HEIDELBERG UNIVERSITY HOSPITAL

imbi Heidelberg Institute of Medical Biometry

Basic Statistics for Biologists

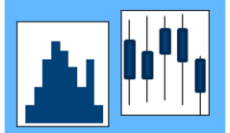
26th of February 2025

Recap of Day 1

Descriptive analysis

Data can be described by:

- measures of location
- measures of variation
- graphics

scale	statements	location	variation	graphics	examples
ratio	= ≠ < > + - ● :	mean median	Q1 & Q3 min & max variance SD IQR range		age, weight, length
interval	= ≠ < > + -	mean median	Q1 & Q3 min & max variance SD IQR range		date of birth
ordinal	= ≠ < >	median mode	Q1 & Q3 min & max (IQR[1] range[1])		scores
nominal	= ≠	mode			gender, fur color

Exercise: Description

Think of ways to improve the following description:

The study population had a mean age of 44 years and was mostly female (68%). On a score of 1-5, the patients had a mean treatment satisfaction of 3.5 (see Figure 1).

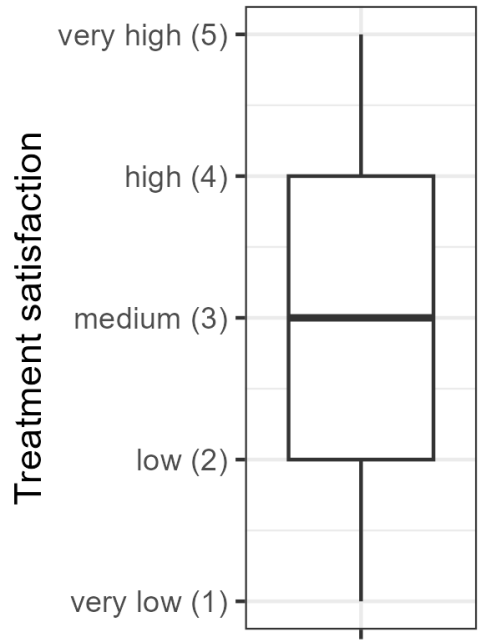


Figure 1

Solution: Description

The study population (n = 20) had a mean age of 44 years (SD = 5.3 years) and was mostly female (68%). On a score of 1-5, the patients had a median treatment satisfaction of 3 and a mode of 4 (see Figure 1).

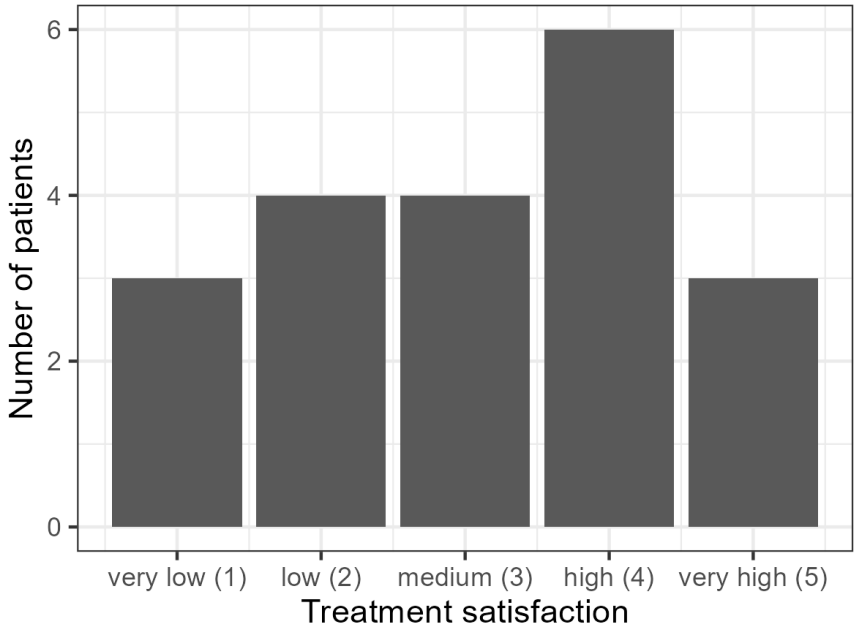


Figure 1

The principle of statistical tests

1. Formulate null- and alternative hypothesis.
2. Collect the data.
3. Find the appropriate probability distribution.
4. Compute the p-value and compare to significance level.

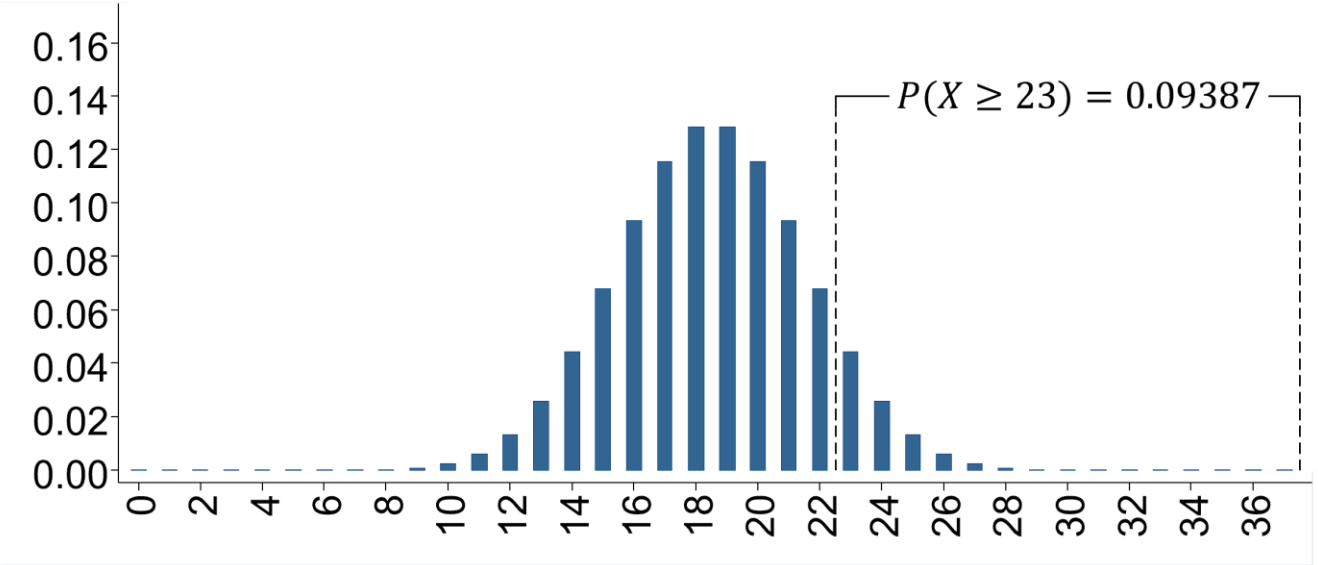
This procedure controls the Type I error.

Some important aspects are missing (Type II error control, multiple testing, …). See later.

P-value

The p-value is the probability of the observed or more extreme values under the assumption of the null hypothesis (H0).

Under the assumption that the male-hamster-probability equals 0.5 (H0), the probability to find at least 23 male hamsters (observed value) or more out of 37 is 0.094 (p-value).



Error probabilities

When testing a hypothesis, two types of false decisions (errors) can be made.

Quelle: Statistical Performance Measures. Statistical performance measures are… | by Neeraj Kumar Vaid | Medium

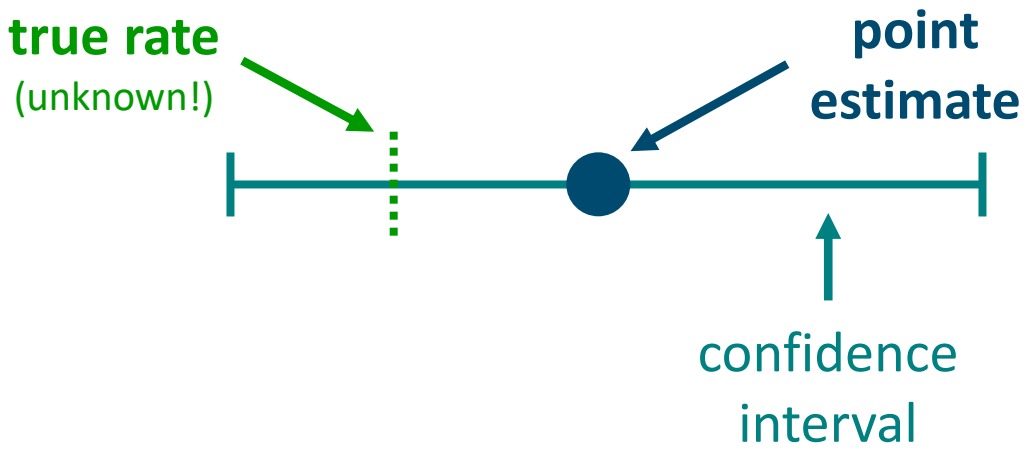
Error probabilities

When testing a hypothesis, two types of false decisions (errors) can be made.

Reality / Decision	H0 is true (no true effect)	H1 is true (true effect)
Decision for H0	Correct	Type II error (= 1 – power)
Decision against H0	Type I error (controlled by α)	Correct

Point estimation and confidence interval

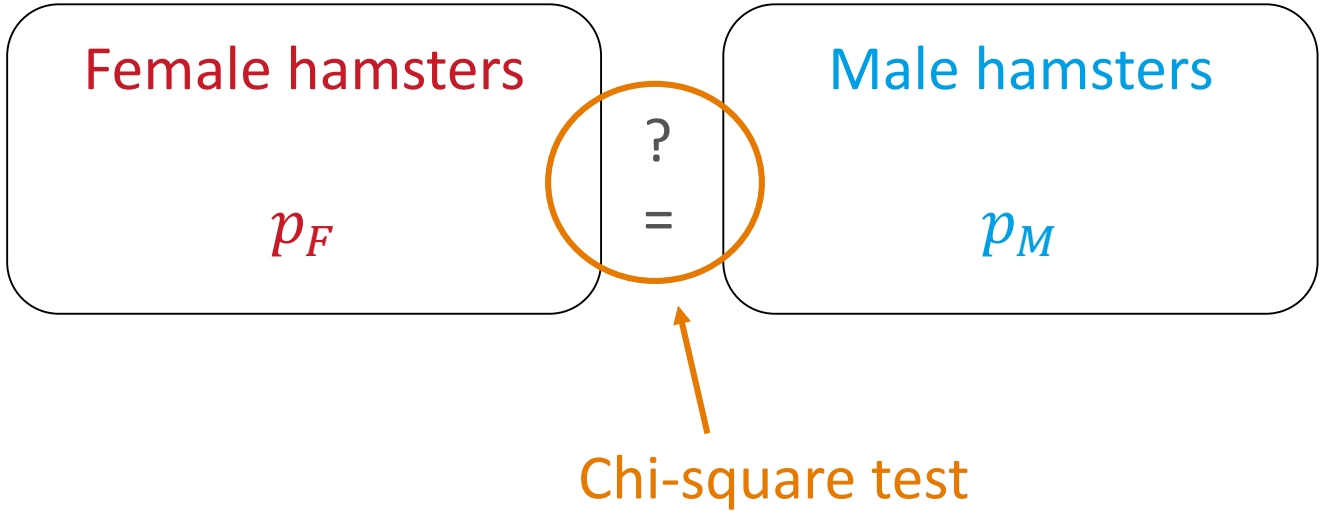
- A point estimate is the best guess for the true value of a parameter (e.g. male-hamster-probability).
- A confidence interval is an interval that likely contains the true effect.



Chi-square test

The chi-square test tests whether event probabilites in two groups are equal.

Probability of event „black fur“



The risk difference is $p_F - p_M$ and can be used to describe the effect.

Day 2

Learning goals

You will be able to

- explain the meaning of the normal distribution.
- interpret parametric and non-parametric reference areas.
- assess the application of t-tests and non-parametric tests.
- interpret results of these tests.
- name corresponding effect sizes.
- explain the principle of sample size estimation and power analysis.
- explain problems and pitfalls of statistical testing and p-values.

The normal distribution

Probability distribution

What is a probability distribution?

„In probability theory and statistics, a **probability distribution** is the mathematical function that gives the probabilities of occurrence of possible outcomes for an experiment. It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).”

https://en.wikipedia.org/wiki/Probability_distribution

“A **probability distribution** describes how the probabilities of different outcomes are assigned to the possible values of a random variable.”

