HelmholtzZentrum münchen
German Research Center for Environmental Health

# Statistical testing

PD Dr. Andreas Beyerlein

Core Facility Statistical Consulting, Institute of Computational Biology,
Helmholtz Zentrum München

andreas.beyerlein@helmholtz-muenchen.de

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

# Outline

1. Theory of hypothesis tests
2. One-sample t-test
3. Two-sample t-test
4. Chi-square test
5. More about tests
6. Multiple testing

# What is your hypothesis?

# Hypothesis testing

**Example:** Consider two groups of study participants who often experience headache. One group (n=17) gets a drug against headache, the other one (n=17) a placebo. We are mainly interested in the effect of the drug on the headache, but we also measured the systolic blood pressure of every proband as a safety measure.

For the final report, we need to address the following questions:

1. *Before the treatment started, did the participants have a blood pressure that was on average normal? (E.g.* $\mu = 120\, mmHg$*, comparable to a reference population)*
2. *Is the blood pressure affected by the treatment?*
3. *Does the treatment reduce the occurrence of headache?*

# Hypothesis testing

Ad 1.) You have estimated the average blood pressure of the participants before treatment as $\hat{\mu} = 121.7\, mmHg$.

→ Is $121.7\, mmHg$ sufficiently close to $120\, mmHg$?

Ad 2.) Under treatment, the average blood pressures of the control and treatment group were $121\, mmHg$ and $124\, mmHg$, respectively.

→ Does the drug have an effect on the blood pressure?

Ad 3.) 4 out of 17 patients from the placebo group reported less occurrence of headache than before, compared to 12 out of 17 participants in the treatment group.

→ Does the drug have an effect on the headache?

# Hypothesis testing

"I've narrowed it to two hypotheses: it grew or we shrunk."

https://i.pinimg.com/originals/6b/ef/f8/6beff8ae8c9d5ab356efe22d113aaa7e.jpg

# Hypothesis testing

**Idea**: Test whether a specific hypothesis is true or not

**Problem**: Usually only data sample available → Uncertainty (truth is unknown)

**Usually:** Formulate $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$ such that the hypothesis you are actually interested in is $\boldsymbol{H_1}$.

$\boldsymbol{H_0}$ (Null hypothesis) *vs.* $\boldsymbol{H_1}$ (Alternative hypothesis)

→ **Decision:** Reject or not reject $\boldsymbol{H_0}$?
$\boldsymbol{H_0}$ can be rejected or not rejected but not proven to be true!

**Example 3:** Does the drug have an effect on the headache?

$\boldsymbol{H_0}$**:** The drug has no effect on the headache

vs. $\boldsymbol{H_1}$**:** The drug has an effect on the headache

# Hypothesis testing: Procedure

- **Define hypothesis**: One- or two-sided? What should be controlled?
- Choose an appropriate test and the **significance level** $\alpha$
- Check required **assumptions** for this test
- Calculate **test statistic**
- **Test decision** can be based on:
  - Critical region
  - P-value
  - Confidence interval

# Hypothesis testing: Decisions

Always two hypotheses: $\boldsymbol{H_0}$ (Null hypothesis) *vs.* $\boldsymbol{H_1}$ (Alternative hypothesis)

| | | "Truth" | |
|---|---|---|---|
| | | **$H_0$ is true** | **$H_1$ is true** |
| Test decision | **$H_0$ not rejected** | Correct decision | $\beta$ or type II error (false negative) |
| | **$H_0$ rejected** | $\alpha$ or type I error (false positive) | Correct decision |

**Type I error:** The error that we want to control
$\alpha$ often set to 0.05 (5%)

**Type II error:** This error can't be controlled
$\beta$ results of assumed distribution, sample size, $\boldsymbol{H_0}$ and $\alpha$.
Related to statistical power $(1 - \beta)$

# Hypothesis testing: Decisions

Always two hypotheses: $\boldsymbol{H_0}$ (Null hypothesis) *vs.* $\boldsymbol{H_1}$ (Alternative hypothesis)

| | | "Truth" | |
|---|---|---|---|
| | | **$\boldsymbol{H_0}$ is true** | **$\boldsymbol{H_1}$ is true** |
| Test decision | **$\boldsymbol{H_0}$ not rejected** | Specificity<br>$(1-\alpha)$ | Type II error<br>$(\beta)$ |
| | **$\boldsymbol{H_0}$ rejected** | Type I error<br>$(\alpha)$ | Sensitivity / Power<br>$(1-\beta)$ |

**Sensitivity:** Test's ability to correctly reject $\boldsymbol{H_0}$.
How likely does the test determine a true effect?
Sensitivity $=$ Power $= 1-\beta$

**Specificity:** Test's ability to correctly not reject $\boldsymbol{H_0}$.
How likely does the test not reject a truly not existent effect?
Specificity $= 1-\alpha$

# Hypothesis testing: Decisions

$H_0$: Drug has no effect *vs.* $H_1$: Drug has an effect

| | | "Truth" | |
|---|---|---|---|
| | | **$H_0$ is true** | **$H_1$ is true** |
| Test decision | **$H_0$ not rejected** | Drug has no effect;<br>Drug is not released | Drug has an effect;<br>Drug is not released |
| | **$H_0$ rejected** | Drug has no effect;<br>Drug is released | Drug has an effect;<br>Drug is released |

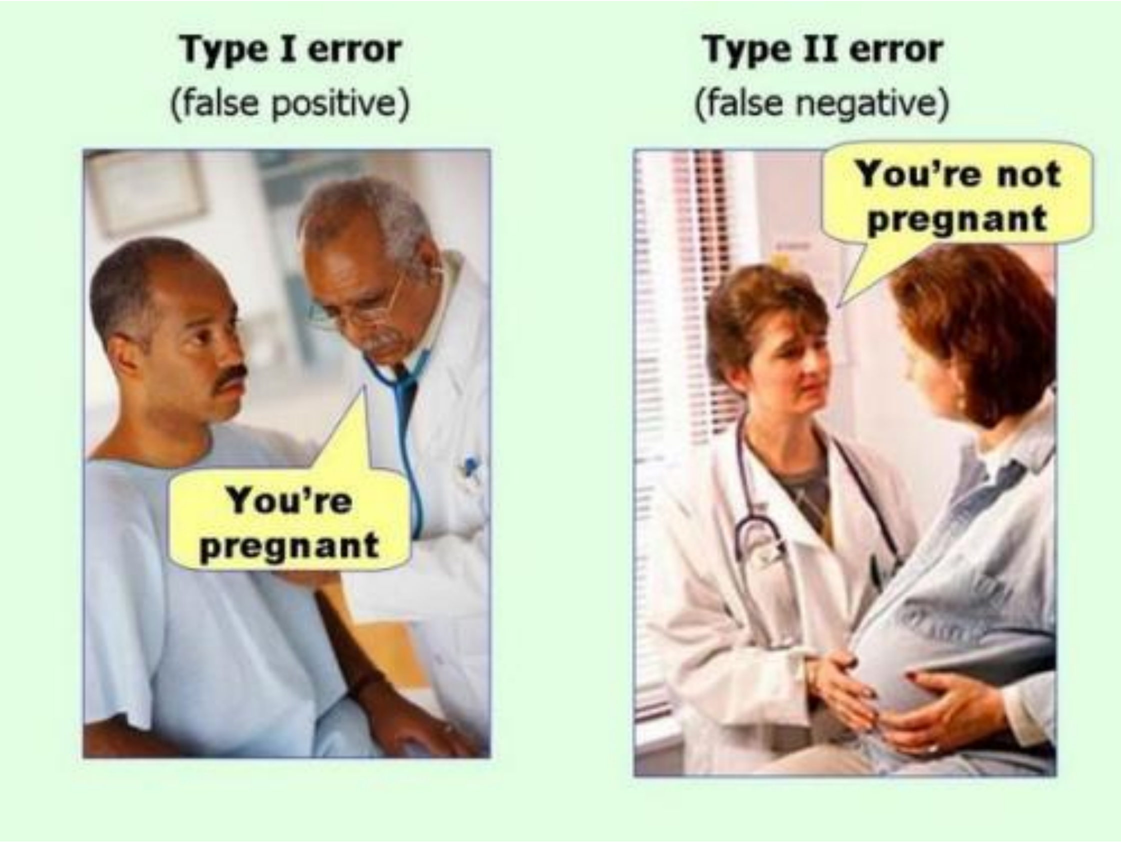
Truth

Decision

**Type I error:** In truth the drug has no effect, but it is still released to the market
→ Side effects, costs, …

**Type II error:** In truth the drug has an effect, but it is not released
→ No treatment available, costs to develop the drug are lost,…

# Hypothesis testing: Decisions



*Paul Ellis: http://effectsizefaq.com*

# Hypothesis testing: Decisions

$\boldsymbol{H_0}$: Drug has no effect *vs.* $\boldsymbol{H_1}$: Drug has an effect

"Truth"

| Test decision | $\boldsymbol{H_0}$ **is true** | $\boldsymbol{H_1}$ **is true** |
|---|---|---|
| $\boldsymbol{H_0}$ **not rejected** | Drug has no effect;<br>Drug is not released | Drug has an effect;<br>Drug is not released |
| $\boldsymbol{H_0}$ **rejected** | Drug has no effect;<br>Drug is released | Drug has an effect;<br>Drug is released |

→ $\alpha$ is pre-defined, $\beta$ is not. The hypothesis you are interested in should therefore be formulated in $\boldsymbol{H_1}$. $\beta$ is usually controlled by the sample size.

