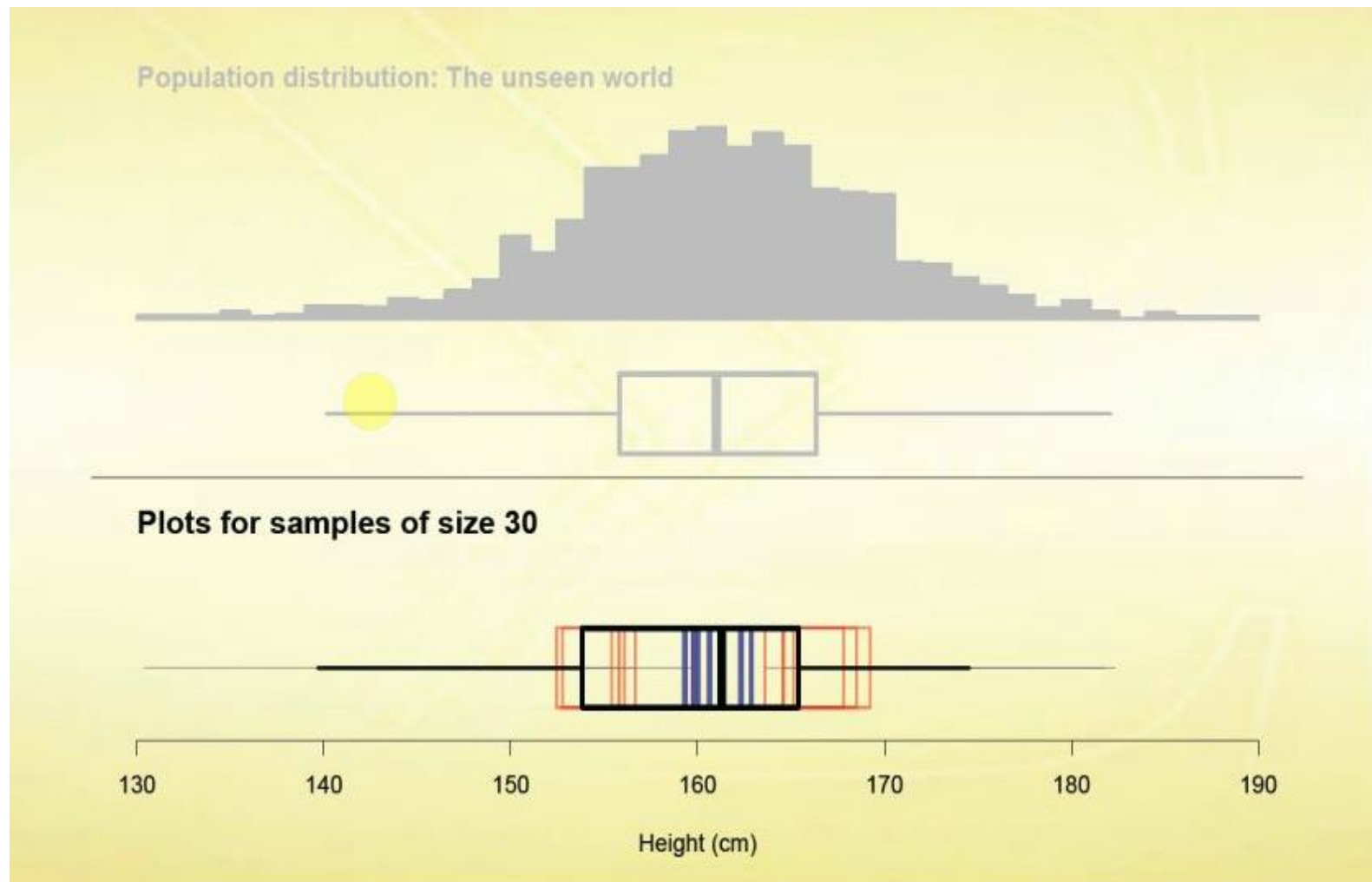
Visualize boxplot with memory

Where is the center of the population?



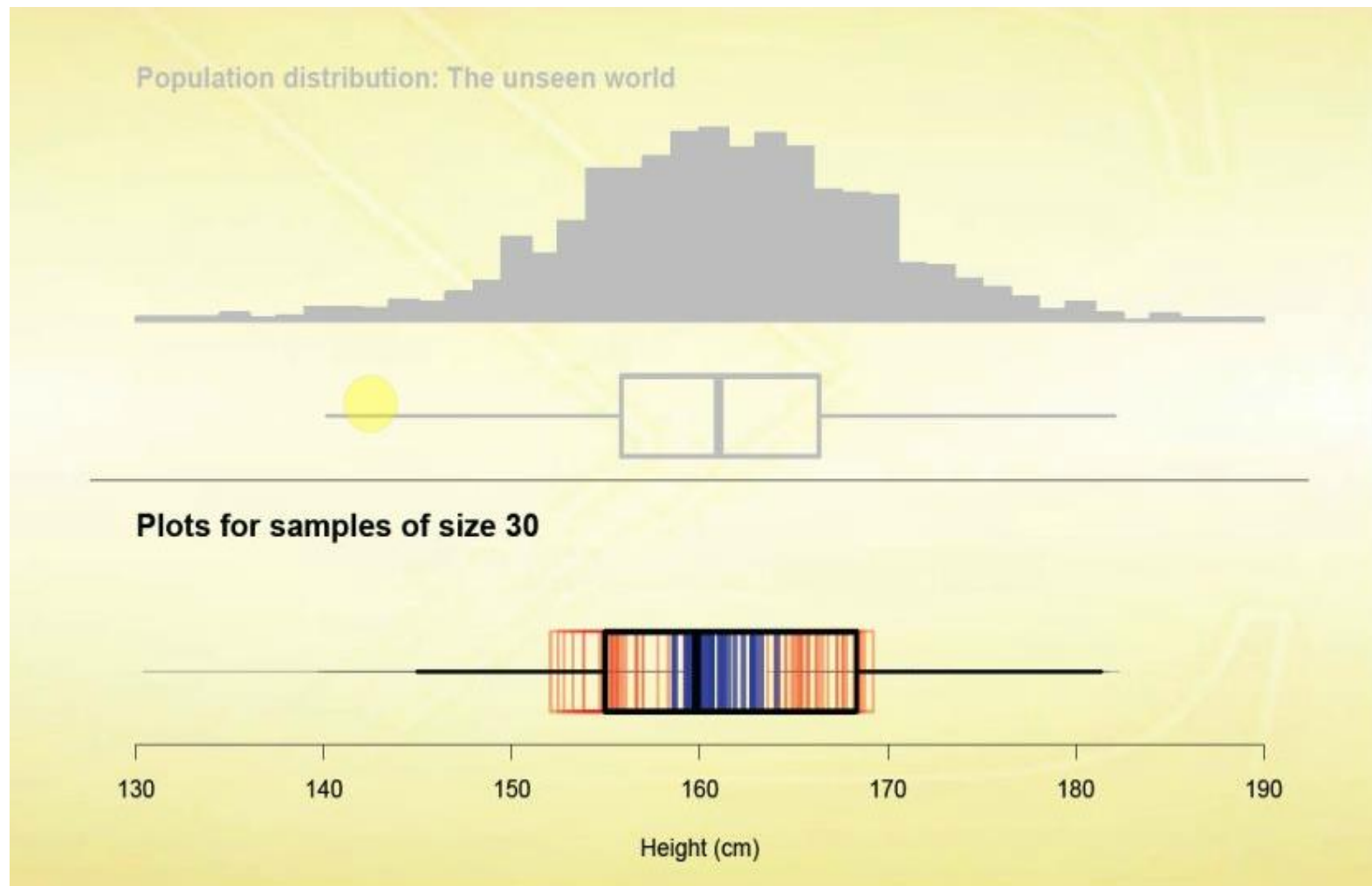
Visualize boxplot with memory

Where is the center of the population?