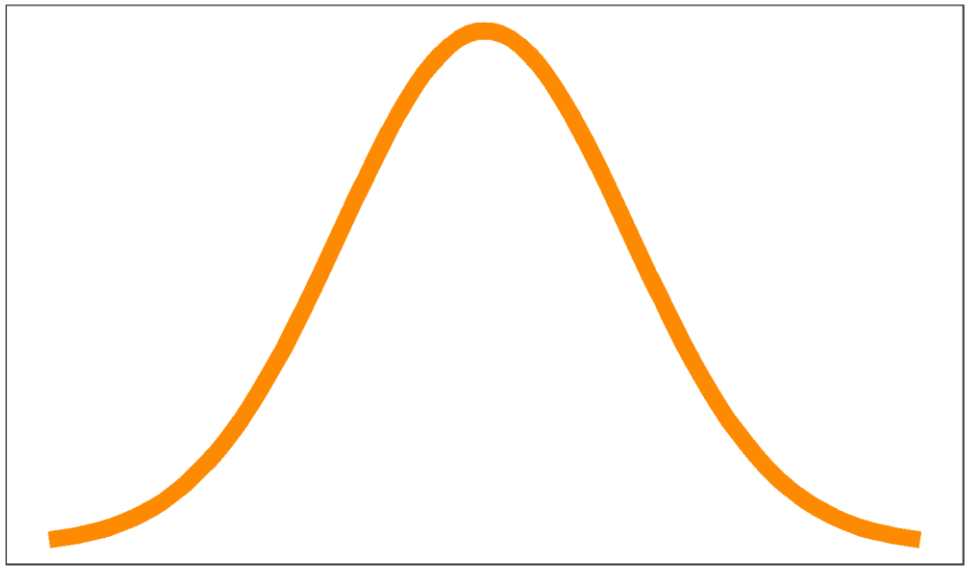
https://www.geeksforgeeks.org/probability-distribution/

Probability distribution

The *probability distribution* (specifically the *density*) arises if we create a histogram from infinitely many observations.

Normal distribution

The most famous example of a probability distribution is the **normal distribution**.

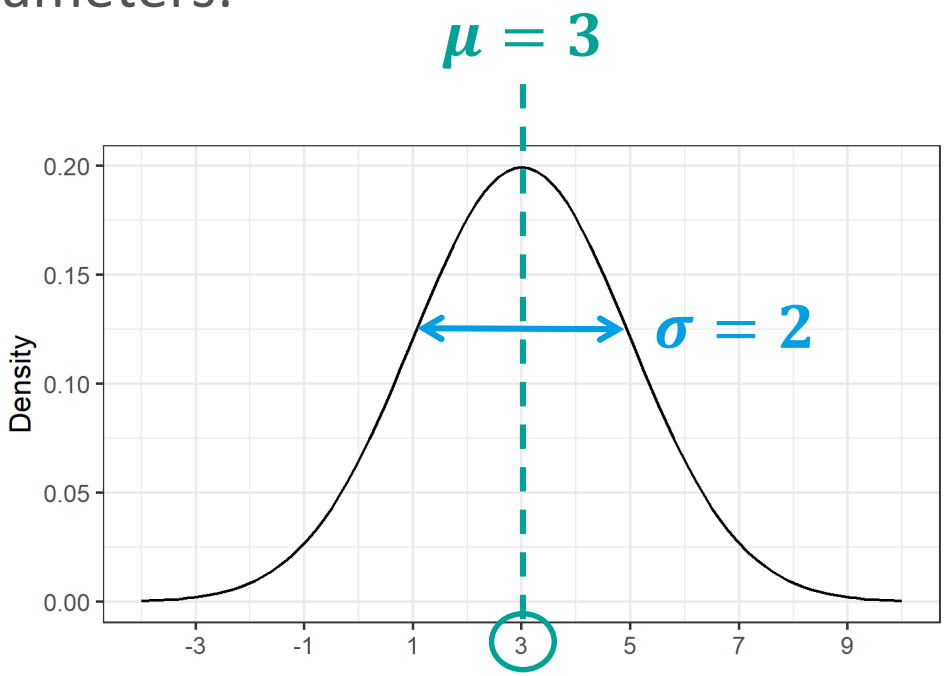


Its probability density is the well-known **Bell curve**.

Theory: Normal distribution

- The normal distribution is defined by two parameters: The **mean μ** and the **standard deviation σ**.
- The **mean** gives the **location** of the distribution.
- The **standard deviation** gives the **spread** of the distribution.
- The curve is defined by the function

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



What makes the *normal distribution* normal?

Two reasons:

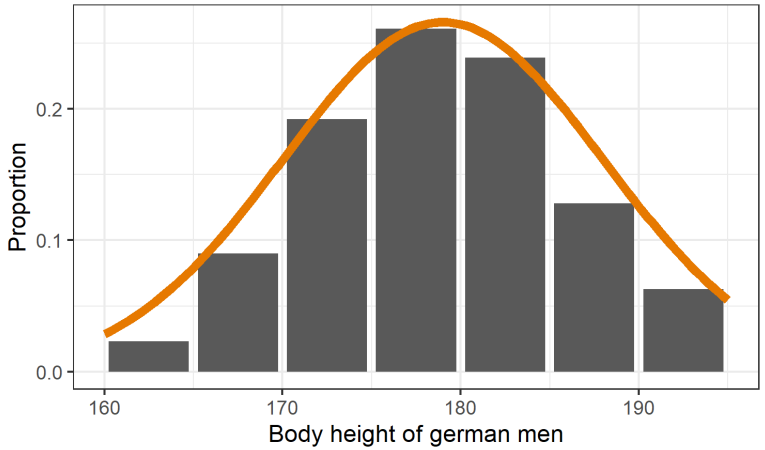
Many „natural“ observations are normally distributed.

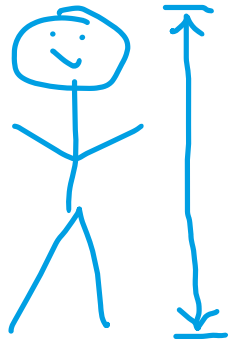
The mean of many independently measured values is approximately normally distributed.

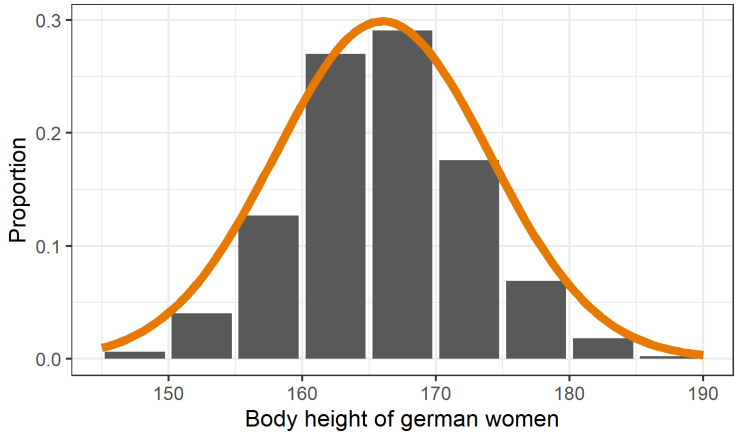
Central limit theorem

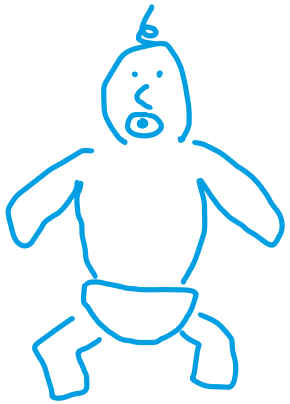
See later.

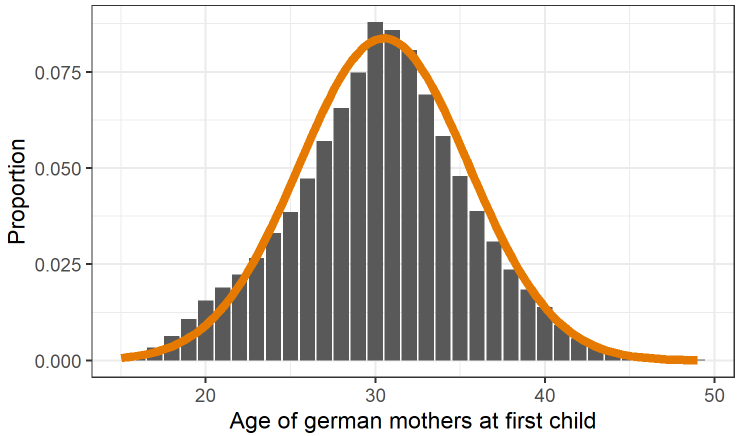
Normal distribution in real life











Reference area

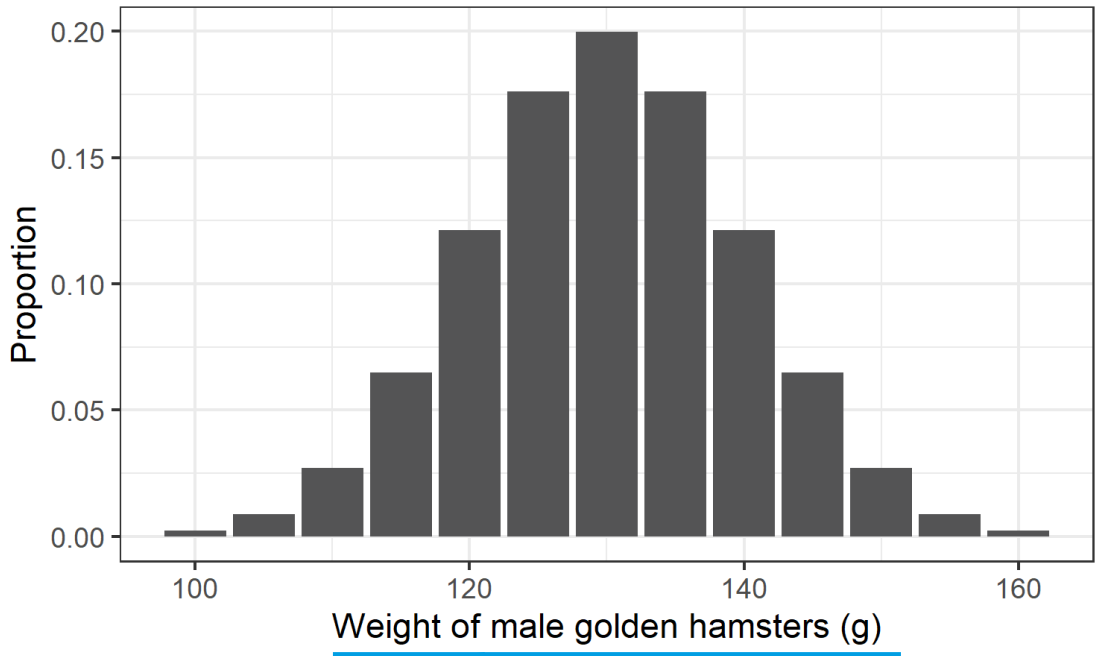
Motivation

Question:

Is my hamster overweight?

Reminder - Histogram

Relative frequency



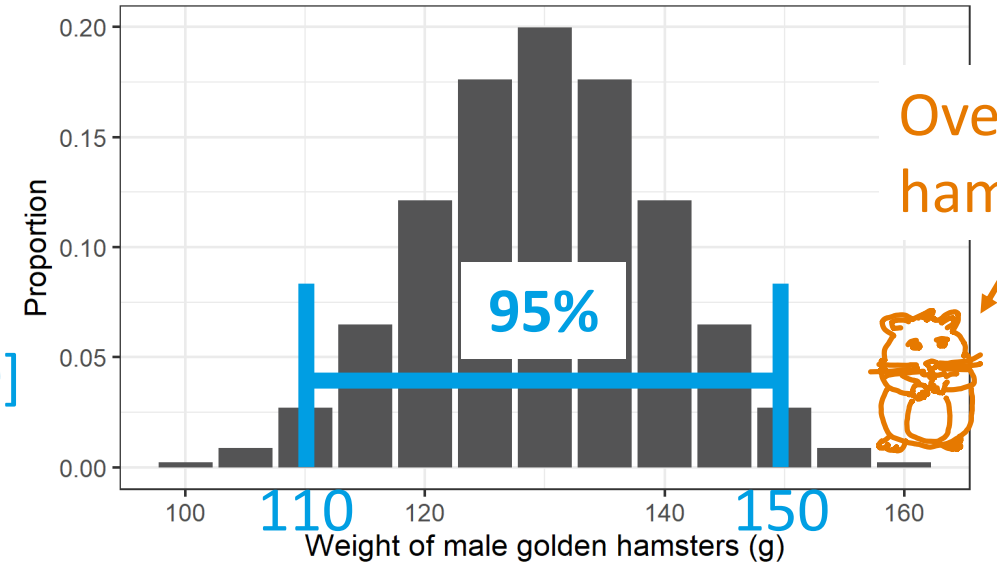
Observed values

Reference areas

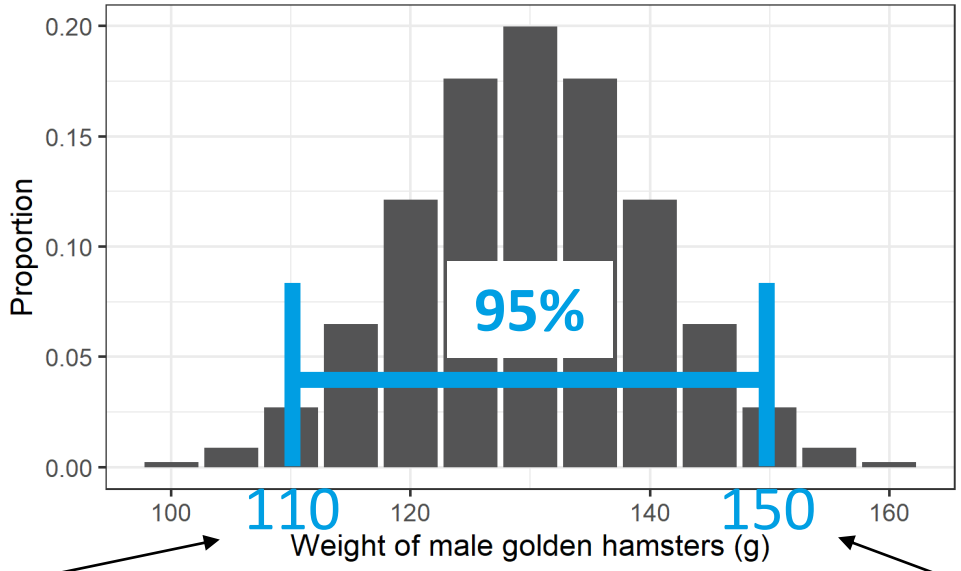
What weight values are normal? What values are abnormally high or low?