→ Formulate hypothesis in the way that the Type I error ($\alpha$) is the error that we want to control for

$\boldsymbol{H_0}$ can be rejected or not rejected - but not proven to be true!

# Hypothesis testing: Decisions

- **p-value**: The probability, under the null hypothesis, of sampling a test statistic at least as extreme as the one which was observed
- Reject $\boldsymbol{H_0}$ if p-value $< \alpha$ (significance level)

- The **confidence interval** determines the area, such that it will include the overall true value in $(1-\alpha)$ out of 100 replications.
- In other words, the confidence interval represents the range of values for the parameter of interest which do not yield a statistically significant difference at the $\alpha$ level.
- The conclusion from a $(1-\alpha)$ confidence interval is equivalent to the conclusion from a test at the $\alpha$ level.

# One sample t-test

# One sample t-test

**Assumptions:** - normal distribution ($X_1, \dots, X_n \sim N(\mu, \sigma^2)$)
- unknown variance $\sigma^2$

$\mu$ = expected value, to be estimated

$\mu_0$ = theoretic value, to be known/assumed

**Hypothesis:** $\boldsymbol{H_0}: \mu = \mu_0$ vs. $\boldsymbol{H_1}: \mu \neq \mu_0$ (two-sided)

**Test statistic:** $T = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}} \sqrt{n} \sim t(n-1)$

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

**Test decision:**

- *Critical region:* Is the test statistic larger/smaller than the quantile of the t distribution?
- *p-value:* Is the p-value smaller than $\alpha$?
- *Confidence interval:* Does the confidence interval contain the theoretic value?

# One sample t-test

$\boldsymbol{H_0}: \mu = \mu_0$ vs. $\boldsymbol{H_1}: \mu \neq \mu_0$

***Reject $H_0$ if:***

*Critical region:* $|T| > \boldsymbol{t_{1-\alpha/2}(n-1)}$

*p-value:* $p = P(T^{new} > |T| \,|\boldsymbol{H_0}\ is\ true) + P(T^{new} < -|T| \,|\boldsymbol{H_0}\ is\ true) < \boldsymbol{\alpha}$

*Confidence interval:* $\boldsymbol{\mu_0} \notin \left[\hat{\mu} - t_{1-\alpha/2}(n-1) \cdot \frac{\hat{\sigma}}{\sqrt{n}} \quad ; \quad \hat{\mu} + t_{1-\alpha/2}(n-1) \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right]$

The test decision will be the same for all three methods.

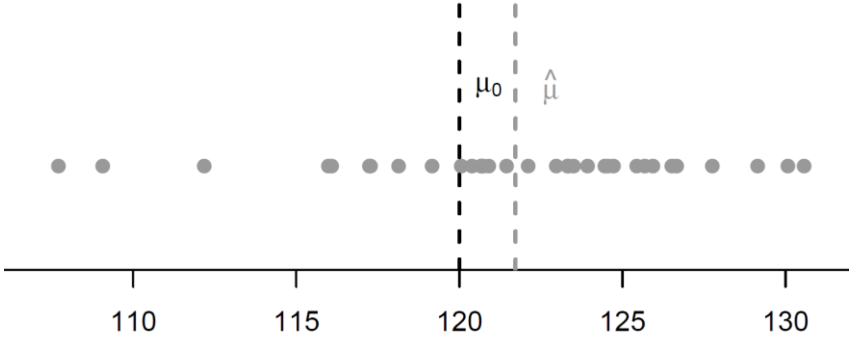
The cut point always depends on $\alpha$

$$\begin{aligned}
\alpha = 0.10&: \quad |T| > t_{0.95}\ (n-1) \ \widehat{=}\ p < 0.10 \ \widehat{=}\ \mu_0 \notin 90\%\ CI \\
\alpha = 0.05&: \quad |T| > t_{0.975}(n-1) \ \widehat{=}\ p < 0.05 \ \widehat{=}\ \mu_0 \notin 95\%\ CI \\
\alpha = 0.01&: \quad |T| > t_{0.995}(n-1) \ \widehat{=}\ p < 0.01 \ \widehat{=}\ \mu_0 \notin 99\%\ CI
\end{aligned}$$

# One sample t-test

**Example:** Participants are expected to have a blood pressure of $\mu_0 = 120\, mmHg$ before treatment. There are 34 measurements:

We choose $\alpha = 0.05$
(also for all other examples)



***Hypothesis:***

$\boldsymbol{H_0}: \mu = 120$ vs. $\boldsymbol{H_1}: \mu \neq 120$

$\boldsymbol{H_0}$ states that the blood pressures don't differ from the theoretic value

***Test statistic:***

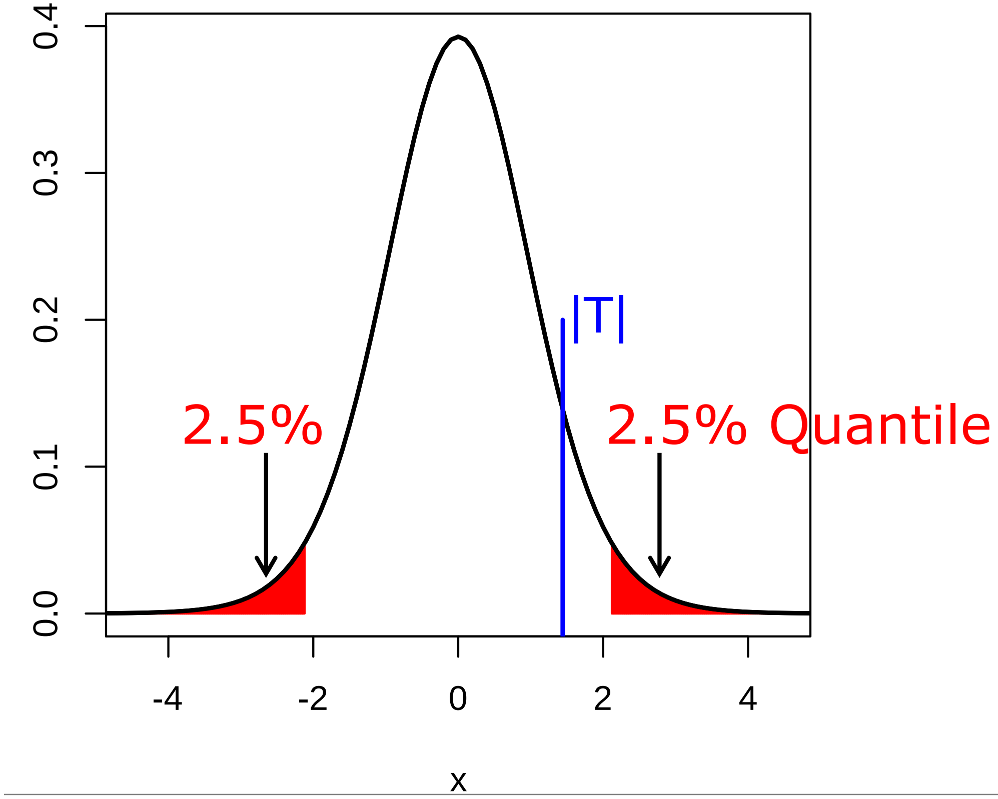
$$\hat{\mu} = \bar{X} = 121.7 \;\; ; \;\; \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 28.9$$

$$T = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}}\sqrt{n} = \frac{121.7 - 120}{\sqrt{28.9}}\sqrt{34} = 1.862$$

# One sample t-test: Critical Region

Reject $\boldsymbol{H_0}$ if test statistic $T$ is more extreme than the $(1-{}^{\alpha}/_{2})$-quantile of the t-distribution (shaded area)