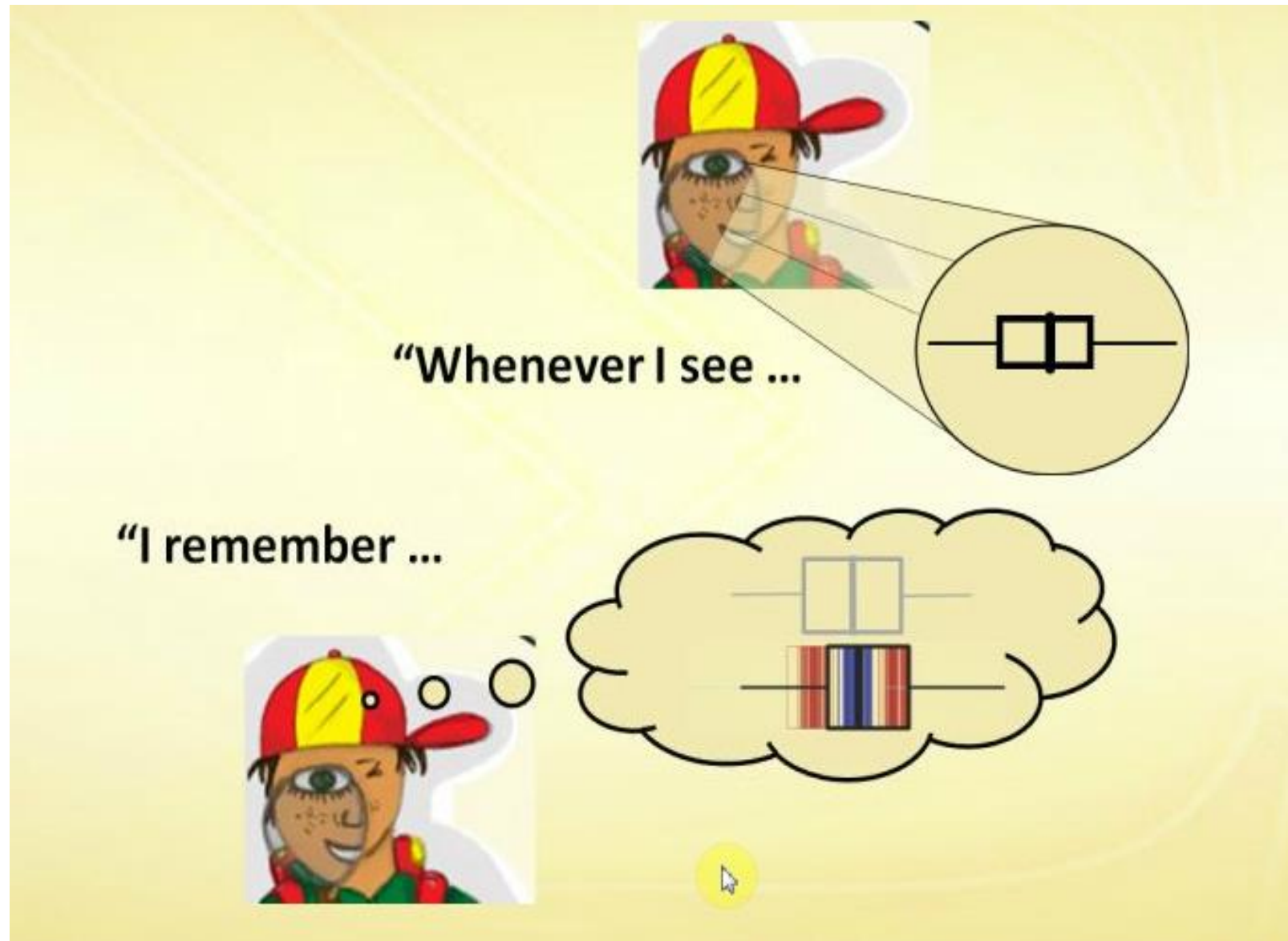
Visualize boxplot with memory

Where is the center of the population?

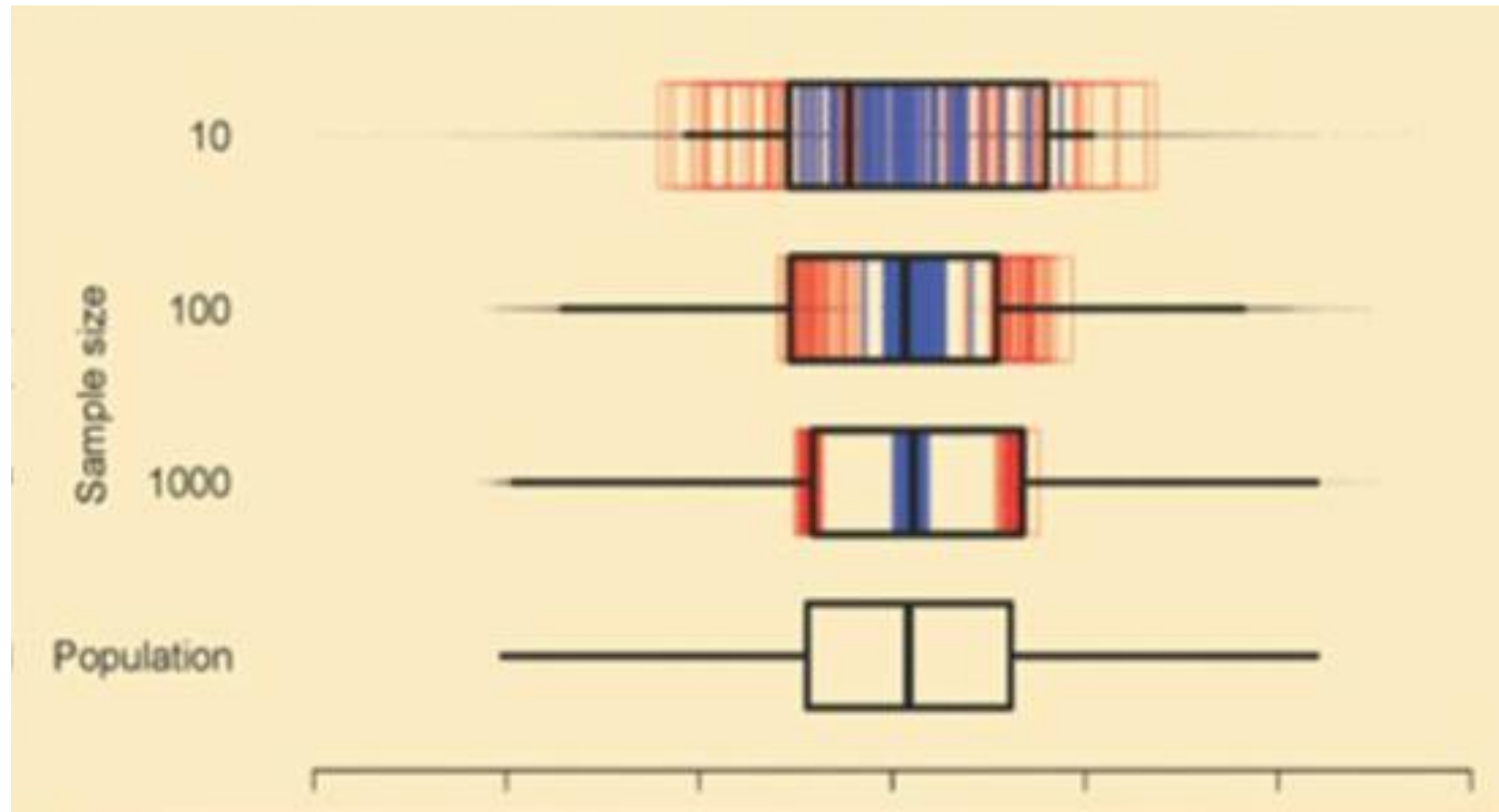


Vizualize boxplot with memory

Where is the center of the population?