- A 95% reference area or 95% reference interval of a variable is the area containing the middle 95%.
- Values outside are usually regarded abnormal.

[110, 150]



Quantiles



Below this value are 2.5% of values. It is the *2.5%-quantile*:

$$q_{0.025} = 110$$

Below this value are 97.5% of values. It is the *97.5%-quantile*:

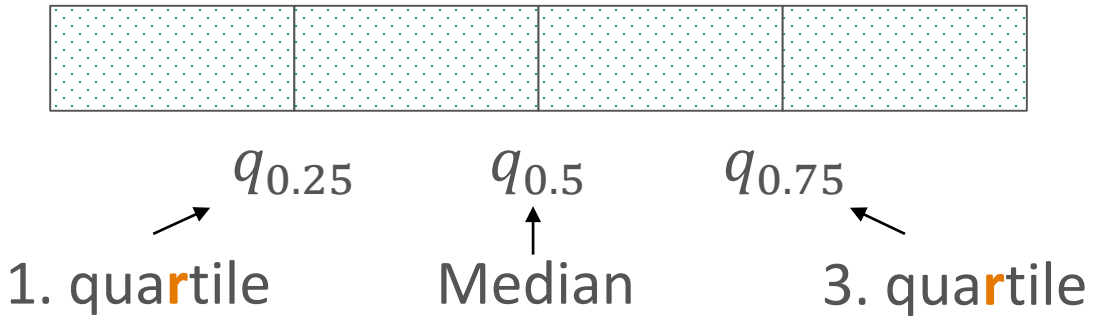
$$q_{0.975} = 150$$

Essential: Reference area and quantiles

The 95% reference area of a variable contains the middle 95% and is defined by the 2.5% and 97.5% quantile:

$$95\%\ \text{reference area} = [q_{0.025}, q_{0.975}]$$

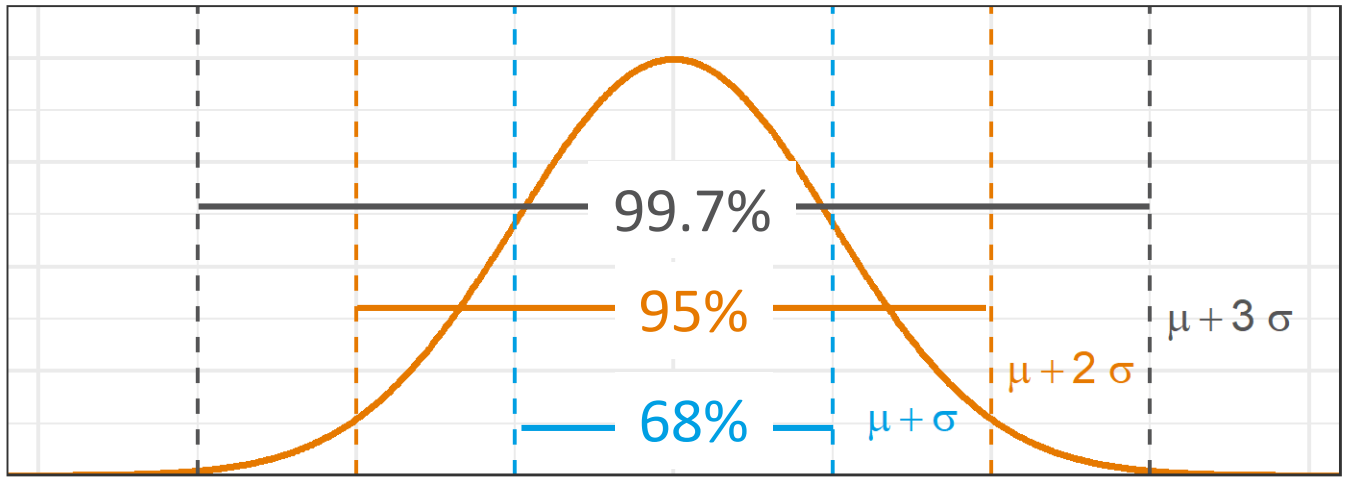
A quantile divides a set of values in two proportions. Below the quantile q_ε is the proportion ε.



Essential: Parametric reference area

If the data is normally distributed (= parametric), reference areas can be calculated easily from mean μ and standard deviation σ.

$$95\%\ \text{reference area} = [\mu - 1.96 \cdot \sigma, \quad \mu + 1.96 \cdot \sigma]$$



Exercise: Reference area

The systolic blood pressure of adults is normally distributed with a mean of 128 and a standard deviation of 20.

What blood pressure values belong to the highest or lowest 2.5%?

Solution: Reference area

The systolic blood pressure of adults is normally distributed with a mean of 128 and a standard deviation of 20.

What blood pressure values belong to the highest or lowest 2.5%?

Parametric 95% reference area $\approx \text{mean} \pm 2\text{SD} = [88; 168]$

Confirmatory analysis of non-binary endpoints

THE GROWTH OF THE ODONTOBLASTS OF THE INCISOR TOOTH AS A CRITERION OF THE VITAMIN C INTAKE OF THE GUINEA PIG [1]

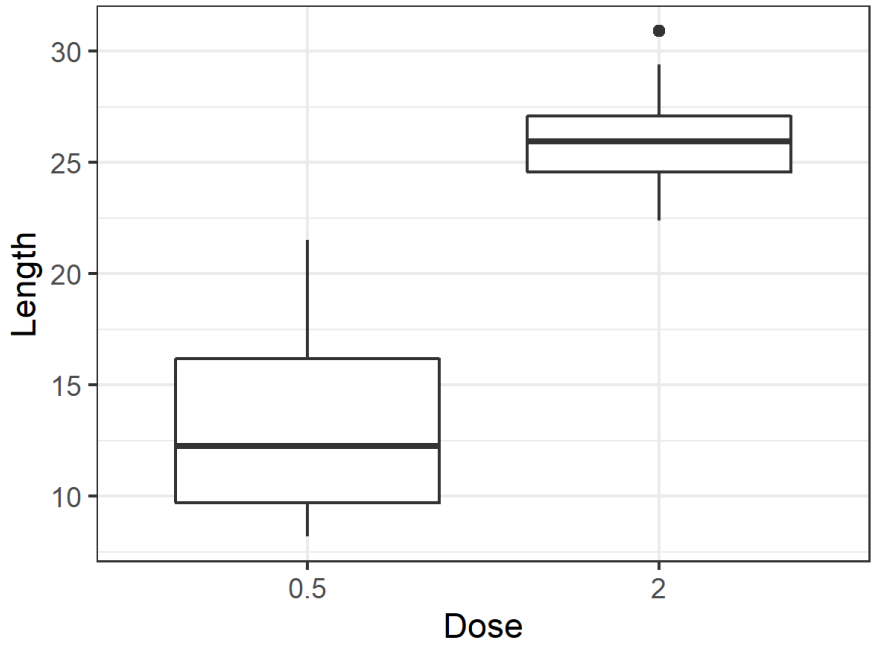
E. W. CRAMPTON

Department of Nutrition, Macdonald College, McGill University, P.O., Prov. Quebec, Canada

Motivation

Does vitamin C supplement dose have an effect on tooth growth in guinea pigs?

- Endpoint: Length of odontoblasts [µm] (cells responsible for tooth growth)
- Treatment: 0.5 mg/day vs. 2 mg/day vitamin C



Is the observed difference by chance?

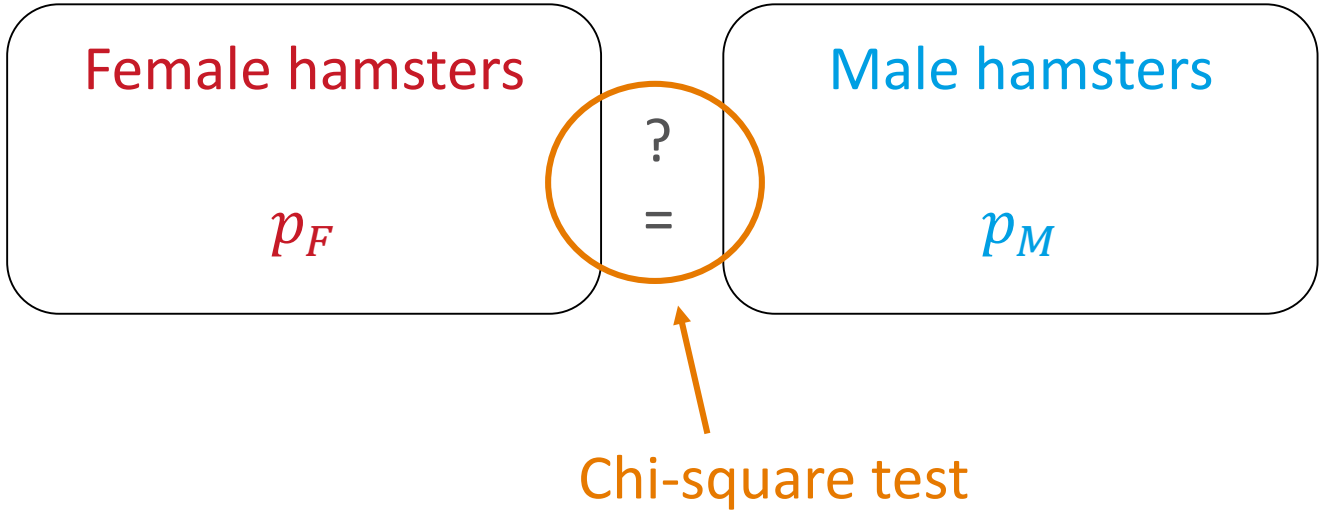
→ hypothesis testing

Reminder: Chi-square test

The chi-square test tests if event probabilites in two groups are equal.

Group

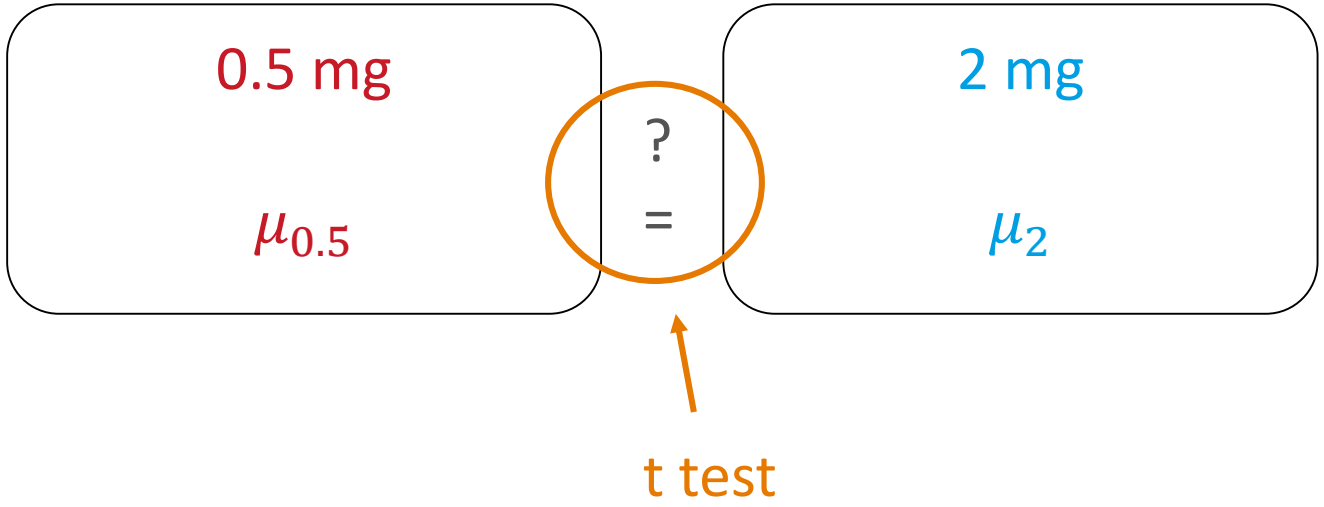
Probability of event „black fur“



Two-group comparison of a binary endpoint.

The t-test

The t-test tests if means in two groups are equal.



Two-group comparison of a numerical endpoint.

Essential: Two-sample t-test

- Setting: Numerical endpoint in two groups A and B.
- Null hypothesis: $\mu_A = \mu_B$ (equality of means)
- Requirements:
  - Normally distributed endpoint
    or
  - Sample size large enough (Rule of thumb: ≥ 30 per group)

Central limit theorem:
The estimated mean (average) is approximately normally distributed if sample size is high. (See later.)