Example:
$|T| = 1.862 < 2.035 = t_{0.975}(33) = t_{1-\alpha/2}(n-1)$



$\alpha = 0.05$
$df = n-1$

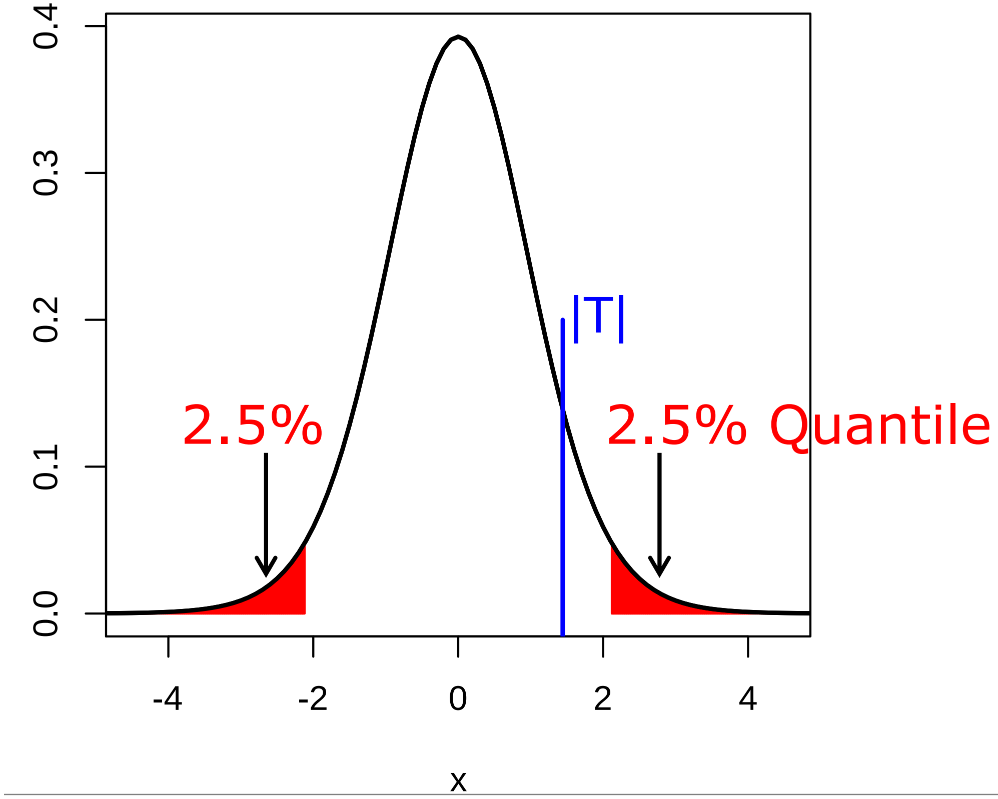
| $1-\frac{\alpha}{2}$ / df | **0.90** | **0.95** | **0.975** | **0.99** |
|---|---|---|---|---|
| 32 | 1.309 | 1.694 | 2.037 | 2.449 |
| 33 | 1.308 | 1.692 | **2.035** | 2.445 |
| 34 | 1.307 | 1.691 | 2.032 | 2.441 |

# One sample t-test: Critical Region

Reject $\boldsymbol{H_0}$ if test statistic $T$ is more extreme than the $(1-{}^{\alpha}/_{2})$-quantile of the t-distribution (shaded area)

Example:

$|T| = 1.862 < 2.035 = t_{0.975}(33) = \ t_{1-\alpha/2}(n-1)$ → $\boldsymbol{H_0}$ **is not rejected**



$\alpha \ = 0.05$

$df \ = n-1$

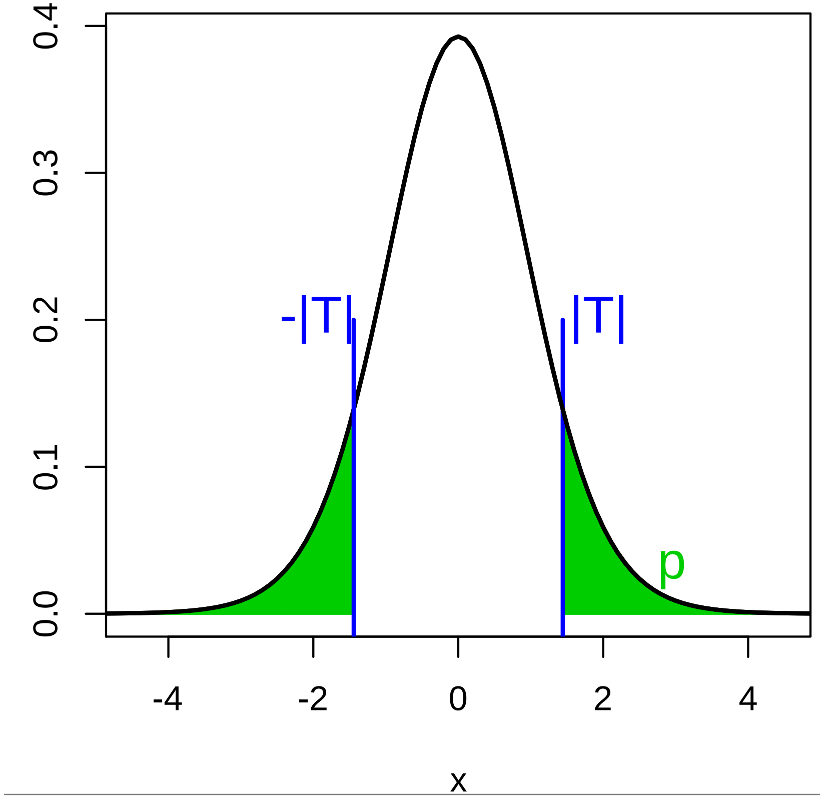
| $1-\frac{\alpha}{2}$ / df | 0.90 | 0.95 | 0.975 | 0.99 |
|---|---|---|---|---|
| 32 | 1.309 | 1.694 | 2.037 | 2.449 |
| 33 | 1.308 | 1.692 | **2.035** | 2.445 |
| 34 | 1.307 | 1.691 | 2.032 | 2.441 |

# One sample t-test: p-value

Reject $\boldsymbol{H_0}$ if p-value (shaded area) is smaller than $\alpha$

→ If we assume that $\boldsymbol{H_0}$ is true, the p-value is the probability of sampling a test statistic at least as extreme as the one which was observed (T = 1.862).

Example: $p = 0.071 > 0.05$ → $\boldsymbol{H_0}$ **is not rejected**



$$p = P(T^{new} > \ \ 1.862|\ blood\ pressure\ is\ normal) + P(T^{new} < -1.862|\ blood\ pressure\ is\ normal)$$

Good statistical practice is to first define $\alpha$ before comparing with $p$!

# One sample t-test: Confidence interval

Reject $\boldsymbol{H_0}$ if the confidence interval does not contain the theoretic value

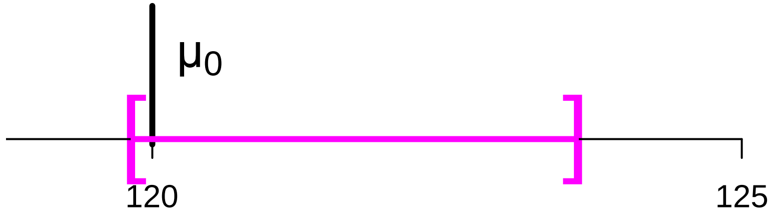
→ The confidence interval determines the area, such that in 100 replications it will overlie the overall true value in $(1-\alpha)$ replications.

Example:

$120 \in [119.8\,;\,123.6]$ → $\boldsymbol{H_0}$ **is not rejected**

$$\left[\hat{\mu} - t_{1-\alpha/2}(n-1)\cdot\frac{\hat{\sigma}}{\sqrt{n}} \quad ; \quad \hat{\mu} + t_{1-\alpha/2}(n-1)\cdot\frac{\hat{\sigma}}{\sqrt{n}}\right] =$$

$$[121.7 - 2.035\cdot\frac{5.376}{\sqrt{34}} \quad ; 121.7 + 2.035\cdot\frac{5.376}{\sqrt{34}}] = [119.8\,;\,123.6]$$



Based on the test decision we do not need to repeat the experiment.

# Two sample t-test

# Two sample t-test

## Idea:

- Two samples **x** and **y**, both normally distributed

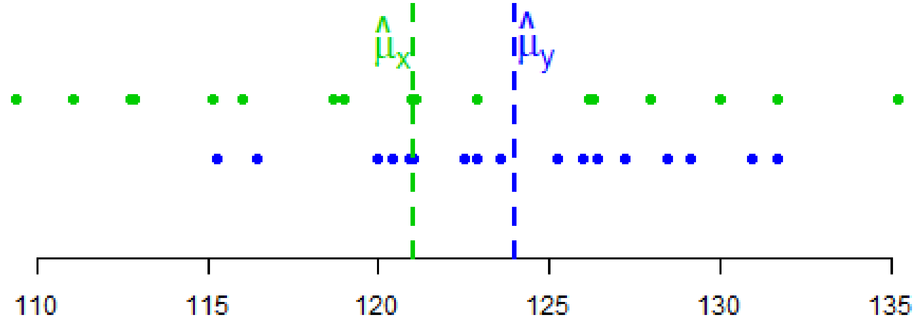
  $x \sim N(\mu_x, \sigma_x^2)$ and $y \sim N\left(\mu_y, \sigma_y^2\right)$ (sample sizes $n_x$ and $n_y$)

- Test for difference in mean(x) = $\mu_x$ and mean(y) = $\mu_y$:
  e.g. $\boldsymbol{H_0}$**:** $\mu_y - \mu_x = 0 \quad \widehat{=} \quad \mu_x = \mu_y$ *vs.* $\boldsymbol{H_1}$**:** $\mu_x \neq \mu_y$

## Example 2:

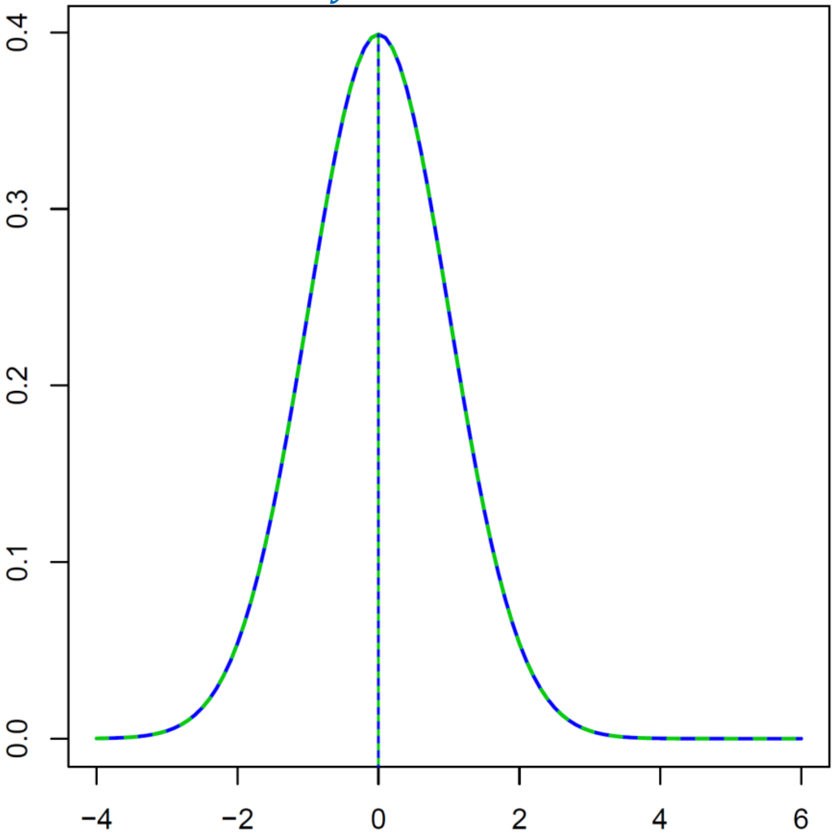
- Under treatment, the treatment group has an average blood pressure of $\hat{\mu}_y = 124$ and the control group of $\hat{\mu}_x = 121$.