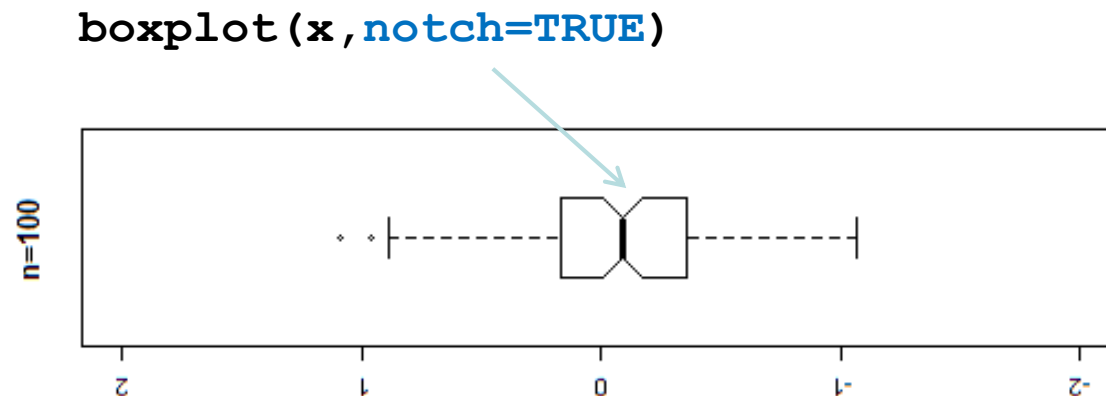
Where is the center of the population?
We get more certain with increasing sample size



How sure can I be about the true parameter value?

Goal:

We would like to determine from our sample/observations an interval, which covers the true parameter value with a probability of 95%.



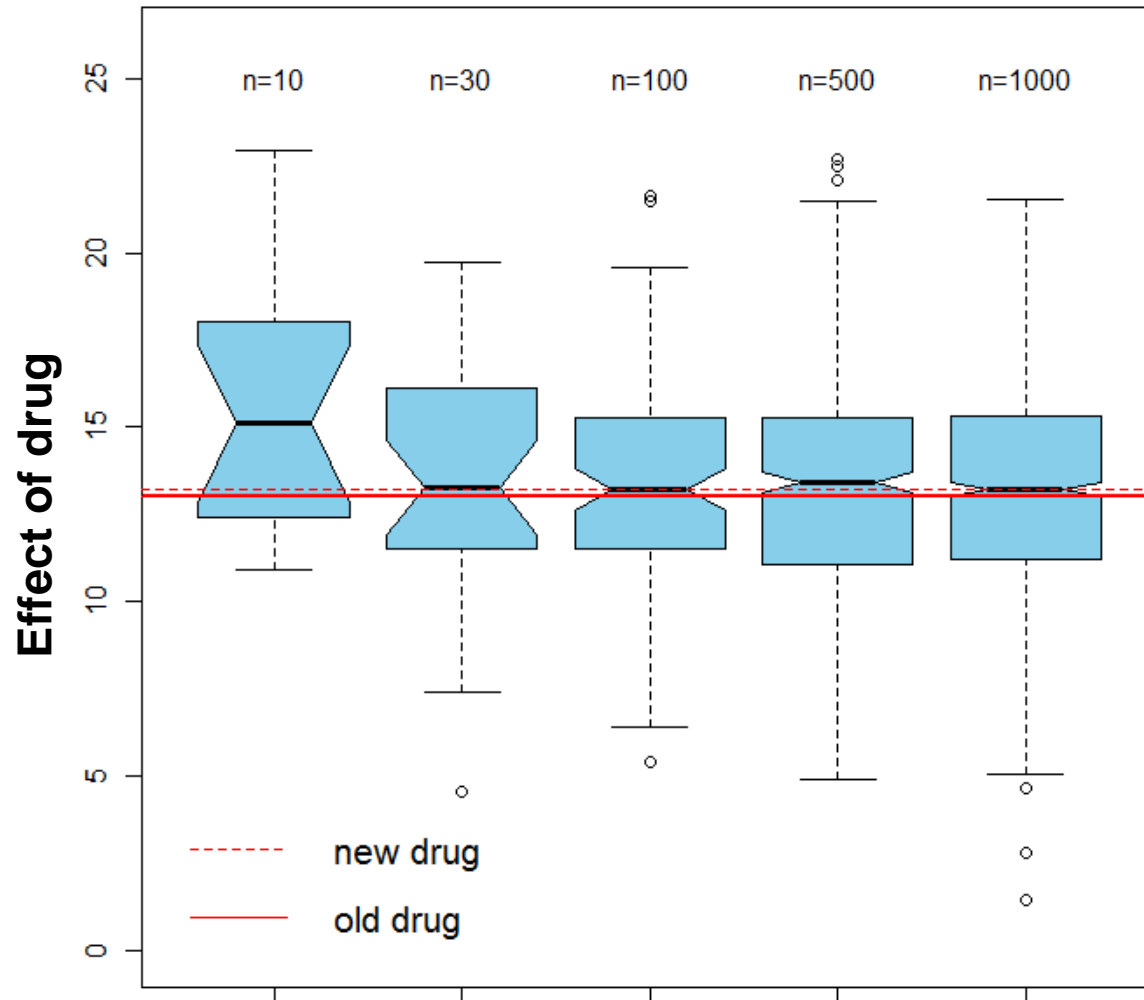
$\pm 1.58 \text{ IQR} / \sqrt{n}$

The notch
covers the
population
median «quite
certain»

Significance does not imply relevance

Everything gets significant if the sample size is large enough

Sample with different sample sizes drawn from a Normal distribution with expected value of 13.1



$$H_0: \mu_0 = \mu_{\text{old-drug}} = 13$$

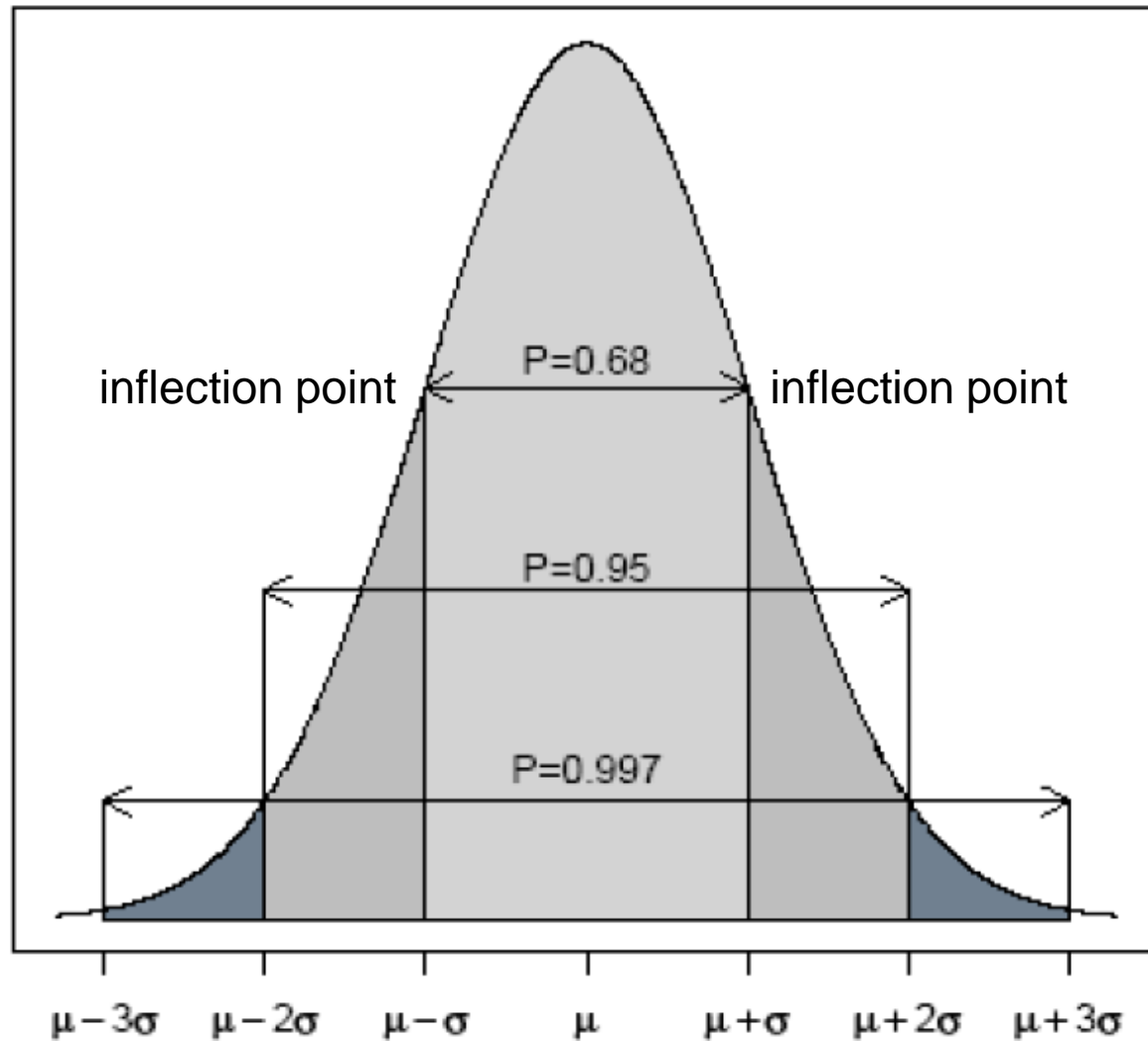
$$H_A: \mu > 13$$

Assume true median of the new drug is 13.1 which would be no relevant improvement compared to old drug value 13