T-test in R

```
> t.test(x = length_0.5, y = length_2)

        Welch Two Sample t-test

data:  length_0.5 and length_2
t = -7.817, df = 14.668, p-value = 1.324e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.335241  -9.324759
sample estimates:
mean of x mean of y
    13.23     26.06
```

p-value

Confidence interval (CI) of the *effect size*

Essential: Effect size – mean difference

The canonical effect size regarding the t-test is the observed mean difference:

$$\hat{\mu}_A - \hat{\mu}_B$$

Estimated mean values

- The p-value only quantifies *how likely the observed difference is random*.
- The effect size is the *best estimator for the true difference*.
- The 95% CI of the effect gives a range that *probably contains the true effect*.

Effect size together with 95% CI contains the most important information:

The mean difference is -12.8 (95% CI [-16.3, -9.3]).

Essential: Standardized mean difference

The mean difference can be standardized to make it unit-independent:

Estimated mean values

Estimated standard deviation

$$\frac{\mu_A - \mu_B}{\sigma}$$

- Also known as Cohen‘s *d*.
- Values ≈ 0.5 are considered medium effect sizes.

Reporting test results - Example

Odontoblast length in the high dose group was significantly higher than in the low dose group with a mean difference of 12.8 µm *(95% CI [9.3, 16.3], $p < 0.001$).*

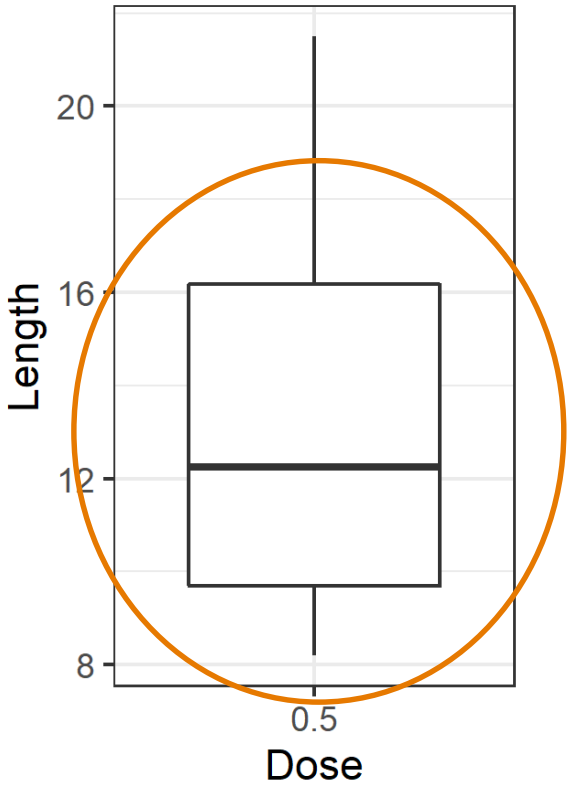
The standardized mean difference of 3.6 indicates a strong effect.

Other effect sizes

The choice of effect size depends on the type of endpoint and the objective.

Setting	Effect size
Binary endpoint, two-group comparison	Risk difference, risk ratio, odds ratio
Normally distributed endpoint, two-group comparison	(standardized) mean difference
Ordinal endpoint, two-group comparison	Common language effect size
Correlation of two endpoints	Pearson‘s correlation coefficient

One sample t-test



Is the length in the low dose group systematically different from 16?

Essential: One-sample t-test

- Setting: Numerical endpoint in one group A.
- Null hypothesis: $\mu_A = \mu_0$ (Comparison of mean to *reference value* μ_0)
- Requirements:
  - Normally distributed endpoint
    or
  - Sample size large enough (Rule of thumb: ≥ 30)

One sample t-test in R

```
> t.test(x = length_0.5, mu = 16)

        One Sample t-test

data:  length_0.5
t = -1.9641, df = 9, p-value = 0.0811
alternative hypothesis: true mean is not equal to 16
95 percent confidence interval:
 10.03972 16.42028
sample estimates:
mean of x
    13.23
```

p-value

Confidence interval (CI) of the *effect size*

Essential: Paired sample t-test

- Setting: Paired values (example: length before and after supplementation)
- Consequence: Paired values are dependent.
- Solution: Calculate difference of paired values and test with one sample t-test.

ID	Length before	Length after
1	12	15
2	19	21
3	16	17
4	12	10

Change?

Equivalent

ID	Length difference
1	3
2	2
3	1
4	-2

Different from 0?

Exercise: Reporting test results

Think of ways to improve the description of test results.

Guinea pigs in the high dose group experienced a significant length gain over two weeks ($p = 0.032$).

Solution: Reporting test results

Ways to improve the description of test results.

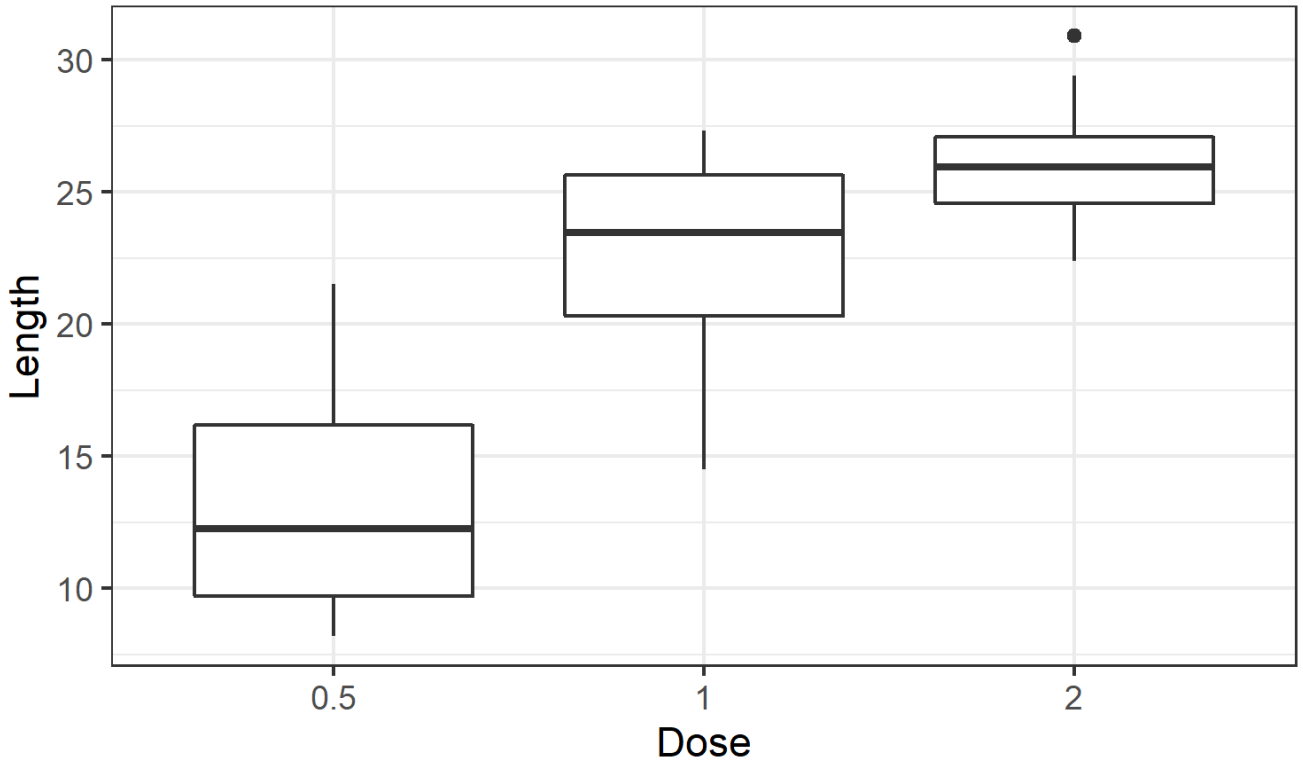
Guinea pigs in the high dose group experienced a mean length gain of 2.1 µm over two weeks (95% CI [0.2; 4.0], p = 0.032).

ANOVA (Analysis of variance)

Comparison of 3 dose groups.

- 0.5 mg
- 1 mg
- 2 mg

Results



Differences by chance?

→ Hypothesis testing

Essential: ANOVA

- Setting: Numerical endpoint in more than two groups.
- Null hypothesis: $\mu_A = \mu_B = \mu_C = \cdots$ (all means are equal)
- Requirements:
  - Normally distributed endpoint
    or
  - Sample size large enough (Rule of thumb: ≥ 30)

Note: Only reasonable if main interest is difference between any of the groups.

If pairwise group comparisons are of interest, plan multiple comparisons (with adjustment for multiple testing).

ANOVA in R

```
> summary(aov(len ~ dose, data = df))
            Df Sum Sq Mean Sq F value   Pr(>F)
dose         1  711.9   711.9   36.01 1.82e-06 ***
Residuals   28  553.5    19.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value

Effect size?

There is no „one“ effect size.

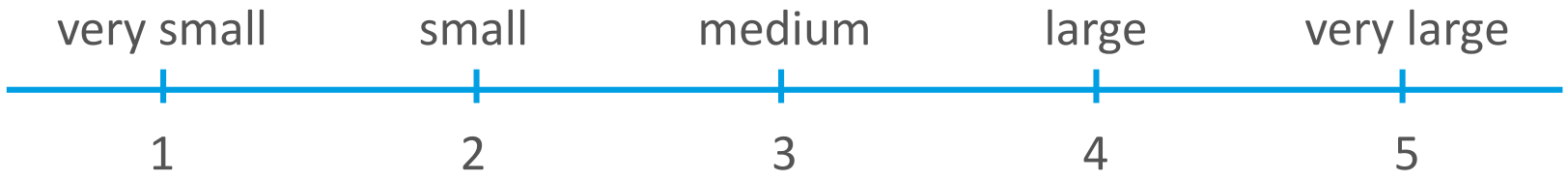
Pairwise t-tests in R

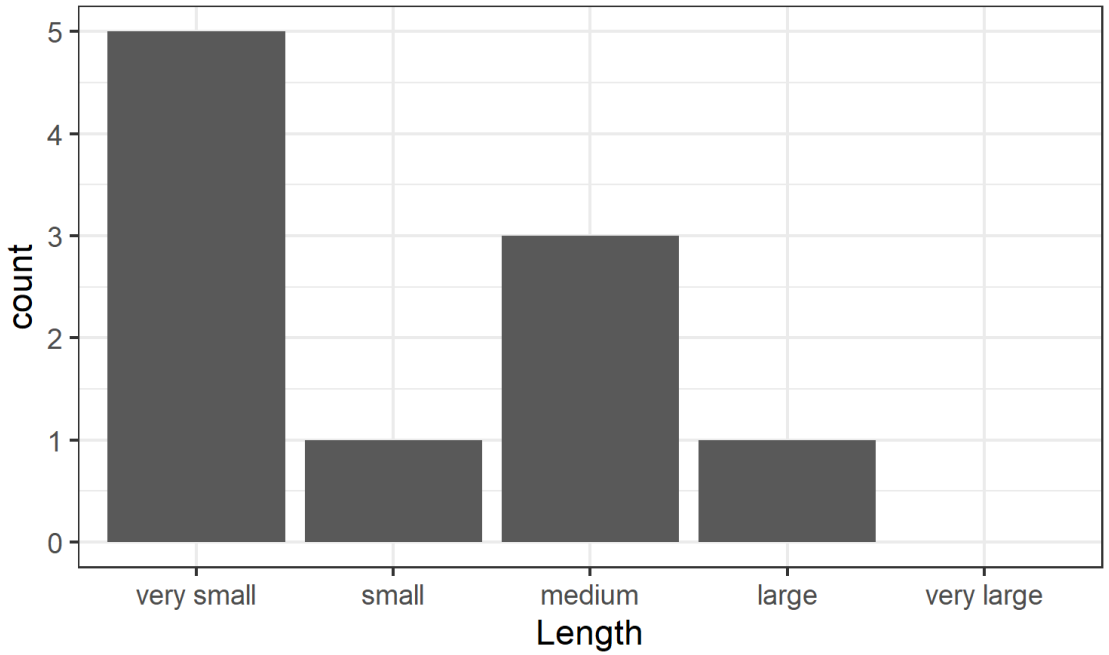
```
> pairwise.t.test(df$len, df$dose)