- Is this difference significant?

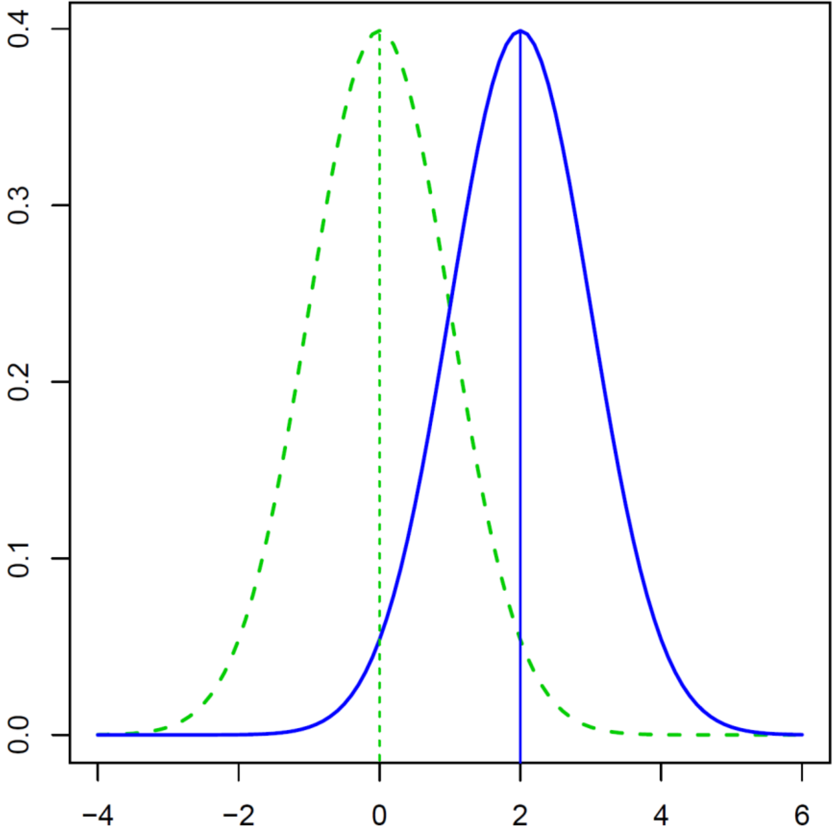


# Two sample t-test

$\boldsymbol{H_0}$: $\mu_y - \mu_x = 0$



$\boldsymbol{H_1}$: $\mu_y - \mu_x \neq 0$



# Two sample t-test: independent samples

## Test statistic:

Both $\sigma_x^2$ and $\sigma_y^2$ known
(***Gauß Test***)

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

$\sigma_x^2$ and $\sigma_y^2$ unknown, but assumed to be the same
(***t Test***)

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\frac{(n_x - 1)\hat{\sigma}_x^2 + \left(n_y - 1\right)\hat{\sigma}_y^2}{n_x + n_y - 2}}}$$

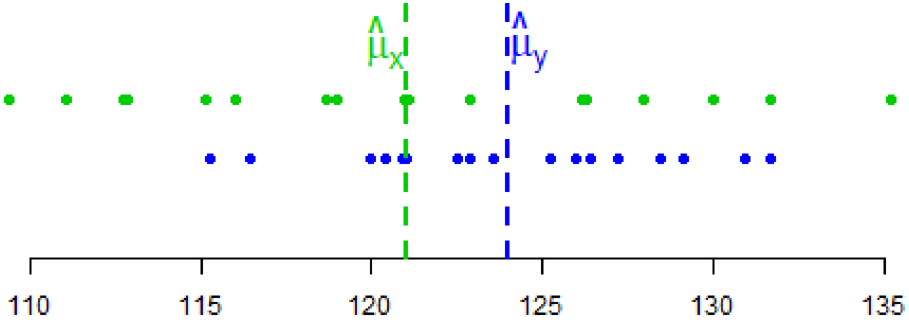
$\sigma_x^2$ and $\sigma_y^2$ unknown, and **not** assumed to be the same
(***Welch Test***)

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}}$$

(***Welch Test***)

# Two sample t-test

## Example:



- The treatment group has an average blood pressure of $\hat{\mu}_y = 124$ and the control group of $\hat{\mu}_x = 121$.

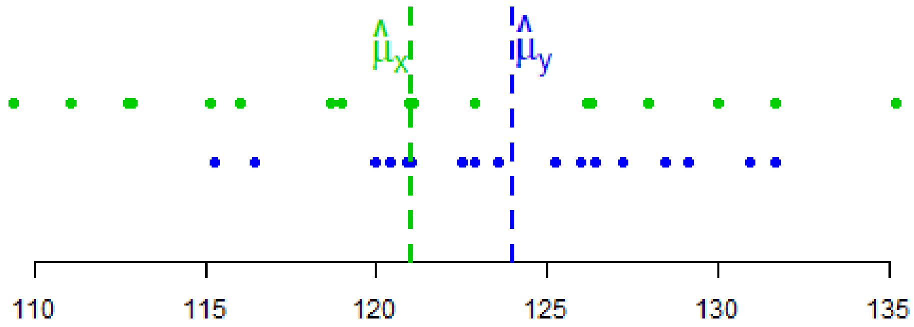
- Check whether the difference is significant, assuming normality and equal, but unknown variances. → t-test

$\boldsymbol{H_0}$: $\mu_x = \mu_y$ vs. $\boldsymbol{H_1}$: $\mu_x \neq \mu_y$ $\qquad T \sim t(n_x + n_y - 2)$

```
> t.test(x, y, var.equal = TRUE)
        Two Sample t-test

data: x and y
t = 1.3716, df = 32, p-value = 0.1797
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.452996 7.443724
sample estimates:
mean of x mean of y
 124.0010  121.0056
```

# Two sample t-test

## Example:



- The treatment group has an average blood pressure of $\hat{\mu}_y = 124$ and the control group of $\hat{\mu}_x = 121$.

- Check whether the difference is significant, assuming normality and equal, but unknown variances. → t-test

$\boldsymbol{H_0}$: $\mu_x = \mu_y$ vs. $\boldsymbol{H_1}$: $\mu_x \neq \mu_y$ $\qquad T \sim t(n_x + n_y - 2)$

```
> t.test(x, y, var.equal = TRUE)
        Two Sample t-test

data: x and y
t = 1.3716, df = 32, p-value = 0.1797
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.452996 7.443724
sample estimates:
mean of x mean of y
 124.0010  121.0056
```

→ $\boldsymbol{H_0}$ **is not rejected:** We found no evidence that the drug affects blood pressure.

# Step-by-step to a two-sample t-test

1. Check for obviously non-normally distributed data using plots
   Nothing visible: continue with

2. Test for normal distribution in each group
   (e.g. *Shapiro–Wilk*, *Kolmogorov–Smirnov*)
   1 rejected hypothesis: *non-parametric test*
   no rejected hypothesis: continue with

3. Test for homogeneous variances (e.g. *Bartlett test, Levene test*)
   hypothesis is not rejected: *two-sample t-test* (more power)
   hypothesis is rejected: *Welch test*

# Chi-square test

# Chi-square test for independence

**Example 3:** Is headache reduced by drug vs. placebo?

4 out of 17 patients from the placebo group reported less occurrence of headache than before, compared to 12 out of 17 participants in the treatment group.