- To ensure relevance of an significant test results should form
- Non-significance could be caused by small sample size. 95%

Density of the Normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}$$



Rule of thumb:

A random value $X \sim N(\mu, \sigma^2)$ has 95% of its probability mass within the following interval:

$$[\mu - 2\sigma, \mu + 2\sigma]$$

Or equivalently:

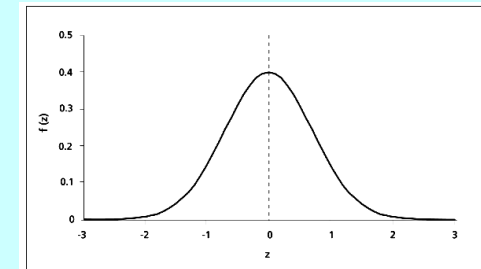
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 95\%$$

Distribution of the standardized mean in case of normal distributed observations

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

Variance σ_x^2 is known.

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



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

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

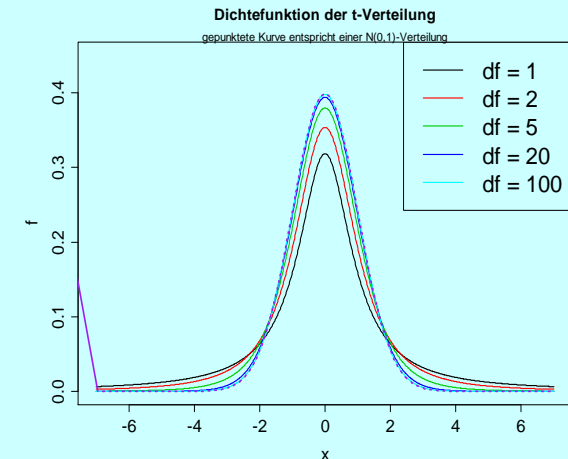
Variance σ_x^2 is unknown and is estimated from the data

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

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

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

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 $t_{df=n-1}$

Construct an exact confidence interval for the expected value if values are normally distributed

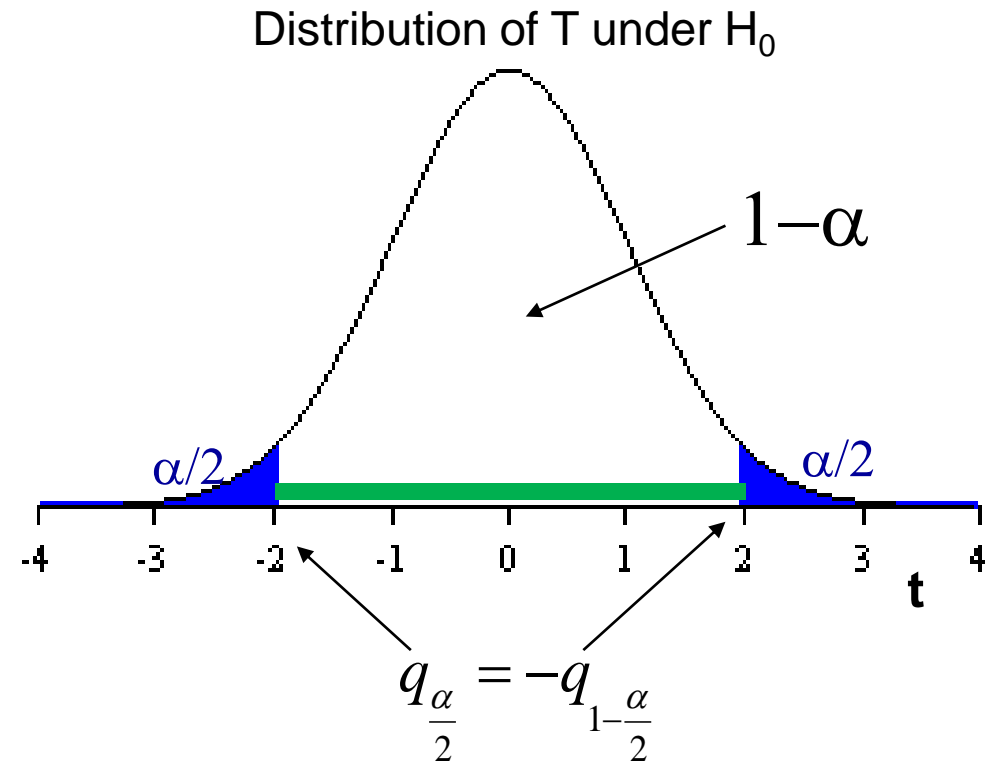
$$X_i \text{ i.i.d. } \sim N(\mu, \sigma^2), E(X) = \mu_x, \text{Var}(X) = \sigma_x^2$$

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

$$P\left(q_{\frac{\alpha}{2}}^t \leq \frac{\bar{X} - \mu_x}{\frac{\sigma_x}{\sqrt{n}}} \leq q_{1-\frac{\alpha}{2}}^t\right) = 1 - \alpha$$

$$P\left(-\frac{\sigma_x}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}^t \leq \bar{X} - \mu_x \leq \frac{\sigma_x}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}^t\right) = 1 - \alpha$$

$$P\left(\bar{X} - \frac{\sigma_x}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}^t \leq \mu_x \leq \bar{X} + \frac{\sigma_x}{\sqrt{n}} \cdot q_{1-\frac{\alpha}{2}}^t\right) = 1 - \alpha$$



95% CI for μ_x $\left[\bar{X} - \frac{\sigma_x}{\sqrt{n}} \cdot q_{0.975}^{t_{n-1}} ; \bar{X} + \frac{\sigma_x}{\sqrt{n}} \cdot q_{0.975}^{t_{n-1}} \right]$

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

$$\bar{x} \pm t_{n-1} q_{97.5\%} \cdot \frac{sd(x)}{\sqrt{n}} \stackrel{n>25}{\approx} \bar{x} \pm z q_{97.5\%} \cdot \frac{sd(x)}{\sqrt{n}}$$