        Pairwise comparisons using t tests with pooled SD

data:  df$len and df$dose

  0.5     1
1 1.1e-05 -
2 9.5e-08 0.055

P value adjustment method: holm
```

adjusted p-values of pairwise comparisons

What if variable is not normally distributed?

Length of odontoblasts:



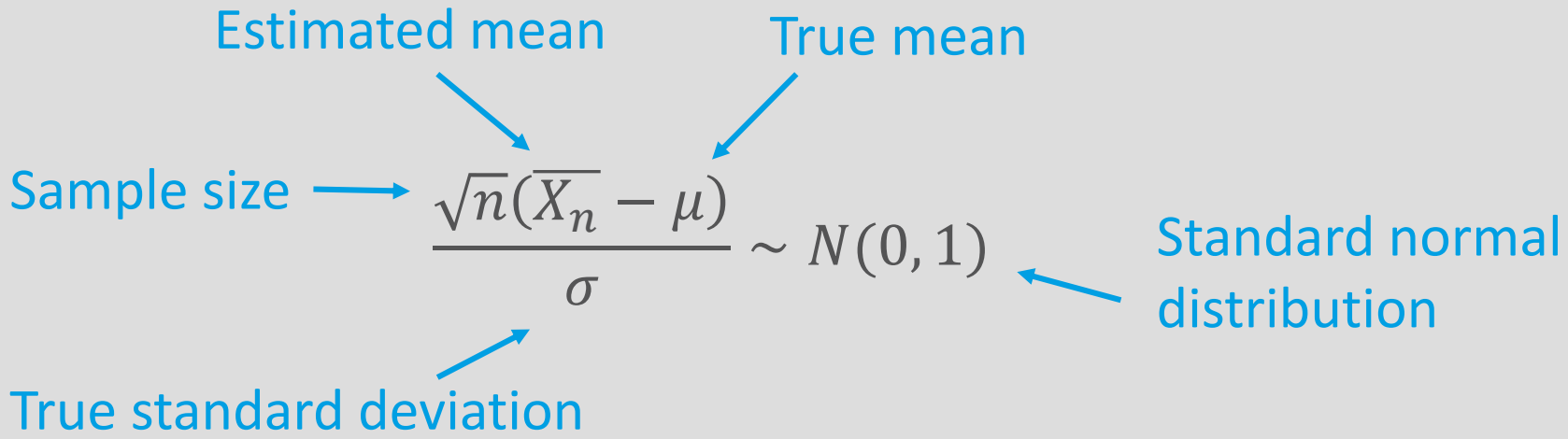


Essential: Central limit theorem

The estimated mean is asymptotically normally distributed.

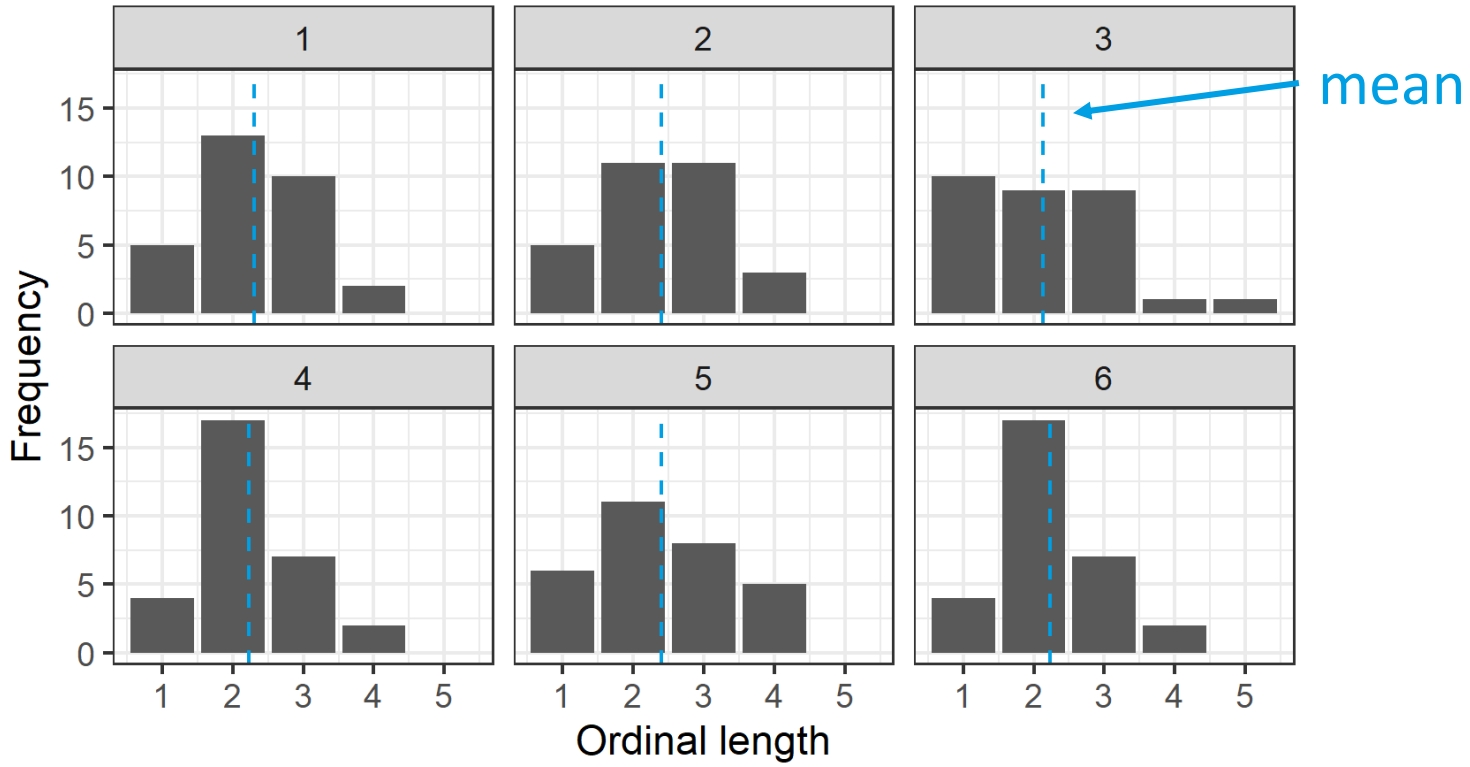
asymptotically = „for large sample sizes“

Theory



Central limit theorem - Example

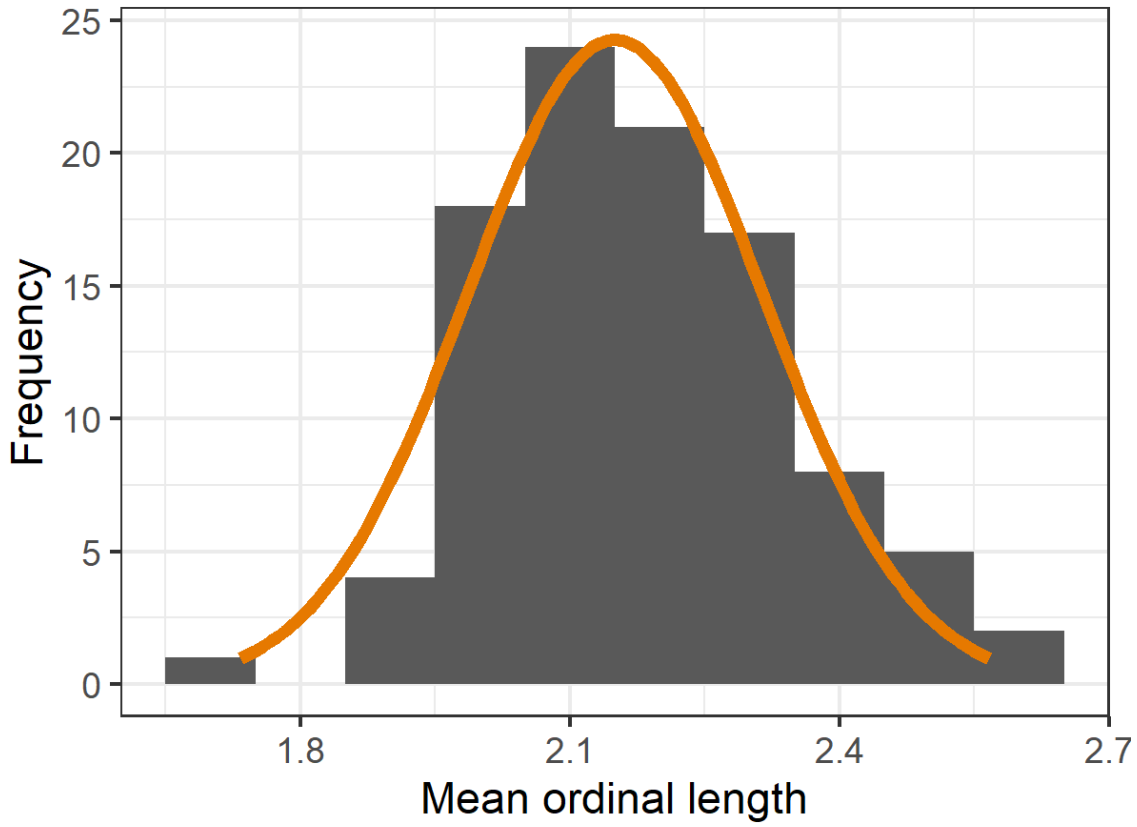
Many different experiments: each measures ordinal length of 30 guinea pigs



Central limit theorem - Example

Histogram of the means:

Normally distributed!

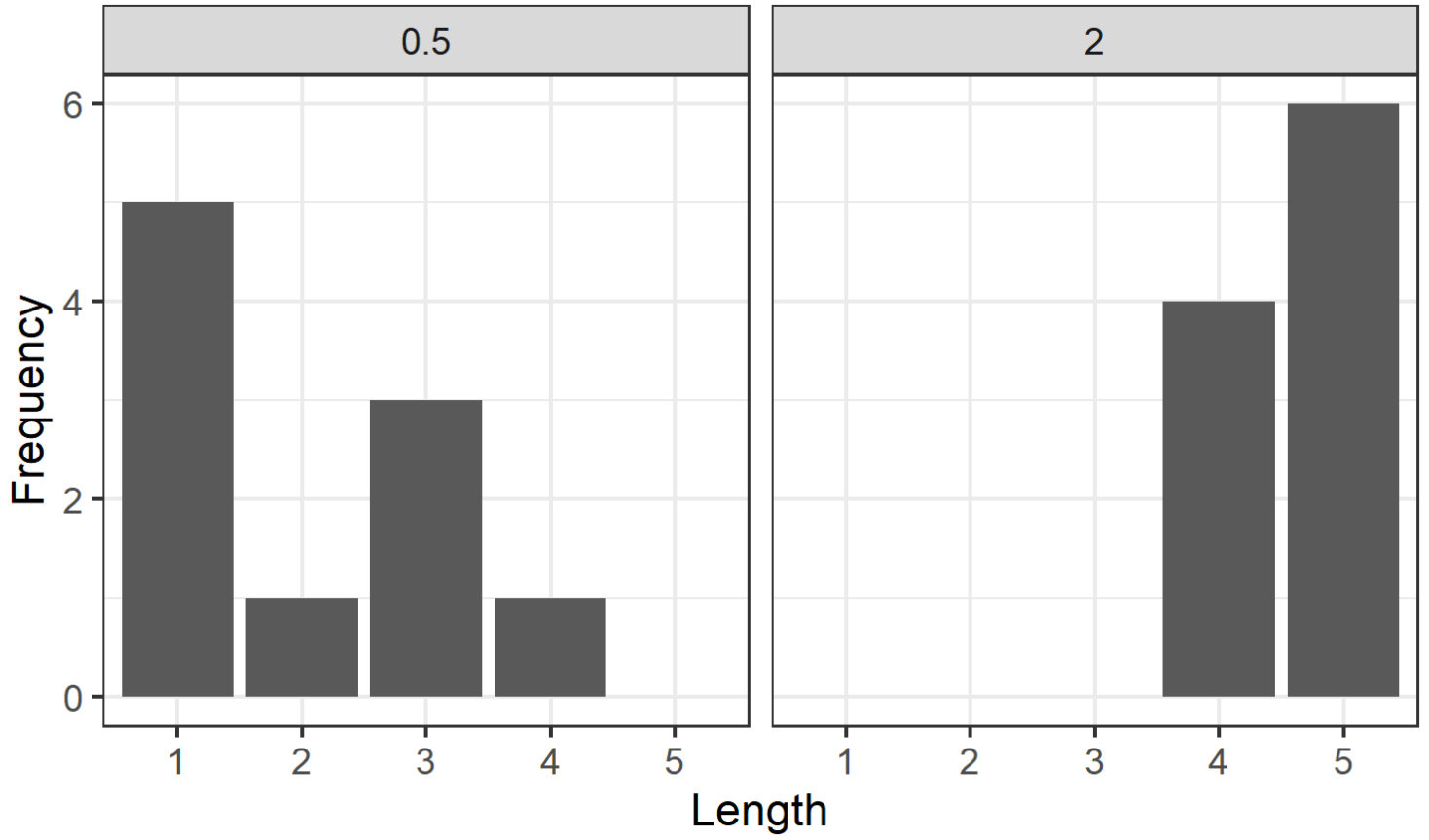


Non-parametric tests

Non-parametric tests are used if

- endpoint is not normally distributed (includes ordinal endpoints).
- sample size is small.
- mean is not of main interest.

Effect of vitamin C dose on length (ordinal)



Essential: Mann-Whitney U test

- Setting: At least ordinal endpoint in two groups A and B.
- Null hypothesis: Equality of distributions.
- Requirements: None.

Mann-Whitney U test in R

```
> wilcox.test(length ~ dose, data = df.MWU)