Observed Frequencies:

| | | **X: Treatment** | | |
|---|---|---|---|---|
| | | Placebo | Drug | |
| **Y: Headache** | Improved | 4 | 12 | 16 |
| | Not improved | 13 | 5 | 18 |
| | | 17 | 17 | 34 |

# Chi-square test for independence

**Example 3:** Is headache reduced by drug vs. placebo?

$\boldsymbol{H_0}$*:* No difference in headache between drug and placebo

Observed Frequencies:

| | | **X: Treatment** | | |
|---|---|---|---|---|
| | | Placebo | Drug | |
| **Y: Headache** | Improved | 4 | 12 | 16 |
| | Not improved | 13 | 5 | 18 |
| | | 17 | 17 | 34 |

Expected Frequencies:

| | | **X: Treatment** | | |
|---|---|---|---|---|
| | | Placebo | Drug | |
| **Y: Headache** | Improved | 8 | 8 | 16 |
| | Not improved | 9 | 9 | 18 |
| | | 17 | 17 | 34 |

*Expected frequencies* given that the observations at the margins are fixed and there is **no** association between **X** and **Y**

# Chi-square test for independence

**Formally:** Extension to three or more categories per variable

| | | X | | | |
|---|---|---|---|---|---|
| | | Category 1 | Category 2 | Category 3 | |
| Y | Category 1 | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{1\cdot}$ |
| | Category 2 | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{\mathbf{2}\cdot}$ |
| | Category 3 | $n_{31}$ | $n_{32}$ | $n_{33}$ | $n_{3\cdot}$ |
| | | $n_{\cdot 1}$<br>$= (\mathrm{n11} + \mathrm{n21} + \mathrm{n31})$ | $n_{\cdot 2}$<br>$= (\mathrm{n12} + \mathrm{n22} + \mathrm{n32})$ | $n_{\cdot 3}$<br>$= (\mathrm{n13} + \mathrm{n23} + \mathrm{n33})$ | $n$ |

I

J

- Number of observations per cell: $n_{ij}$
- Calculate the *expected* number of observations per cell: $e_{ij} = \frac{n_{i\cdot} * n_{\cdot j}}{n}$
- Test whether the two variables **X** and **Y** are independent, i.e. the frequencies spread similarly in the table

# Chi-square test for independence

**Hypothesis:** $\boldsymbol{H_0}$**:** $n_{11} = e_{11},\ n_{12} = e_{12},\ \dots$ vs. $\boldsymbol{H_1}$**:** $n_{ij} \neq e_{ij}$ in one cell

**Test statistic:** Squared differences between expected $e_{ij}$ and observed frequencies $n_{ij}$

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}$$

**Test decision:**

$$X^2 \sim \chi^2(df)$$

Reject, if: $X^2 > \chi^2_{1-\alpha}(df)$

Degrees of freedom: $df = (I-1)(J-1)$

**Assumptions:** - Independent couples $(X_k, Y_k)$ grouped in a table
- Adequate cell counts expected:
5 or more observations expected in all cells

# Chi-square test for independence

**Example 3:** Is headache reduced by drug vs. placebo?

$\boldsymbol{H_0}$: No difference in headache between drug and placebo

| | | **X: Treatment** | | |
|---|---|---|---|---|
| | | Placebo | Drug | |
| **Y: Headache** | Improved | 4 | 12 | 16 |
| | Not improved | 13 | 5 | 18 |
| | | 17 | 17 | 34 |

```
> M1 <- matrix(c(4, 12, 13, 5), byrow = TRUE, nrow = 2, ncol = 2)
> chisq.test(M1, correct=FALSE)

        Pearson's Chi-squared test

data:  M1
X-squared = 7.5556, df = 1, p-value = 0.005983
```

$$\chi^2_{0.95}(1 * 1) = 3.84$$

# Chi-square test for independence

**Example 3:** Is headache reduced by drug vs. placebo?

$H_0$: No difference in headache between drug and placebo

|  |  | **X: Treatment** |  |  |
|---|---|---|---|---|
|  |  | Placebo | Drug |  |
| **Y: Headache** | Improved | 4 | 12 | 16 |
|  | Not improved | 13 | 5 | 18 |
|  |  | 17 | 17 | 34 |

```
> M1 <- matrix(c(4, 12, 13, 5), byrow = TRUE, nrow = 2, ncol = 2)
> chisq.test(M1, correct=FALSE)

        Pearson's Chi-squared test

data:  M1
X-squared = 7.5556, df = 1, p-value = 0.005983
```

→ $H_0$ **is rejected:**
We conclude that the drug has an effect on headache

# More about tests

# Thoughts about p-values

- *P*-values can indicate how incompatible the data are with the null hypothesis in a specified statistical model.
- *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- A *p*-value, or statistical significance, does not directly measure the size of an effect or the importance of a result.
- Any effect, no matter how tiny, can produce a small *p*-value if the sample size or measurement precision is high enough.

*Wasserstein / Lazar: The ASA's Statement on p-Values*

# One-sided or two-sided hypothesis

- **Important decision** – needs to be decided before the test is performed

- Only relevant for tests on differences, but not on independence (e.g. Chi-Square test)

- One-sided hypothesis implies that you are sure that if there is a difference, it can only be (or is only relevant) to one direction.

- **In almost all cases** (at least in life-sciences), **two-sided tests are performed**. Exception: Some clinical trial designs

- By default, also the p-value is calculated in that way.

# Hypothesis testing

## Standard hypotheses:

- Two-sided: $\boldsymbol{H_0}$: " = " vs. $\boldsymbol{H_1}$: " ≠ "
- One-sided: $\boldsymbol{H_0}$: " ≤ “ vs. $\boldsymbol{H_1}$: " > "

  $\boldsymbol{H_0}$: " ≥ " vs. $\boldsymbol{H_1}$: " < "

Usually: Formulate $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$ such that the hypothesis you are actually interested in is $\boldsymbol{H_1}$.

**Example**: Average blood pressure of participants before the experiment:

- Two-sided: $\boldsymbol{H_0}: \mu = 120\, mmHg$ vs. $\boldsymbol{H_1}: \mu \neq 120\, mmHg$
- One-sided: $\boldsymbol{H_0}: \mu \leq 120\, mmHg$ vs. $\boldsymbol{H_1}: \mu > 120\, mmHg$

  $\boldsymbol{H_0}: \mu \geq 120\, mmHg$ vs. $\boldsymbol{H_1}: \mu < 120\, mmHg$

# One sample t-test: one-sided hypothesis

*(a)* $\boldsymbol{H_0}: \mu = \mu_0$ vs. $\boldsymbol{H_1}: \mu \neq \mu_0$
*(b)* $\boldsymbol{H_0}: \mu \geq \mu_0$ vs. $\boldsymbol{H_1}: \mu < \mu_0$
*(c)* $\boldsymbol{H_0}: \mu \leq \mu_0$ vs. $\boldsymbol{H_1}: \mu > \mu_0$

***Reject $H_0$ if:***

*Critical region:*

(a) $|T| > \boldsymbol{t_{1-\alpha/2}(n-1)}$
(b) $T < \boldsymbol{t_{\alpha}(n-1)}$
(c) $T > \boldsymbol{t_{1-\alpha}(n-1)}$

Remember: $T = \frac{\hat{\mu}-\mu_0}{\hat{\sigma}}\sqrt{n}$

*p-value:*

(a) $p = P(T^{new} > |T| \, |\boldsymbol{H_0} \, is \, true) + P(T^{new} < -|T| \, |\boldsymbol{H_0} \, is \, true) < \boldsymbol{\alpha}$
(b) $p = P(T^{new} < T \, |\boldsymbol{H_0} \, is \, true) < \boldsymbol{\alpha}$
(c) $p = P(T^{new} > T \, |\boldsymbol{H_0} \, is \, true) < \boldsymbol{\alpha}$

*Confidence interval:*

(a) $\boldsymbol{\mu_0} \notin \left[\hat{\mu} - t_{1-\alpha/2}(n-1) \cdot \frac{\hat{\sigma}}{\sqrt{n}} \quad ; \quad \hat{\mu} + t_{1-\alpha/2}(n-1) \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right]$
(b) $\boldsymbol{\mu_0} \notin \left[-\infty \quad ; \quad \hat{\mu} + t_{1-\alpha}(n-1) \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right]$
(c) $\boldsymbol{\mu_0} \notin \left[\hat{\mu} - t_{1-\alpha}(n-1) \cdot \frac{\hat{\sigma}}{\sqrt{n}} \quad ; \quad \infty\right]$

# Independent or dependent samples?

- Remember question 2:
  *Is the blood pressure affected by the treatment?*

- We analyzed this in the following way:
  *Under treatment, the treatment group has an average blood pressure of* $\hat{\mu}_y = 124$ *and the control group of* $\hat{\mu}_x = 121$*.* → *Is this difference significant?*

- Obviously, we did not take the blood pressure measurements before treatment into account here, although we had them for each participant!

- How can we make use of this additional information?

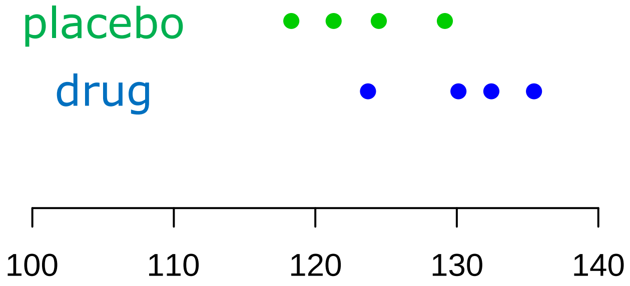
# Independent or dependent samples?