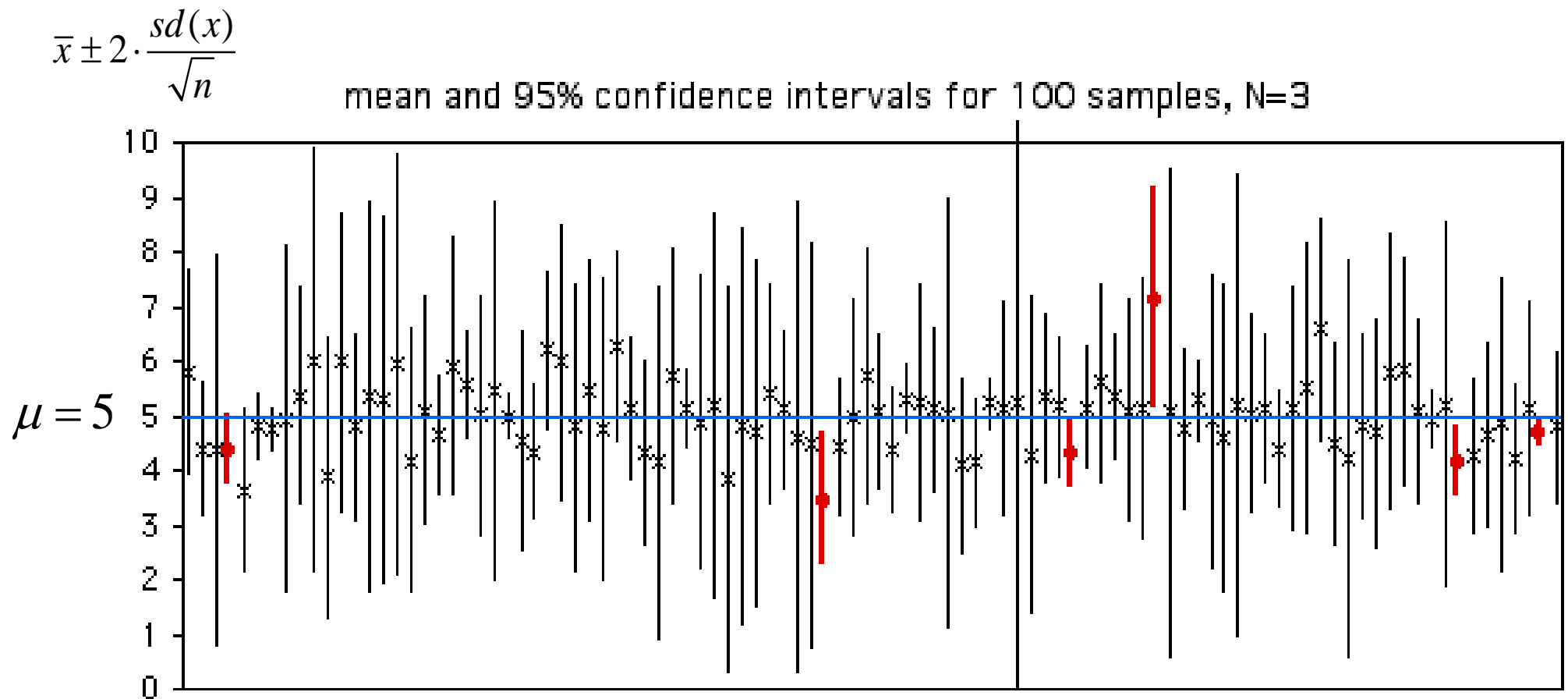
$se(\bar{x})$: standard error of the mean

Please note that for this CI 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.

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

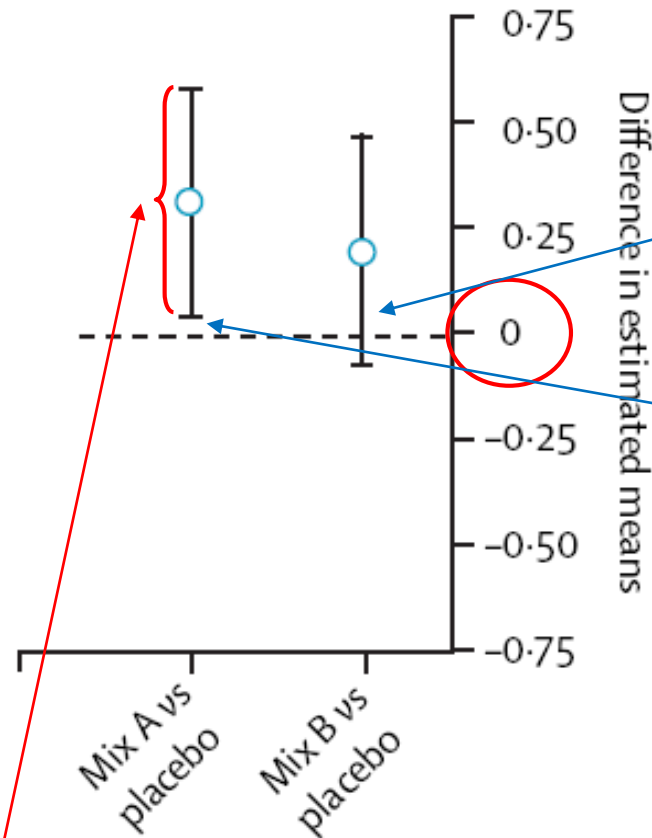


95 out of 100 95%-CI for μ do cover the true population parameter $\mu=5$ when simulating 100 random samples from a population following $N(\mu=5, \sigma^2)$.

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.

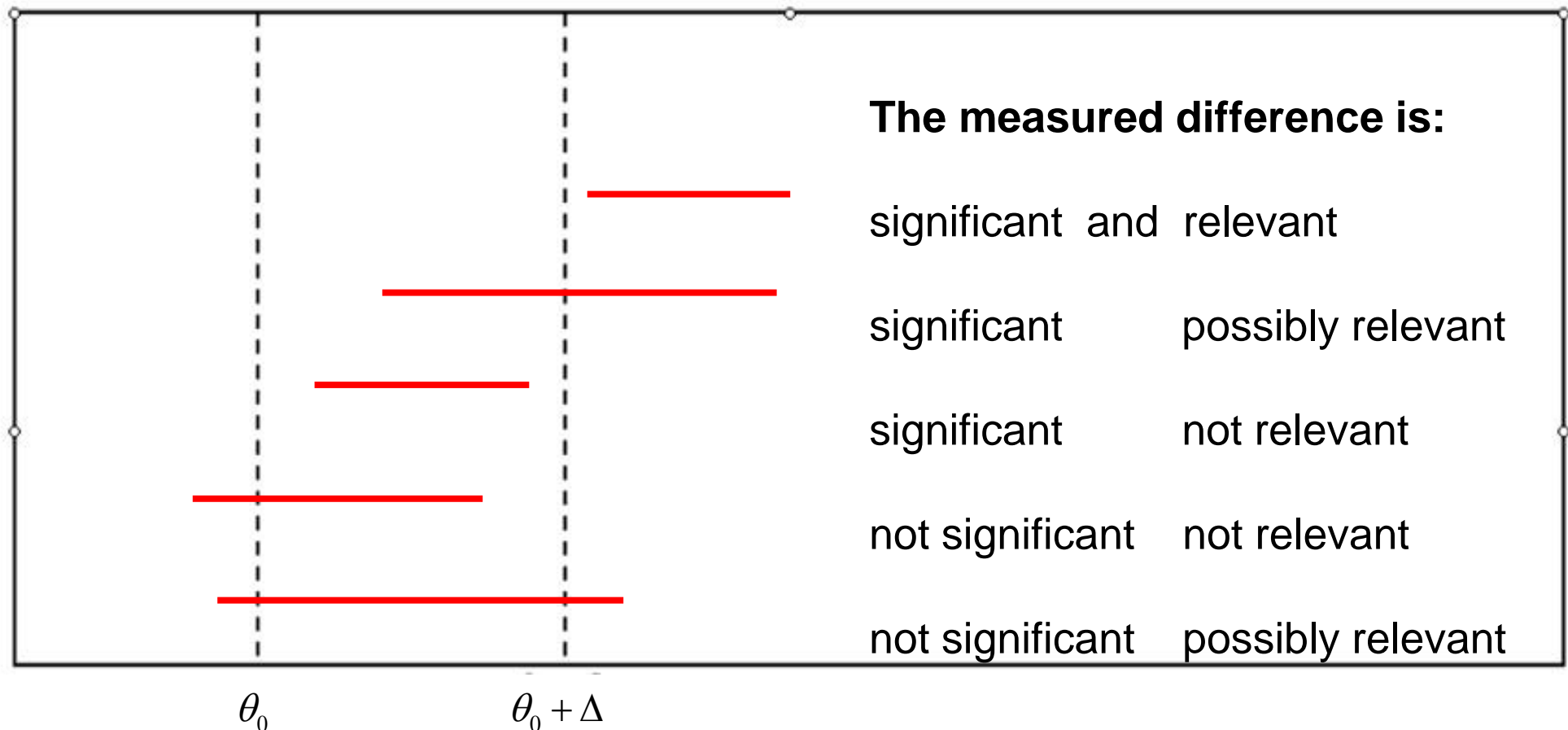


- The CI covers all plausible values for the true mean-difference – here the true treatment effect
- If 0 is covered by the CI it is plausible that the treatment effect is 0 – we have no evidence against H_0 , that the treatment has no effect.
- If 0 is **not** covered by the CI, we say that the treatment 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.