        Wilcoxon rank sum test with continuity correction

data:  length by dose
W = 2, p-value = 0.0002137
alternative hypothesis: true location shift is not equal to 0
```

p-value

The corresponding effect size is the *common language effect size*.

What test should I use?

It’s complicated!

- Often: don’t use a test! Instead, description and visualization of effects.
- What hypothesis should be tested/rejected?
- How many groups?
- Level of measurement?
- Assumptions about the distribution of endpoint?

Seek consultation!

So far...

- We know how to perform an appropriate statistical test for one hypothesis.
- Controls Type I error for this hypothesis.

What is missing?

- Control of Type II error -> Sample size calculation
- Testing of multiple hypotheses -> Adjusting for multiple testing

Reminder: Error probabilites

When testing a hypothesis, two types of false decisions (errors) can be made.

Reality / Decision	No true effect	True effect
Negative	Correct	Type II error (= 1 – power)
Positive	Type I error (controlled by α)	Correct

Power to the tests!

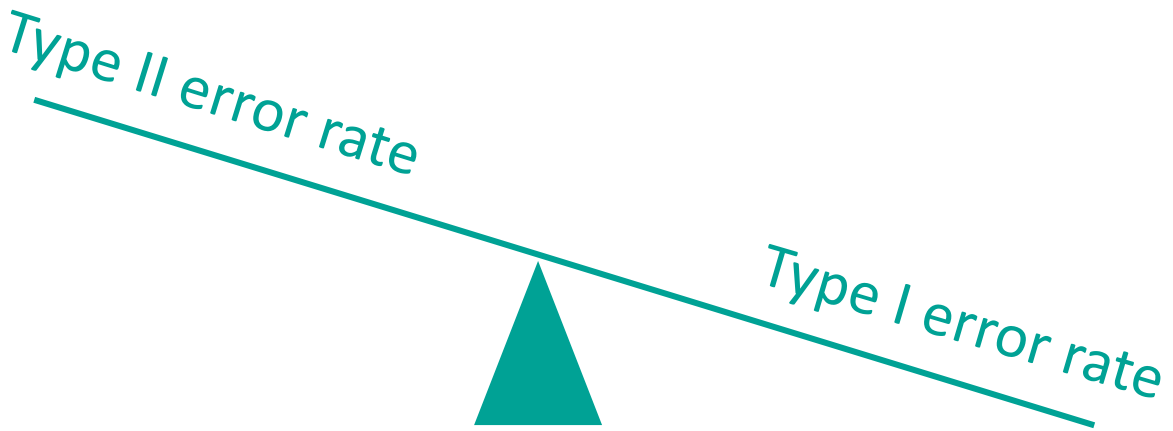
Statistical tests are used to make decisions with low error probabilities.

- Significance level 0.05 → Type I error rate ≤ 0.05

But: Testing without sample size calculation/power analysis is pointless since Type II error could be arbitrarily high!

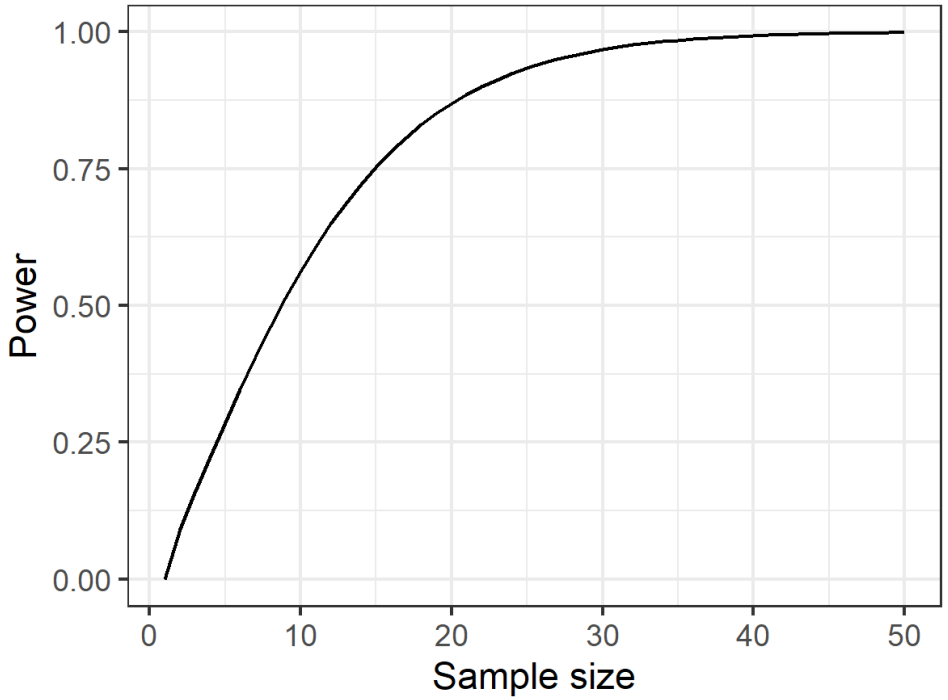
Statistical uncertainty relation

- The decision rule (for or against the hypothesis) cannot be chosen such that both error rates are minimal.
- A smaller Type 1 error rate leads to a higher Type 2 error rate (and vice versa).



The effect of sample size

The only way to decrease type II error without increasing type I error is to **increase the sample size**.



Power of t test

- Standardized mean difference = 1
- Type I error rate = 0.05

Essential: Sample size calculation

Sample size calculation is performed to reduce the probability of a Type II error.

- Choose an appropriate statistical test and significance level.
- Choose an effect size that you want to detect (realistic and relevant).
- Guess the variability in your data.
- Choose the power for showing that effect size.

Can be combined in a standardized effect

→ Calculate sample size

Exercise: Effects on power

How is the power affected by

- a greater sample size?
- a greater significance level (alpha)?
- a greater mean difference?
- a greater standard deviation?

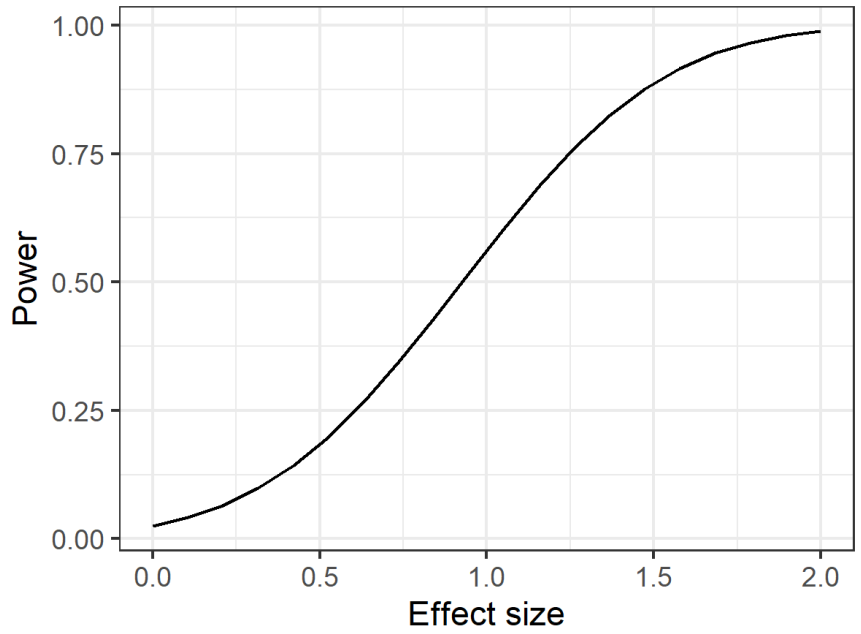
Solution: Effects on power

How is the power affected by

- a greater sample size? Power increases.
- a greater significance level (alpha)? Power increases.
- a greater mean difference? Power increases.
- a greater standard deviation? Power decreases.

Power depends on effect size

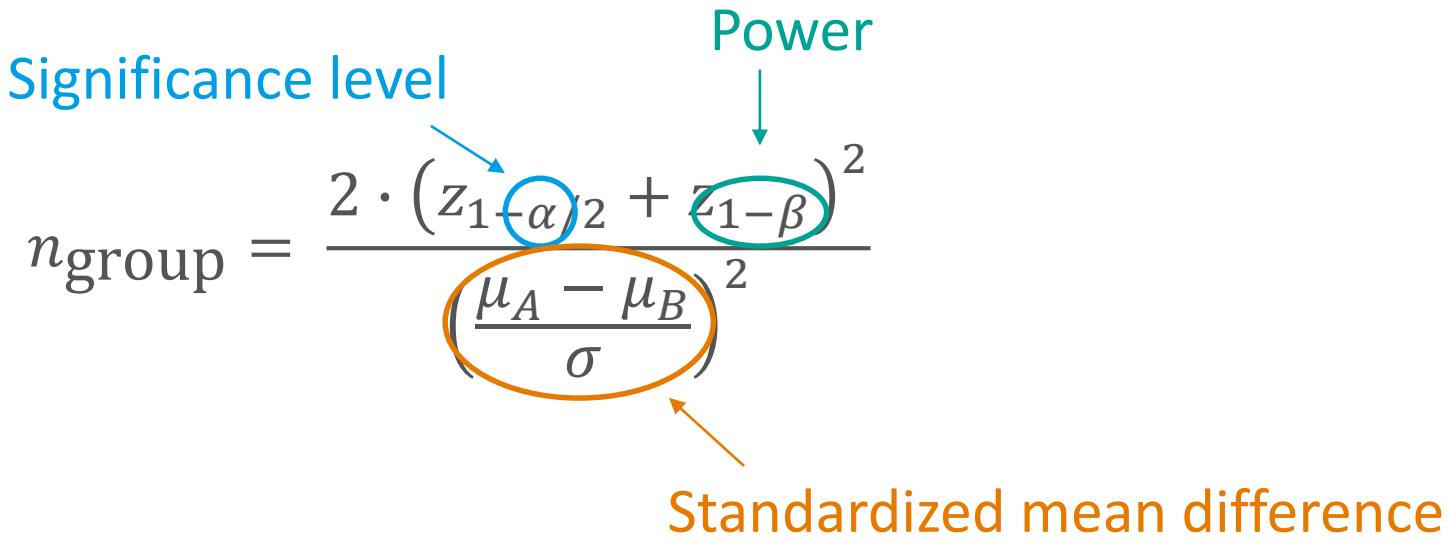
- Power is the probability of „proving“ a certain alternative hypothesis.
- The power depends on *what you want to show (effect size)* .



Power of t test at level 0.05 with sample size 10 per group.

Theory: Sample size of two sample t test

The sample size for the two sample t test can be calculated by:



t test sample size in R

```
> power.t.test(delta = 15, sd = 10, power = 0.8, sig.level = 0.05)
```

mean difference

standard deviation

power

significance level

```
     Two-sample t test power calculation

              n = 8.06031
          delta = 15
             sd = 10
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

sample size per group

```
NOTE: n is number in *each* group
```

Essential: Power analysis

If we have no control over the sample size (*we got only 4 mice*)
but still want to perform a test, we should do a power analysis.

- Choose an appropriate statistical test and significance level.
- Choose an effect size that you want to detect (realistic and relevant).
- Guess the variability in your data.
- Choose the sample size.

→ Calculate power

t test power analysis in R

```
> power.t.test(delta = 15, sd = 10, n = 4, sig.level = 0.05)

     Two-sample t test power calculation

              n = 4
          delta = 15
             sd = 10
      sig.level = 0.05
          power = 0.4299273
    alternative = two.sided

NOTE: n is number in *each* group
```

mean difference

standard deviation

sample size per group

significance level

power

Power to the tests! Again!

If the power is low, there is little chance of getting a small p-value,
even if there is a relevant effect.

So why bother testing?

Instead, describe and visualize effect sizes and uncertainty.

Essential: When (not) to test

Hypothesis testing is for testing (prespecified) hypotheses!

Requirements:

- One prespecified hypothesis (or adjustment for multiple testing, see later).
- Sample size calculation (or power analysis).

Controls Type I error

Controls Type II error

Do not test if objective is exploratory: *Where is an effect? What is the effect size?*

Instead, focus on description and visualization of effect estimates and uncertainty.

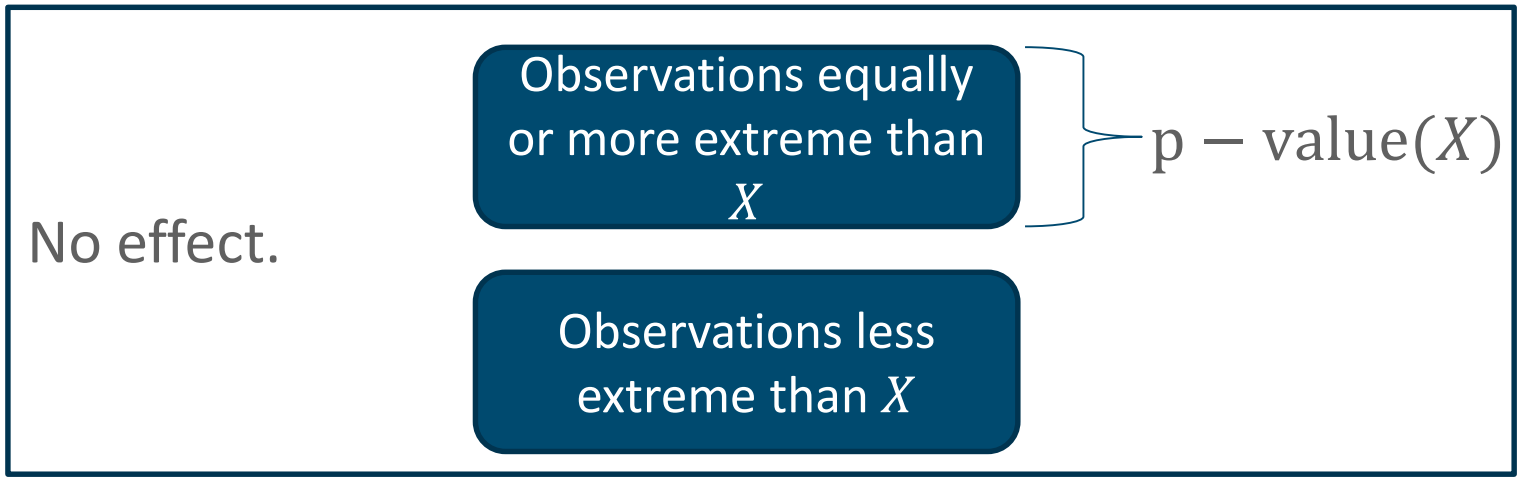
Problems with inappropriate testing

What is the p-value?