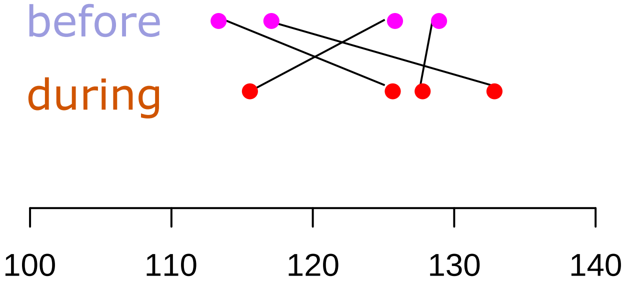
**Independent Sample:**

- Two groups of patients
- One group gets a placebo, the other group a drug.
- Compare blood pressure between both groups

**Dependent Sample:**

- Consider only the drug group
- Repeated measurements of blood pressure
- Compare blood pressure before and during treatment in the <u>same</u> individuals (hence: “dependent”)





# Independent or dependent samples?

**When to use what?**

- **Independent**: Two independent samples, e.g. differences between two groups of individuals
- **Dependent**: Repeated measurements conducted at the same individual, or matched pairs design

**Why is this important?**

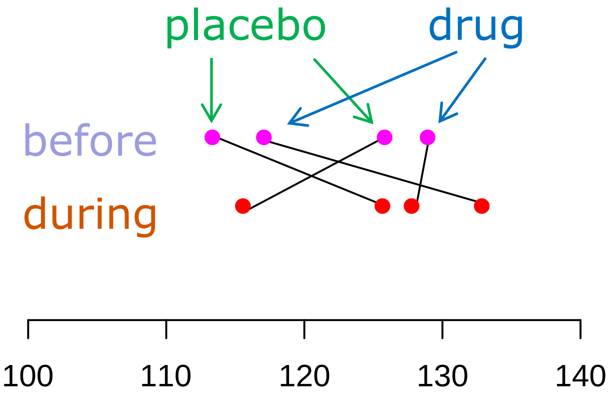
Variance is smaller for dependent samples. This is an advantage that should be considered in the analysis. Paired tests for dependent samples have more power if the assumptions are fulfilled.

→ **Example: Exercises**

# Independent or dependent samples?

**Some notes about t-tests on dependent samples:**

- A t-test on two dependent samples is equivalent to a t-test on whether the mean difference within each individual between the two samples (e.g. before and during treatment) is 0.
- Hence, one can also apply a t-test on dependent samples with two grouping factors (e.g. drug and placebo): It is equivalent to the two-sample t-test on mean differences between the values from the dependent samples.



# Parametric or non-parametric tests?

- The t-test and the Chi-Square test assume that the data follow a certain distribution (e.g. t-test: normal distribution). They are therefore called **parametric tests**.

- However, the distributional assumptions may not always be fulfilled. In that case, it may be necessary to use a test with no specific assumptions about the underlying distribution, a so-called **non-parametric test**.

- For more or less each parametric test, **there is a non-parametric alternative**, e.g. t-test / Wilcoxon test; Chi-square test / Fisher's exact test.

- If the distribution assumptions are valid, the **parametric test has more statistical power** than the non-parametric test (i.e. it is more likely to get a significant result in case there is a true effect).

- The **non-parametric test has less assumptions** and is therefore the safer choice.

# Wilcoxon rank sum test (= Mann–Whitney *U* test)

- **Assumptions**:

  - Ordinal variables, no specific distribution, but interval scaled

  - Two independent variables

- The Wilcoxon rank sum test is often used if the assumptions for a t-test are not fulfilled
- Tests whether two independent samples are based on the same distribution (which does not need to be specified)
- Test statistic is based on ranks

# Wilcoxon rank sum test (= Mann–Whitney *U* test)

**Example**: X = (19, 5, 1, 15, 8) and Y = (18, 6, 3, 10, 2)

- $\boldsymbol{H_0}$: **X** and **Y** are based on the same distribution **vs.**
  $\boldsymbol{H_1}$: **X** and **Y** are based on different distributions
- Ranks of pooled sample:

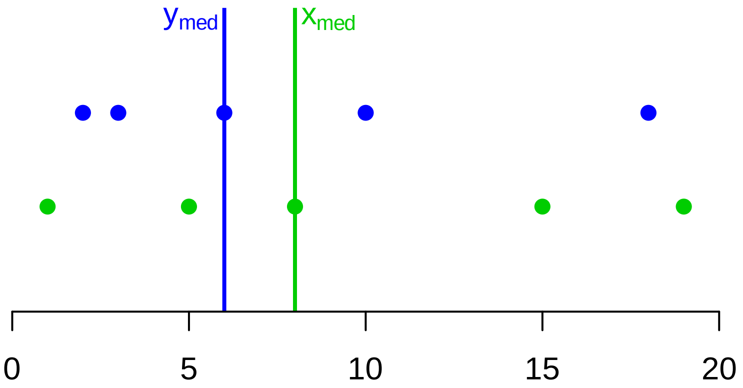
| Value | 1 | 2 | 3 | 5 | 6 | 8 | 10 | 15 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|
| Sample | X | Y | Y | X | Y | X | Y | X | Y | X |
| Ranks | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

- Medians: median(X) = 8; median(Y) = 6
- Sum of ranks of the first group X: $R_x$ = 1+4+6+8+10 = 29

  Test statistic W: $$W = R_x - \frac{n_x(n_x+1)}{2} = 29 - \frac{5(5+1)}{2} = 14$$

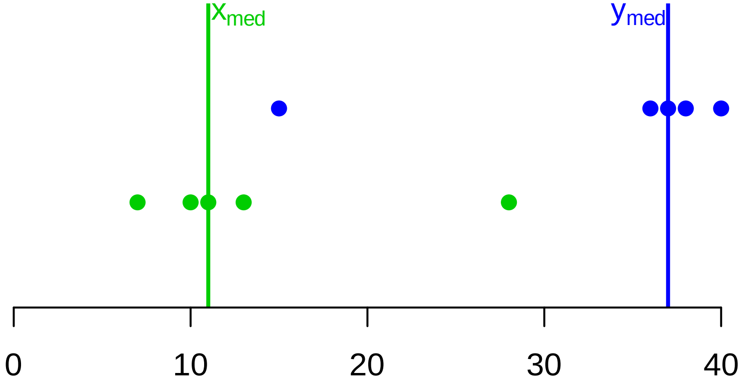
- Reject $\boldsymbol{H_0}$ if: $\mathrm{W} > w_{1-\alpha/2}(n_x, n_y) = 22$ or $\mathrm{W} < w_{\alpha/2}(n_x, n_y) = 3$

# Wilcoxon rank sum test (= Mann–Whitney *U* test)



```
> X <- c(19, 5, 1, 15, 8)
> Y <- c(18, 6, 3, 10, 2)
> wilcox.test(X, Y)

        Wilcoxon rank sum test

data: X and Y
W = 14, p-value = 0.8413
alternative hypothesis: true location shift is not equal to 0

> qwilcox(0.975, 5, 5)
 22
> qwilcox(0.025, 5, 5)
 3
```

# Wilcoxon rank sum test (= Mann–Whitney *U* test)

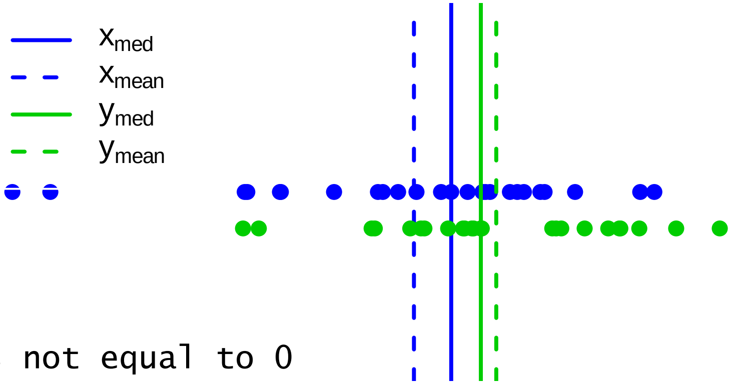


```
> X <- c(10, 28, 13, 11, 7)
> Y <- c(36, 38, 37, 40, 15)
> wilcox.test(X, Y)

        Wilcoxon rank sum test

data: X and Y
W = 1, p-value = 0.01587
alternative hypothesis: true location shift is not equal to 0

> qwilcox(0.975, 5, 5)
 22
> qwilcox(0.025, 5, 5)
 3
```