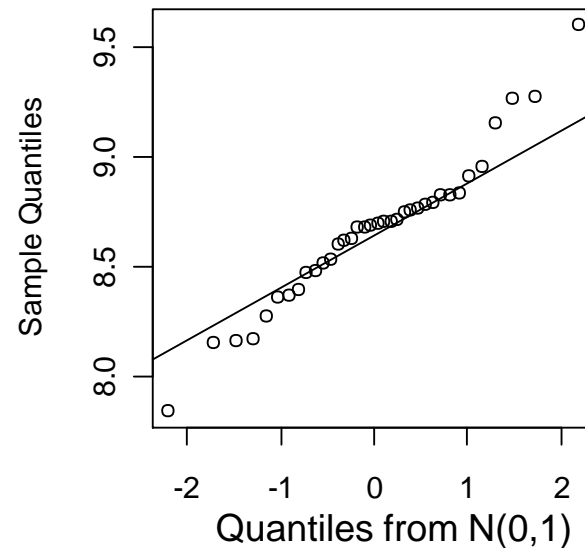
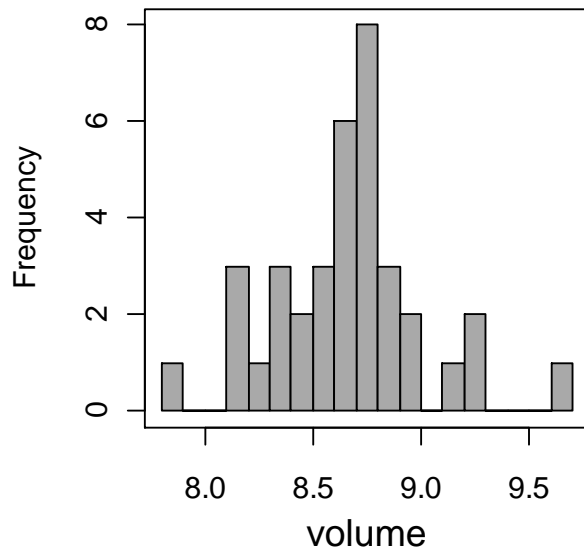
With a confidence interval we can decide:
 Is there a significant difference to a postulated value θ_0 ?
 Is the difference relevant ($>\Delta$)?

Draw CIs that correspond to the description on the right



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

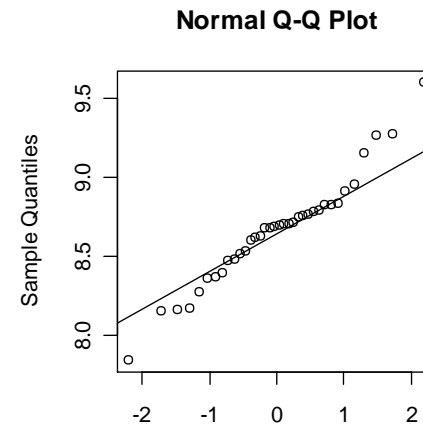
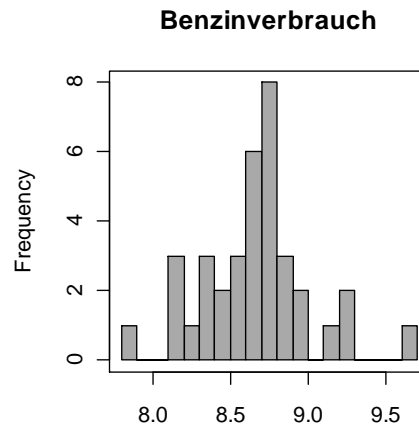
Postulated expected value μ of 8.2 should be tested.

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 $H_A: \mu \neq \mu_0$

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

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

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

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

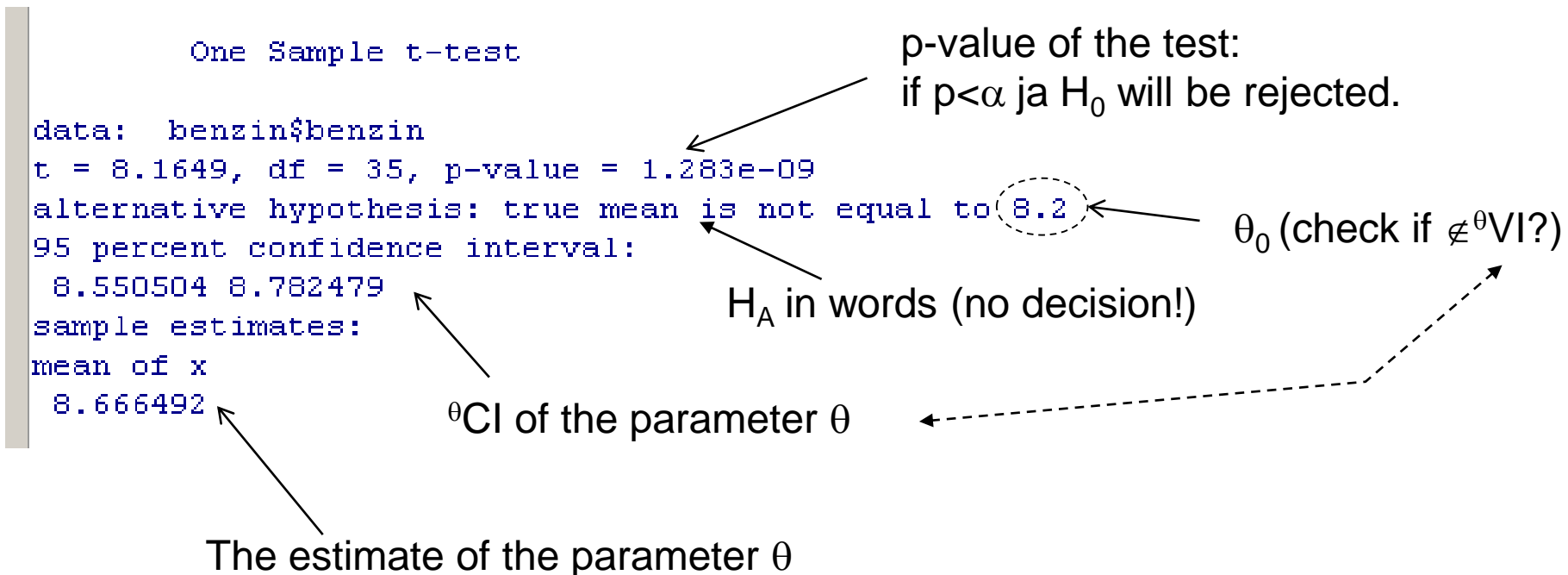
$$m_0 \notin {}^m VI^{1-a} \Rightarrow H_0 \text{ rejected}$$