When testing a **null hypothesis**

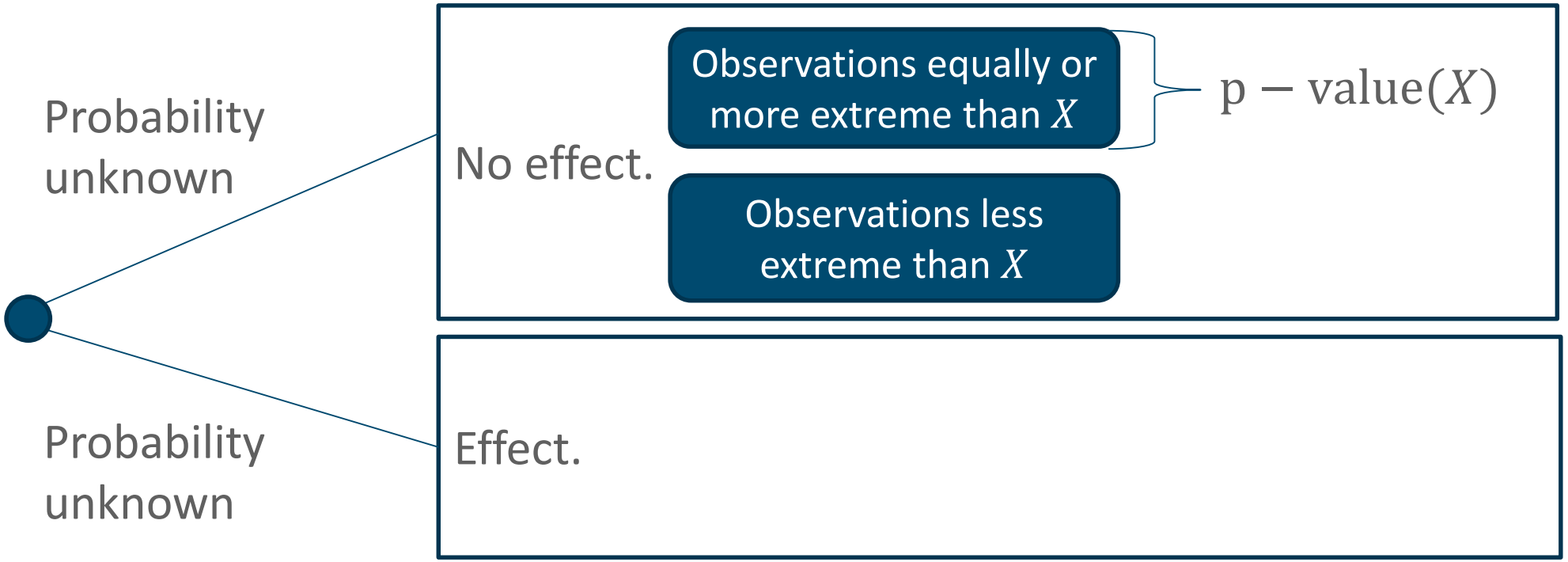
$$\boldsymbol{H_0}\textbf{: There is no true effect,}$$

the **p-value** of an experimental observation X is the probability of an **equally or more extreme** observation than X **under the null hypothesis**.



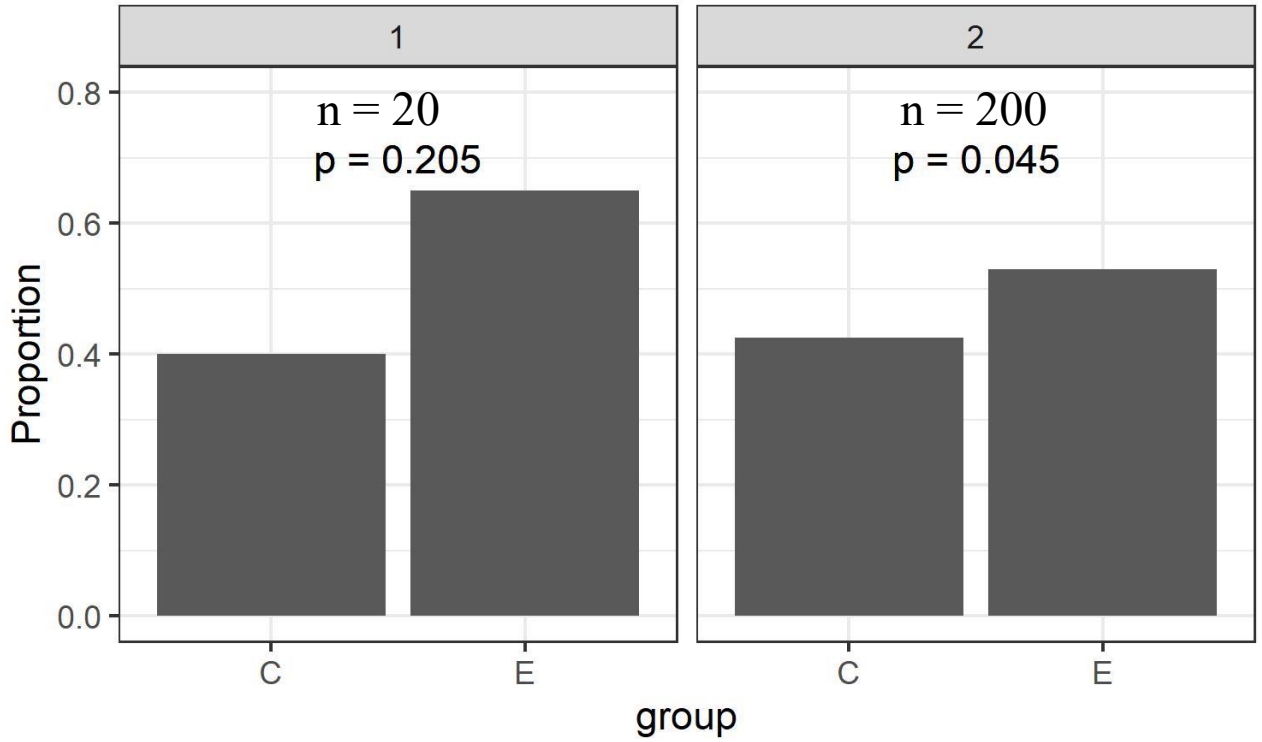
What is the p-value not?

It is **not the probability of H_0 being true.**



What is the p-value not?

It is **not an effect measure.**

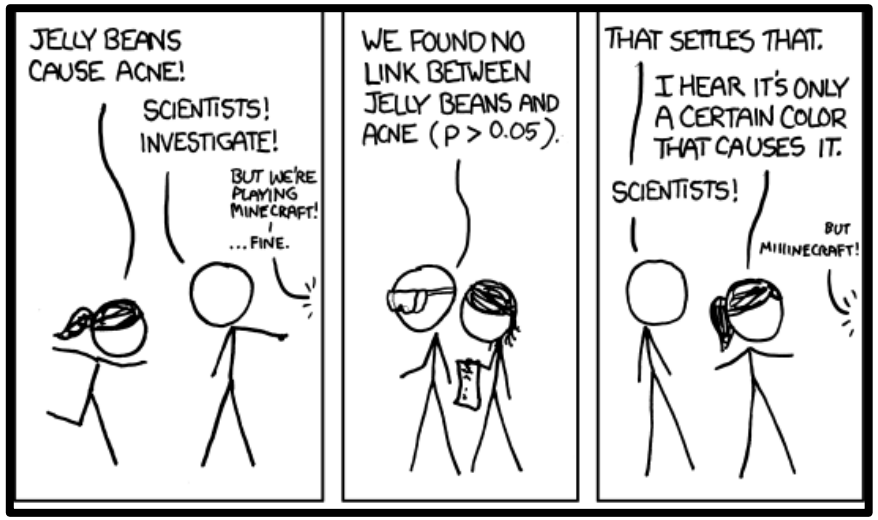


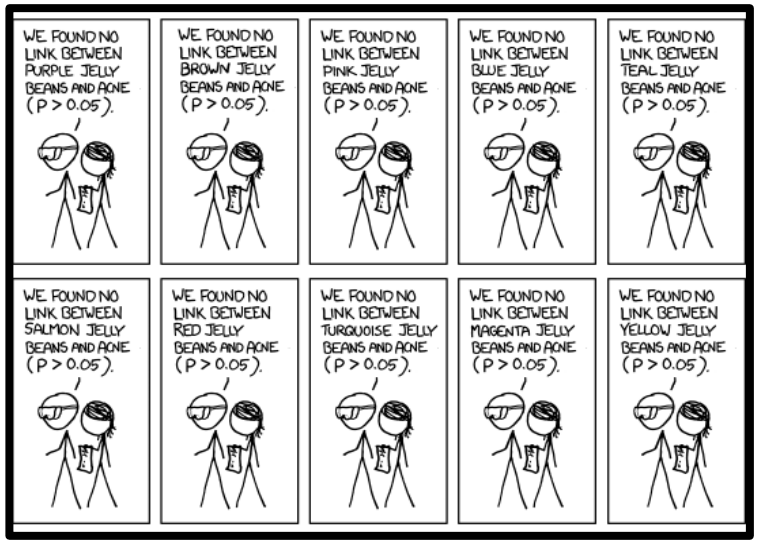
p-hacking

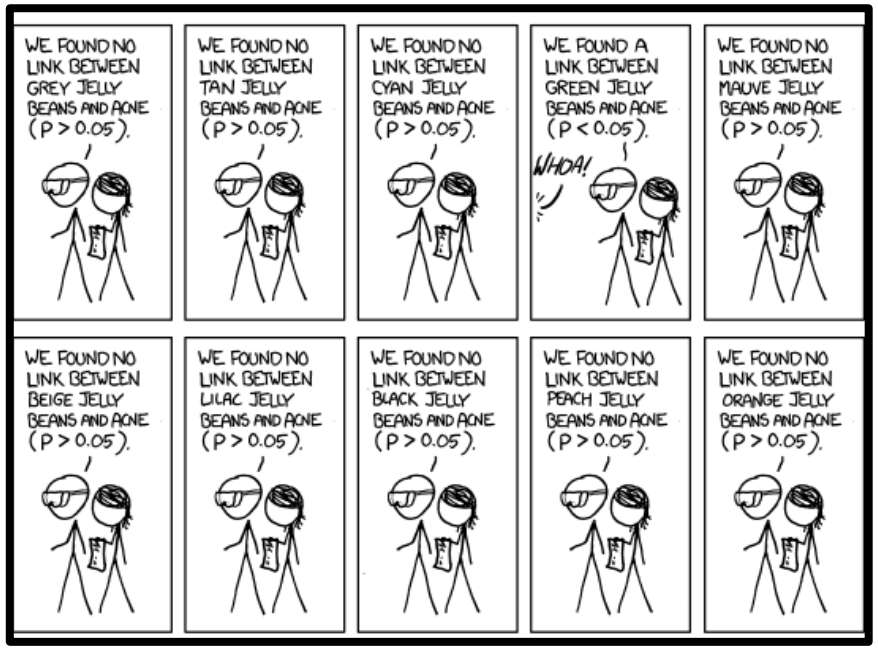
p-hacking denotes post-hoc manipulations with the aim to **fake confirmatory** findings.