# Wilcoxon rank sum test (= Mann–Whitney *U* test)

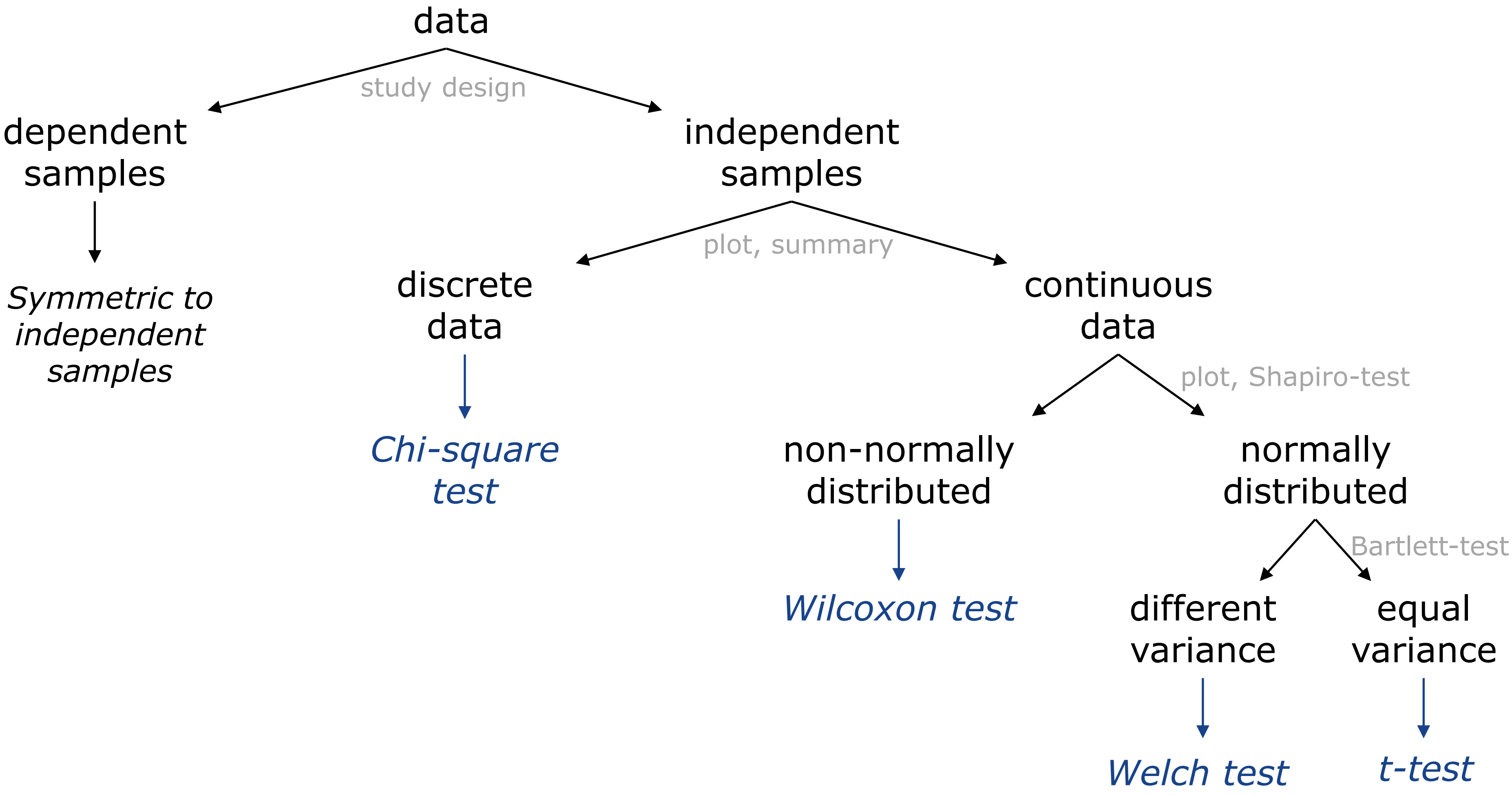
Why still trying to apply the t-test?

- If the assumptions are fulfilled, the t-test has more power, i.e. it is more likely to detect differences.



-
  ```
  > set.seed(1)
  > x <- rnorm(25, mean = 17.5, sd = 4)
  > y <- rnorm(25, mean = 20.0, sd = 4)

  > t.test(x, y, var.equal = TRUE)
          Two Sample t-test
  data: x and y
  t = -2.0634, df = 48, p-value = 0.0445
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
   -3.85852145 -0.05000768
  sample estimates:
  mean of x mean of y
   18.17466 20.12893

  > wilcox.test(x, y)
          Wilcoxon rank sum test
  data: x and y
  W = 234, p-value = 0.131
  alternative hypothesis: true location shift is not equal to 0
  ```

# Decision tree for two-sample tests (simplified)



# More than two samples

- It is also possible to do overall tests on more than two groups.

- The Chi-square test has already been defined for tables with more than 2x2 dimensions.

- Generalization of t-test / Wilcoxon test to more than two groups: ANOVA / Kruskal-Wallis test

  → Hypothesis (ANOVA):

  $\boldsymbol{H_0}$**:** $\mu_1 = \mu_2 = \cdots = \mu_K$

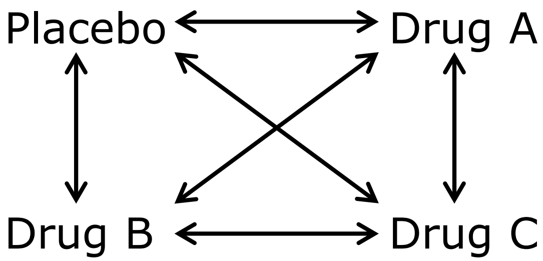
  $\boldsymbol{H_1}$**:** means are different for at least 2 groups

# Multiple Testing

# Multiple testing

## Example:

- Four independent groups of treatments (placebo, drug A, drug B, drug C)
- Measurement of blood pressure
- Do the blood pressures vary between the drugs?
- Pairwise comparisons:



→ 6 pairwise comparisons

- How likely do we erroneously reject at least one of the six $H_0$ given that there is in truth no difference between any of the four treatment groups? 5%?
- This is different from ANOVA / Kruskal-Wallis test, where $H_0$ is an overall hypothesis, i.e. none of the groups differ from each other.

# Multiple testing

Let’s assume there is in truth no difference in blood pressure between any of the four treatment groups.

- Probability to reject $\boldsymbol{H_0}$, if $\boldsymbol{H_0}$ is true, in one test:

$$\alpha = 0.05$$

- Probability not to reject $\boldsymbol{H_0}$, if $\boldsymbol{H_0}$ is true, in one test:

$$1 - \alpha = 0.95$$

- Probability not to reject $\boldsymbol{H_0}$, if $\boldsymbol{H_0}$ is true, in two tests:

$$0.95 \cdot 0.95 = 0.9025$$

- Probability to reject at least one $\boldsymbol{H_0}$, if $\boldsymbol{H_0}$ is true, in two tests:

$$1 - 0.95 \cdot 0.95 = 1 - (1 - 0.05)^2 = 0.0975$$

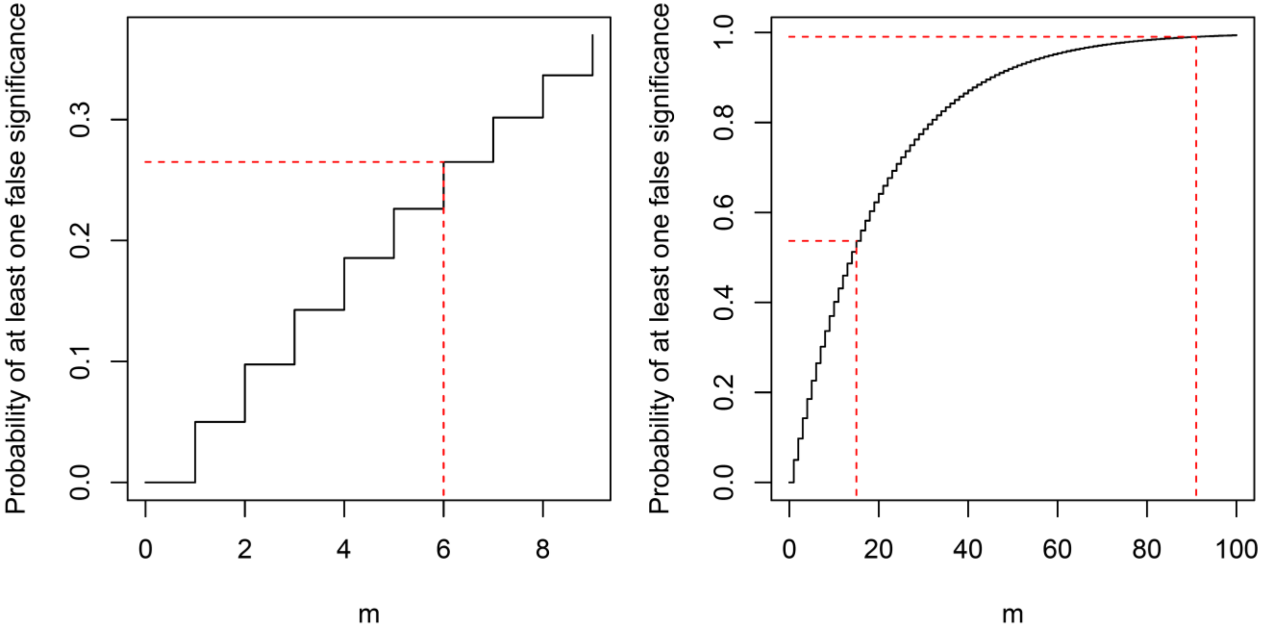
- Probability to reject at least one $\boldsymbol{H_0}$, if $\boldsymbol{H_0}$ is true, in six tests:

$$1 - (1 - 0.05)^6 = 0.2649$$

# Multiple testing

Probability of observing at least one significant result just due to chance:

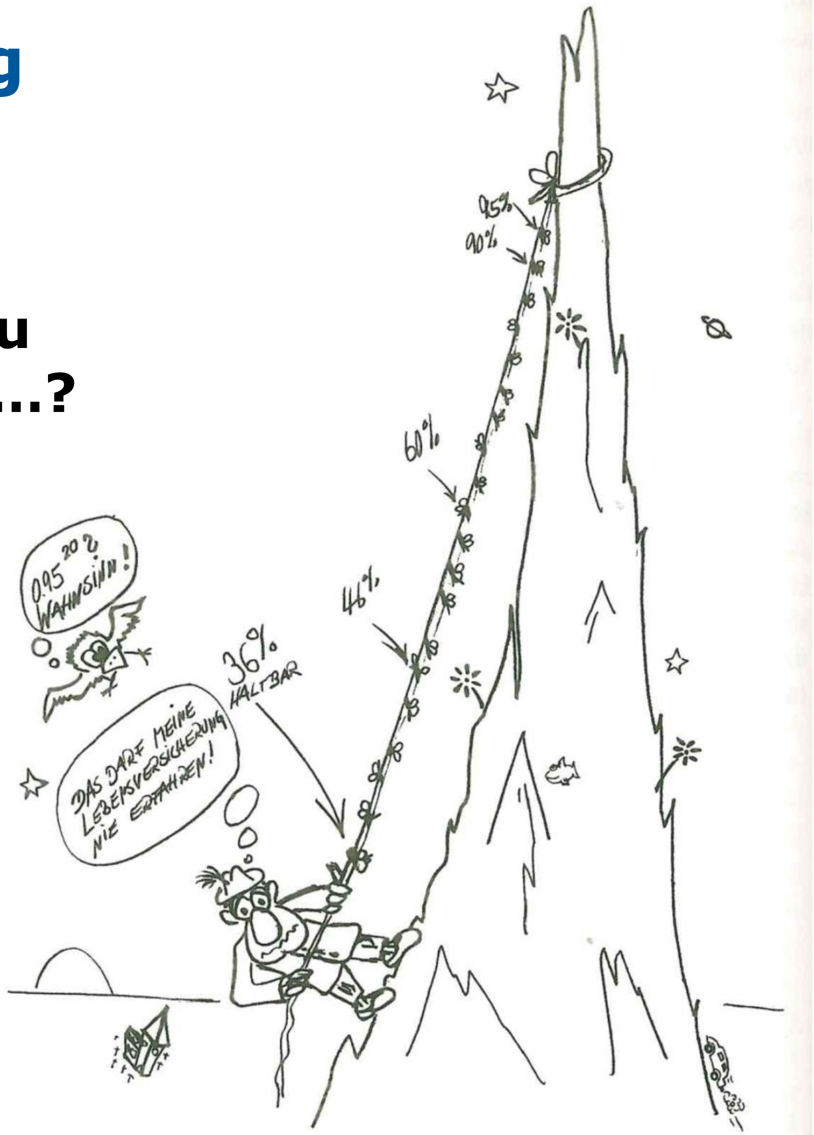
$$\alpha^* = 1 - (1 - \alpha)^m$$



4 groups → 6 pairwise comparisons → $\alpha^* = 26\%$<br>
6 groups → 15 pairwise comparisons → $\alpha^* = 54\%$<br>
14 groups → 91 pairwise comparisons → $\alpha^* = 99\%$

# Multiple testing

**Would you trust this…?**



*Dubben / Beck-Bornholdt: Der Hund, der Eier legt*

# Multiple testing

- Sometimes the study design requires to do multiple tests, e.g. in biomarker discovery studies.

- There are strategies to correct the significance level $\alpha$ (or the p-values) for multiple testing.

- All these strategies keep the significance level $\alpha$ for the overall analysis.

- Example: For our six comparisons of placebo and drugs A, B, C we get the following p-values:

| Placebo / A | Placebo / B | Placebo / C | A / B | A / C | B / C |
|---|---|---|---|---|---|
| 0.2087 | 0.8176 | 0.0225 | 0.2920 | 0.0014 | 0.0097 |

→ Tests A / C, B / C and Placebo / C are considered significant without correction

# Bonferroni correction

- Divide the significance level by the number of tests: $\alpha_{bon} = \frac{\alpha}{m}$
- Compare all p-values to $\alpha_{bon}$
- Intuitive and easy-to-apply approach, but conservative (i.e. losing power)

- **Example:** 6 tests

  → Corrected alpha level: $\alpha_{bon} = 0.05/6 = 0.0083$

  → Original p-values:

| Placebo / A | Placebo / B | Placebo / C | A / B | **A / C** | B / C |
|---|---|---|---|---|---|
| 0.2087 | 0.8176 | 0.0225 | 0.2920 | **0.0014** | 0.0097 |

  → Only test A / C is significant considering Bonferroni correction, instead of additionally Placebo / C and B / C without correction.

# Holm correction

1. Sort p-values from lowest to highest $p_{(k)}$

| | A / C | B / C | Placebo / C | Placebo / A | A / B | Placebo / B |
|---|---|---|---|---|---|---|
| $p_{(k)}$ | 0.0014 | 0.0097 | 0.0225 | 0.2087 | 0.2920 | 0.8176 |

2. Calculate $\alpha_{holm} = \frac{\alpha}{m+1-k}$ with $m$ the number of hypotheses and $k$ the current test

| | $\alpha/(6+1-1)$ | $\alpha/(6+1-2)$ | $\alpha/(6+1-3)$ | $\alpha/(6+1-4)$ | $\alpha/(6+1-5)$ | $\alpha/(6+1-6)$ |
|---|---|---|---|---|---|---|
| $\alpha_{holm}$ | 0.0083 | 0.0100 | 0.0125 | 0.0167 | 0.025 | 0.05 |

3. Consider all tests to be significant until the smallest $k$ with $p_{(k)} > \alpha_{holm}$

| | **A / C** | **B / C** | Placebo / C | Placebo / A | A / B | Placebo / B |
|---|---|---|---|---|---|---|
| $p_{(k)}$ | **0.0014** | **0.0097** | 0.0225 | 0.2087 | 0.2920 | 0.8176 |

→ Now, the comparisons A / C and B / C are considered significant.
**(!)** The testing procedure stops once a failure to reject occurs **(!)**

# Benjamini-Hochberg (FDR)

FDR = False discovery rate = proportion of incorrect rejections among all rejections of the null hypothesis. (→ proportion of significant results being false positives)

1. Sort p-values from lowest to highest $p_{(k)}$

| | A / C | B / C | Placebo / C | Placebo / A | A / B | Placebo / B |
|---|---|---|---|---|---|---|
| $p_{(k)}$ | 0.0014 | 0.0097 | 0.0225 | 0.2087 | 0.2920 | 0.8176 |

2. Calculate $\alpha_{fdr} = \frac{k}{m} \cdot \alpha$ with $m$ the number of hypotheses and $k$ the current test

| | $\frac{1}{6} \cdot \alpha$ | $\frac{2}{6} \cdot \alpha$ | $\frac{3}{6} \cdot \alpha$ | $\frac{4}{6} \cdot \alpha$ | $\frac{5}{6} \cdot \alpha$ | $\frac{6}{6} \cdot \alpha$ |
|---|---|---|---|---|---|---|
| $\alpha_{fdr}$ | 0.0083 | 0.0167 | 0.0250 | 0.0333 | 0.0417 | 0.0500 |

3. Consider all tests to be significant until the largest $k$ with $p_{(k)} < \alpha_{fdr}$

| | **A / C** | **B / C** | **Placebo / C** | Placebo / A | A / B | Placebo / B |
|---|---|---|---|---|---|---|
| $p_{(k)}$ | **0.0014** | **0.0097** | **0.0225** | 0.2087 | 0.2920 | 0.8176 |

→ Tests A / C, B / C and Placebo / C are considered significant.

# Multiple testing - differences of the methods

- Family Wise Error Rates (FWER):
  Probability of getting at least one false positive
  **Bonferroni**, **Holm**

- False Discovery Rate (FDR):
  Expected fraction of false positive results among all rejected hypothesis
  **Benjamini-Hochberg**

| **FWER** | **FDR** |
|---|---|
| More conservative:<br>less likely to accept a false positive | Higher power:<br>more likely to find “differences” |
| If you really don’t want to have any false positive | If you are accepting a few false positives |
| Confirmatory analysis, e.g. Registration of drugs | Exploratory analysis, e.g. Screening of features for further investigation |

# Multiple testing - Summary

- Adjust for multiple testing if you test several hypotheses at once and mention it in your methods / results!
- One can either adjust the significance level $\alpha$ or the p-values

  → reject $\boldsymbol{H_0}$ either if $p \leq \alpha_*$ or if $p_* \leq \alpha$

  → if you adjust for $\alpha$, you can still see the original p-values which makes your analysis more transparent

  → R adjusts the p-values

- **Bonferroni** is quite conservative, i.e. has less power
- **Holm** is less conservative and has always more statistical power than Bonferroni
- **Benjamini-Hochberg** is often used in analyses of high-dimensional data

- Good scientific practice: Decide for one approach <u>before</u> you see the results!

# Summary: Hypothesis tests

**Important issues:**

- Define $\boldsymbol{H_0}$ and $\boldsymbol{H_1}$
- Decide about two-sided vs. one-sided hypothesis (default: two-sided)
- Define significance level $\alpha$ (default: 0.05)
- Consider the following characteristics:
  - One-sample vs. two-sample (vs. multi-sample) tests
  - Dependent vs. independent samples
  - Parametric vs. non-parametric tests
- Be careful with the interpretation of p-values!
- If necessary, decide about strategy to handle multiple testing

# Helpful literature

- Testing statistical Hypotheses (Lehmann, Romano)
- Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133
- Introduction to Statistics and Data Analysis - With Exercises, Solutions and Applications in R (Heumann, C. et al.) (in English)
- Statistik: Der Weg zur Datenanalyse (Fahrmeir, L. et al.)
- Nichtparametrische statistische Methoden (Büning, H.; Trenkler, G.)
- Der Hund, der Eier legt (Dubben, H.-H.; Beck-Bornholdt, H.-P.)