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

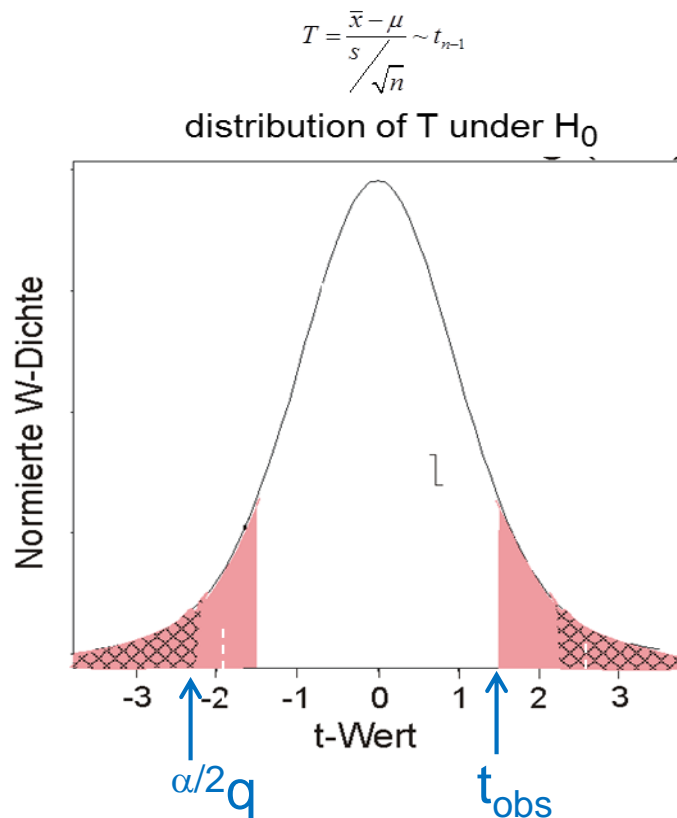
```
> barley
```

	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

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

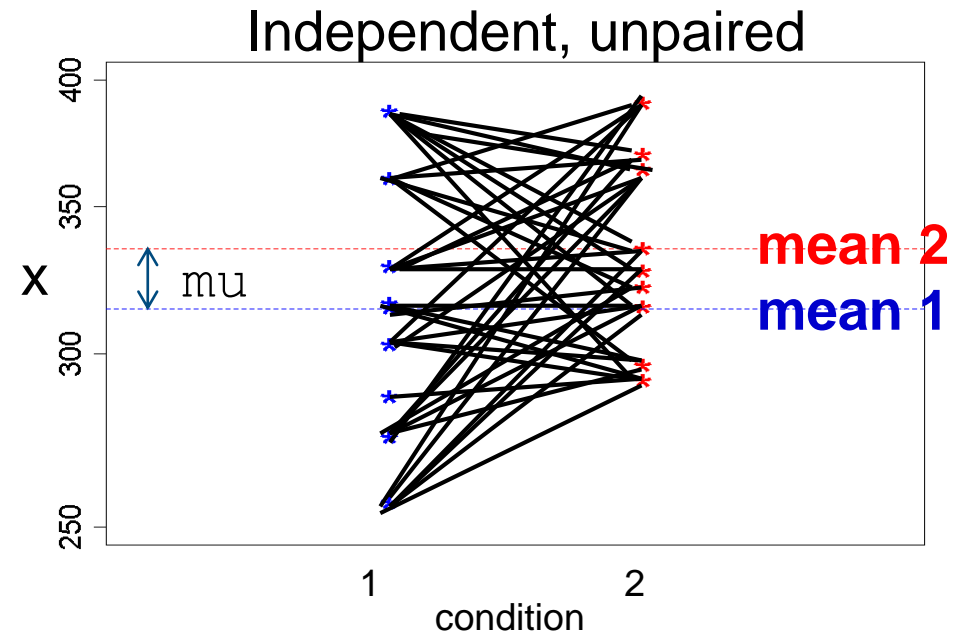
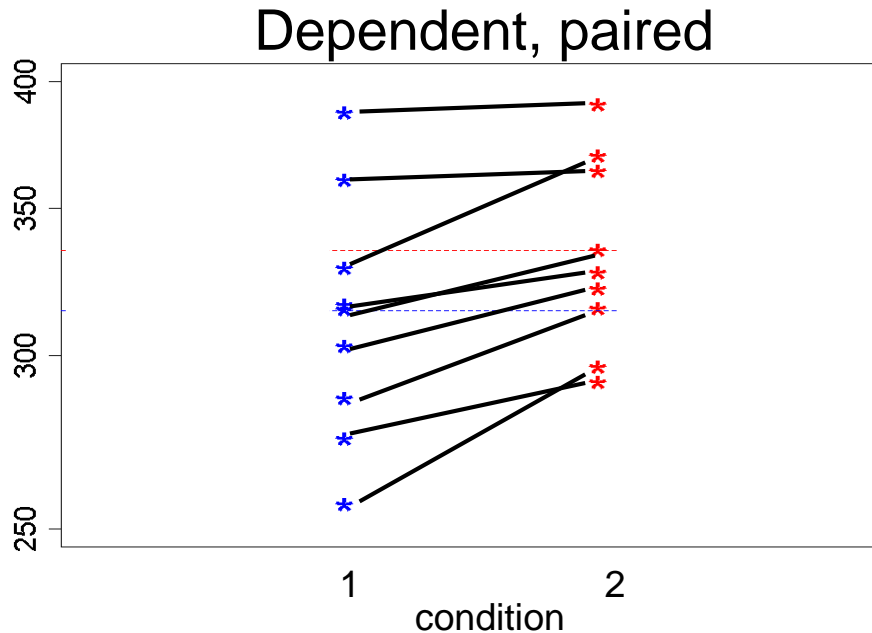
p-value < 0.1 : weak evidence for H_A

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

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

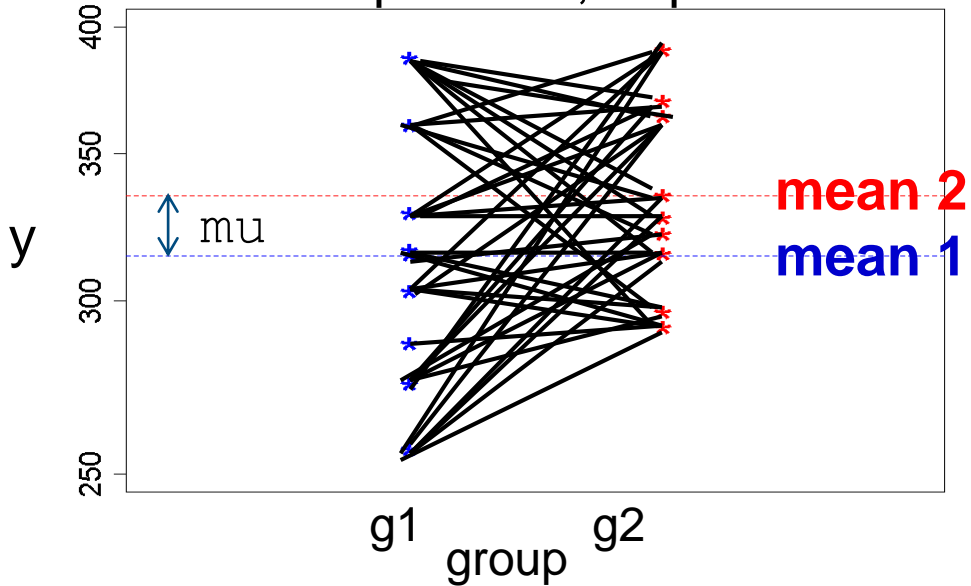


In a **paired design** always a pair of values from group1 and group2 correspond to each other (often 2 treatments were applied to each unit or person) $\sim n_1=n_2$
In a paired design we test if the population **differences within pairs** are zero ($\mu=0$).

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.

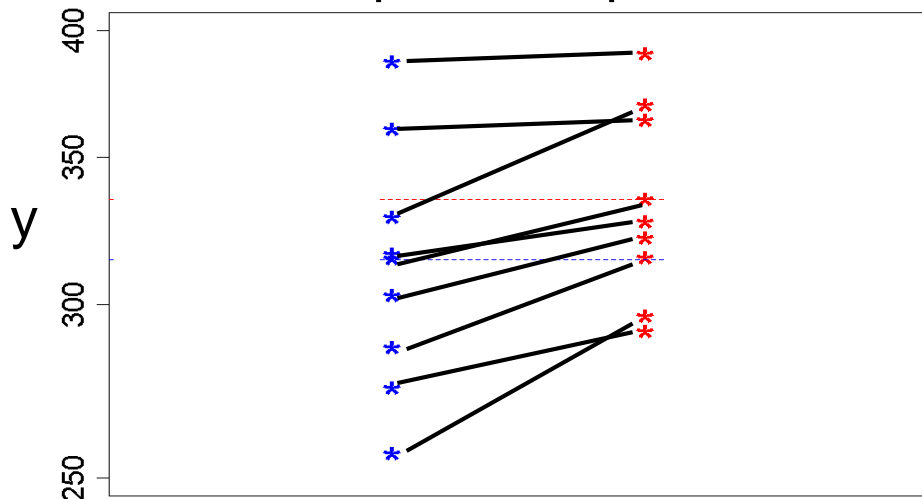
Unpaired and paired data with continuous outcome

Independent, unpaired



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

Dependent, paired

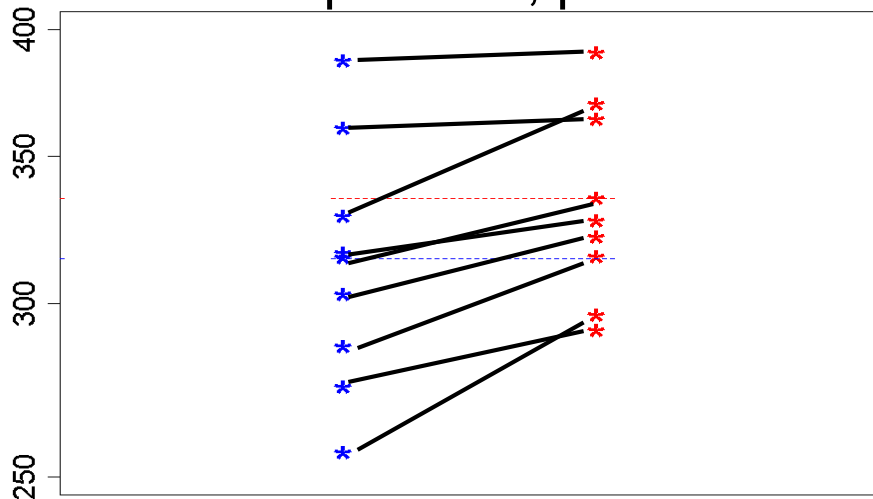


```
t.test(g1,g2, mu=0,  
       var.equal=T, paired=TRUE)
```