It includes techniques like

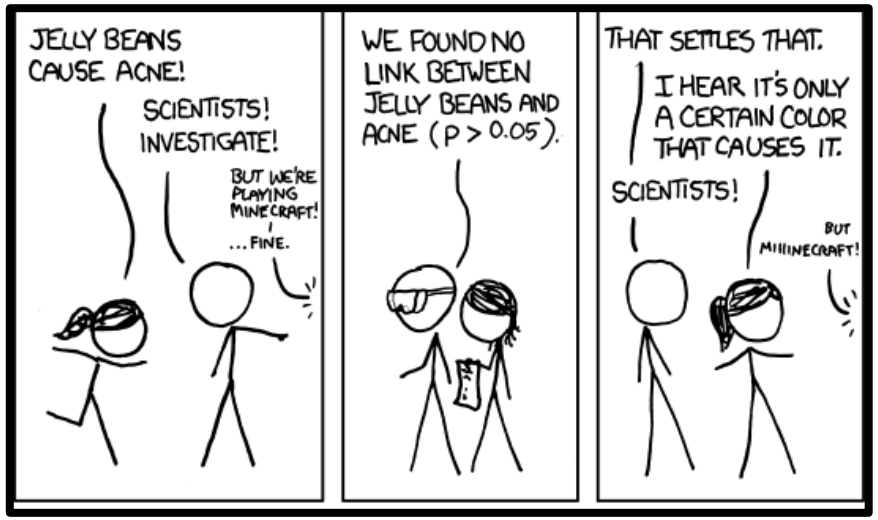
- HARKing (Hypothesizing After Results are Known)
- Unadjusted multiple testing
- Subgrouping
- Many others

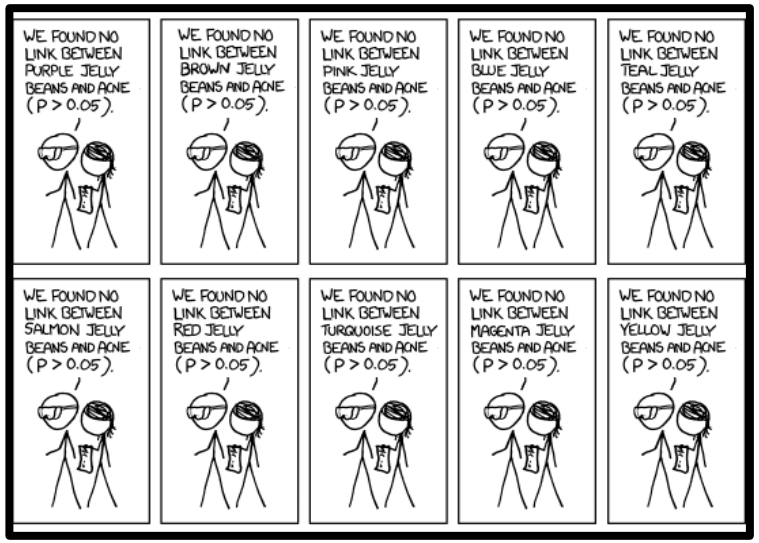


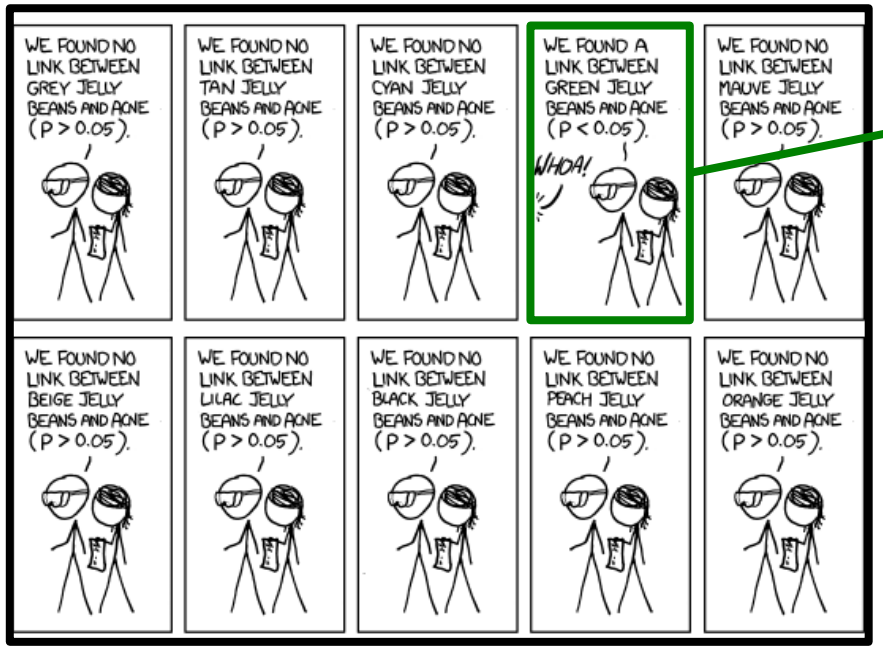


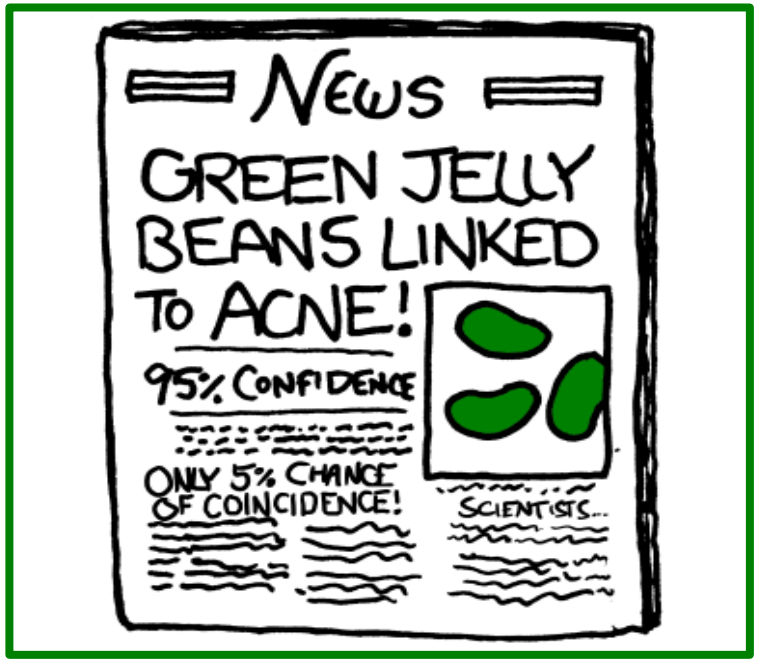


Source:
xkcd.com/882/









Source:
xkcd.com/882/

HARKing

= **H**ypothesizing **A**fter **R**esults are **K**nown takes several forms:

- Actually pretending to have started with the „significant“ hypothesis.
- Hiding the true a priori hypothesis.
- Pretending to have prespecified the analysis.
- Pretending to have planned only the presented analyses (e.g. by only mentioning those in ”methods“-section).
- Using the term ”significant“ inappropriately.

Multiple testing

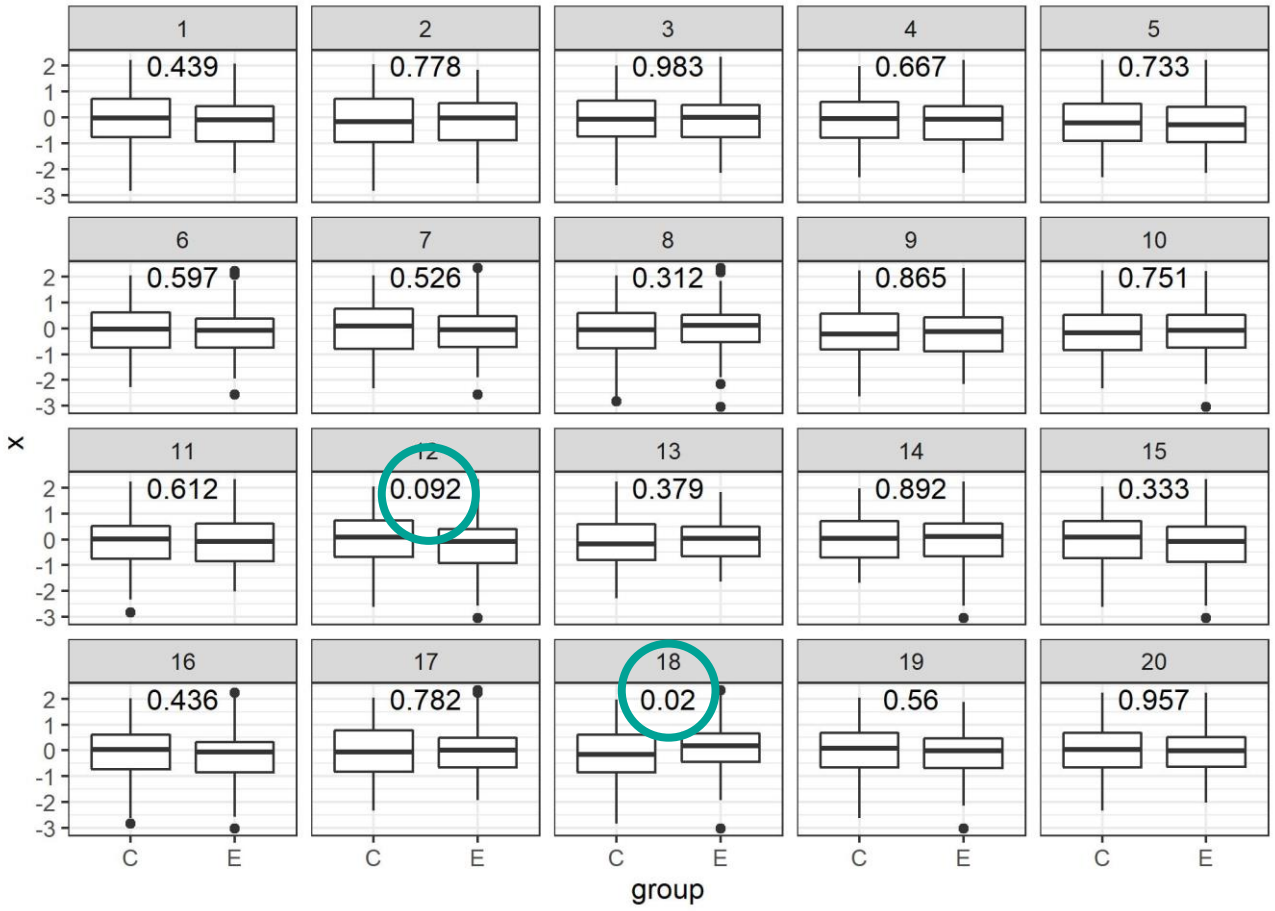
What is the problem?

- 1 in 20 tested (true) hypotheses is falsely significant.
- On average, you just need to perform 20 different tests to get a significant result.

- If multiple tests are carried out and p-values are not corrected for multiple testing, they have to be denoted as descriptive.

- The multiple testing problem can take several forms.

Subgrouping



- Searching for subgroups with an effect leads to a multiple testing problem.
- In some subgroups, there will appear to be an effect by chance only.

Multiple testing – statistical implications

The probability of making any false discovery increases.

Reality / Decision	No effect in any variable	Effect in variable xy
„Significance“ in any variable	$> \alpha$	unknown
„Significance“ in variable xy	α	power 1 - β

not controlled

If no correction for multiple testing is made.

Adjusting for multiple testing

Several methods for adjustment exist.

The easiest one is the Bonferroni correction:

- If you are testing *k* hypotheses, divide all p-values by *k*.

More sophisitcated: Bonferroni-Holm adjustment

Multiple testing – statistical implications

The effect of multiplicity adjustment.

Reality / Decision	No effect in any variable	Effect in variable xy
„Significance“ in any variable	α	unknown
„Significance“ in variable xy	$< \alpha$	$< 1 - \beta$

controlled

decreases

The handling of multiple testing
in an exploratory setting
depends on your balancing of the
different error rates.

Error types

- *Type 1 error rate*
  Probability of falsely rejecting
  one specific hypothesis

  Controlled by significance level.

- *Familywise error rate (FWER)*
  Probability of falsely rejecting
  any of a family of hypotheses

  Controlled by multiplicity adjustment.

- *False discovery rate (FDR)*
  Proportion of false discoveries
  (with regard to all
  discoveries/rejections)

  Controlled by other adjustment methods, e.g. *Benjamini-Hochberg.*

Discussion

What (error) is important to you?

Day 3 - Bring your own data!

Maybe you have encountered a question or problem in your own research that could be tackled with some of the methods learned in this course.

This is why we will have time to discuss some of your real life examples on Day 3 (Samuel Kilian)

→ For this we will need your Data as well as your Research questions

Data

I4

	A	B	C	D	E	F
1	Species_No	Petal_width	Petal_length	Sepal_width	Sepal_length	Species_name
2	2	1	3,5	2	5	Versicolor
3		1	4	2,2	6	Versicolor
4	2	1,5	4,5	2,2	6,2	
5	3	1,5			6	Verginica
6	1	0,3	1,3	2,3		Setosa
7	2	1	3,3	2,3	5	Versicolor
8	2	1,3	4	2,3	5,5	
9	2	1,3	4,4	2,3	6,3	Versicolor
10		1		2,4	4,9	Versicolor
11	2	1	3,7	2,4		Versicolor
12	2	1,1		2,4	5,5	Versicolor
13	2	1,1	3	2,5	5,1	Versicolor
14	2	1,1	3,9	2,5	5,6	Versicolor
15	2	1,3	4	2,5	5,5	Versicolor
16	3	1,7	4,5	2,5	4,9	Verginica
17	2	1,5	4,9	2,5	6,3	Versicolor

- Every column is one variable
- Provide us with .xlsx or .csv formats

Research Question

There are two types of research questions, confirmatory and exploratory ones.

Examples:

- *Confirmatory*: Is there a difference in **means** of the variable body weight between the groups A and B?

- *Exploratory*: Out of the variables A, B, C, D, which one separates the groups U and V the best?

How does it work?

Fill out the provided form

- Specify your research question
- Explain your data as thoroughly as possible
- Clean and attach your data
- Send it all to kilian@imbi.uni-heidelberg.de
- Samuel will choose 1 or 2 examples to discuss in class with you

Thanks for your attention!