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

Pros and Cons of a paired compared to unpaired design

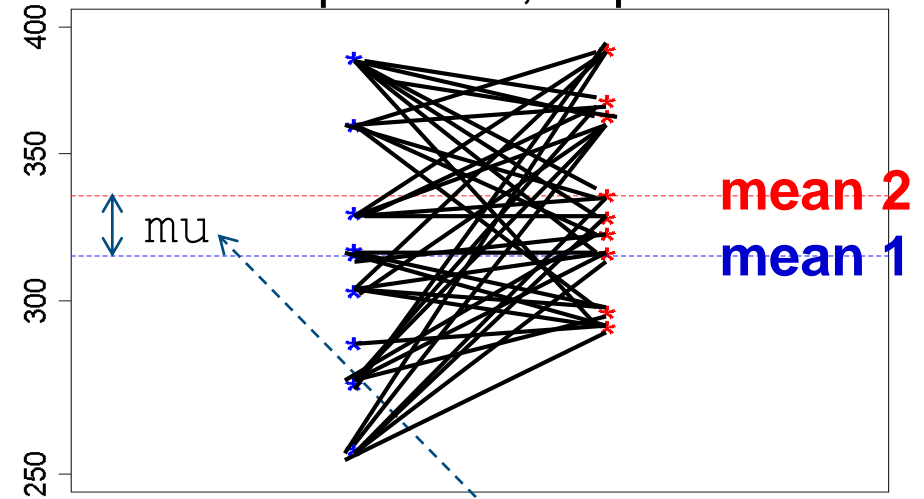
Dependent, paired



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

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

Independent, unpaired



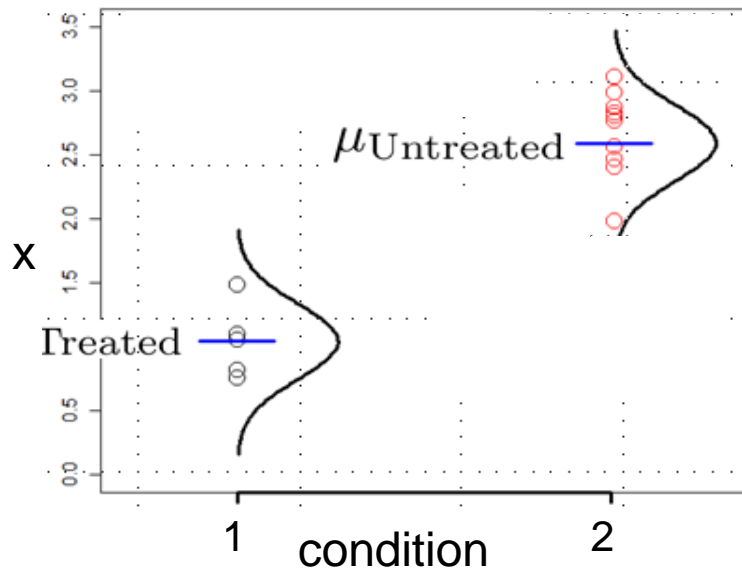
`t.test(g1, g2, mu=0, paired=FALSE)`

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

- + 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.

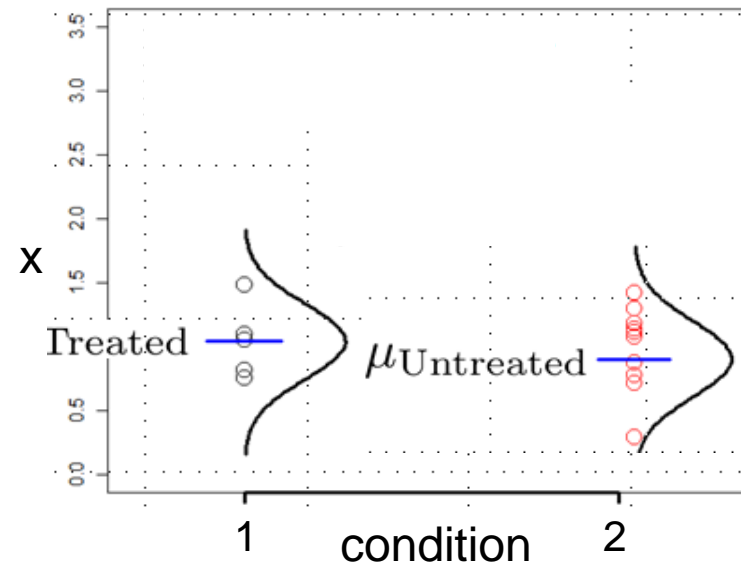
Test for differences between treatments, unpaired design

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

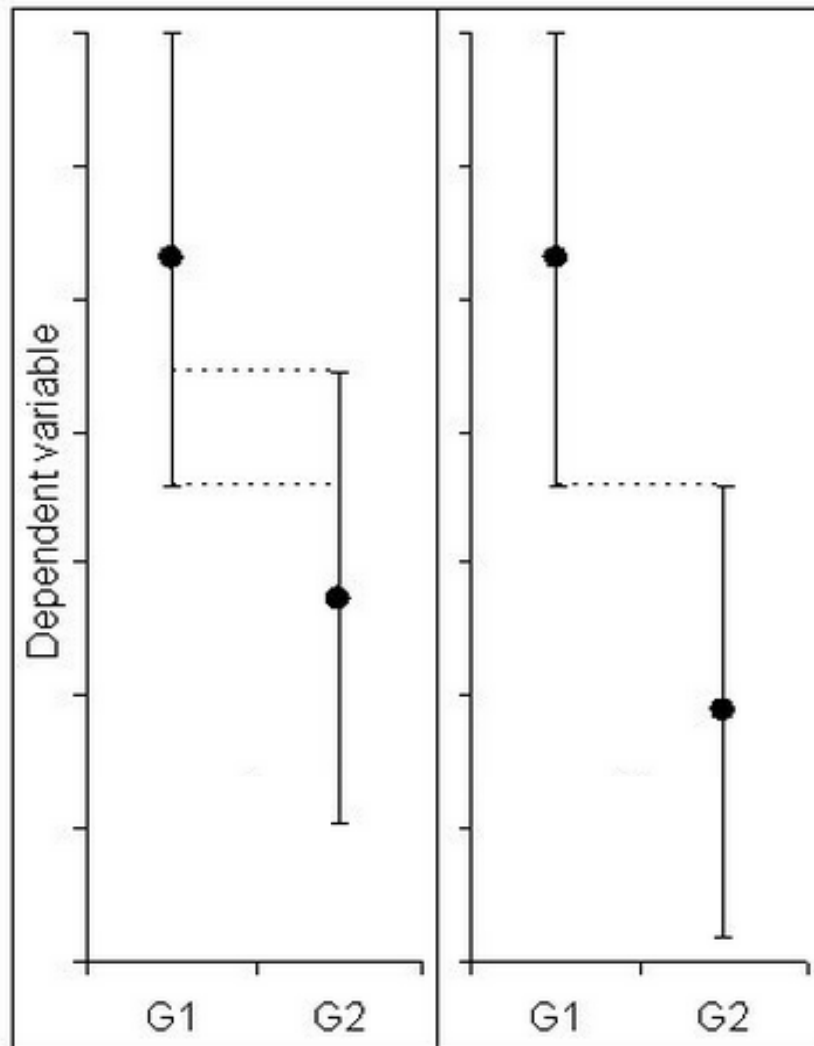


Figure 2: The CIs on the left overlap by about 1/4, half the average margin of error, which corresponds to a p value of $\approx .05$. The CIs on the right are just touching. This corresponds to a p value of $\approx .01$ (Cumming and Finch, 2005).

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

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 (σ 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 (σ is known)
for unpaired
sample

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