Chapter 6:

Evaluation: Inferential Statistics

Overview

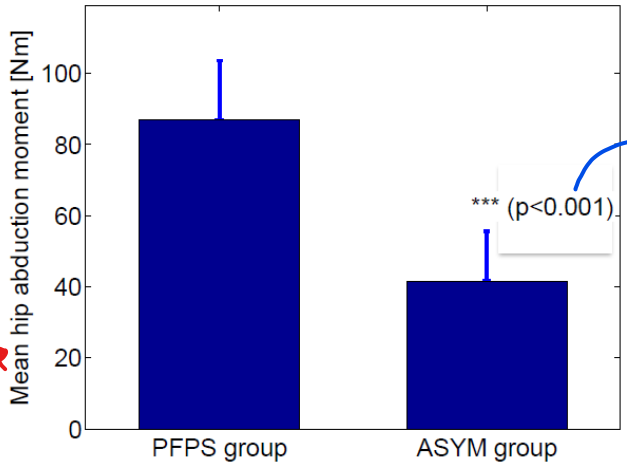
Statistical testing, statistical significance, and errors

During the research process, you do not only want a description of the data. You need to do experiments to do data collection and test specific hypotheses/theories. The goal is to reject or accept the theory. As it is still just a theory, you cannot reprove it but you can show that the theory is accepted as one.

The confidence interval shows whether the results of a study might not give the true mean, or it might not even be within the confidence interval. This already indicates whether your measurement is actually something that is a value or not. However, you might detect effects that do not exist or miss the effects that do exist in your data analytics that's why you need tools to state whether the probability that given the data acquired, our theory/ hypothesis is true/false.

If scores of two samples by different groups/conditions are likely to be not exactly the same, the question is whether the difference is caused by random bias or if actually an effect has been observed. With inferential statistics, you can find out whether the result that you observed is just by chance.

You have 2 different groups of people. One group is asymptomatic (ASYM) and the other one has a condition (PFPS). The graph below shows the mean hip abduction moment.
For the one group, the mean value is visibly different from the other group, but the question is how probable it is, that the effect is just chance (e.g., because not enough people were tested). The *** means that the data has been statistically tested with inferential statistical tools and the p-value shows that the 2 groups are highly significantly different.
The probability that the difference has been observed by chance is quite low.



Statistical significance between two results exists if the probability that the difference is by chance below a certain significance level.

Significance level α:

To do statistical testing, you set a significance level (a level at which you would like to operate). It is somehow arbitrary, but a lower significance level means that you have higher evidence and a lower probability of observing the effect by chance.

Typical levels are: α = 0.05 (*), α =0.01 (**), α = 0.001 (***)

Significant results (p < α)

Significant results mean, that the p-value is smaller than the significance level. When you observe that, you can reject the null hypothesis H_0 and there is a **statistically significant difference** between the results!

Non-Significant results (p ≥ α)

Non-significant results mean, that the p-value is equal or greater than the significance level. You cannot reject the null hypothesis H_0 in this case. However, you cannot conclude that the null hypothesis is true. You basically **cannot conclude anything**!

Error types

Correct predictions:

If the effect that is found in the results is true and the effect really exists, then it is called a true positive. If the effect that was found is false and the effect also does not exist, it is a true negative.

Errors:

There are 2 different types of errors. A false positive or a false negative prediction. A false positive prediction is also called Type I error or α-error. It means, that a non-existing effect was found. A false negative prediction is often called type II error or β-error. Here, the effect exists but was not found.

Effect found in results	An effect exists: False (H_0 is true)	An effect exists: True (H_0 is false)
true	**Type I error** (False Positive) Non-existing effect found (α-error)	**Correct** (True Positive) Existing effect was found
false	**Correct** (True Negative) No effect exists, no effect found	**Type II error** (False Negative) Effect exists but is not found (β-error)

A person gets tested for Covid-19:
True Positive: A Person is a carrier of the virus, and the test predicts that the person is a carrier of the virus.
True Negative: A Person is not a carrier of the virus, and the test predicts that the person is not a carrier of the virus.
False Positive: A person is not a carrier of the virus, and the test predicts that the person is a carrier of the virus.
False Negative: A person is a carrier of the virus, but the test predicts that the person is not a carrier of the virus.

Student t-test

A test that is very often used in statistical testing is the student's t-test. It tests for significant differences in means. If you have 2 uniform distributions, it is tested whether they have a mean that is significantly different from each other. Therefore, a null hypothesis and an alternative hypothesis are defined.

The null hypothesis is defined that the means are equal:

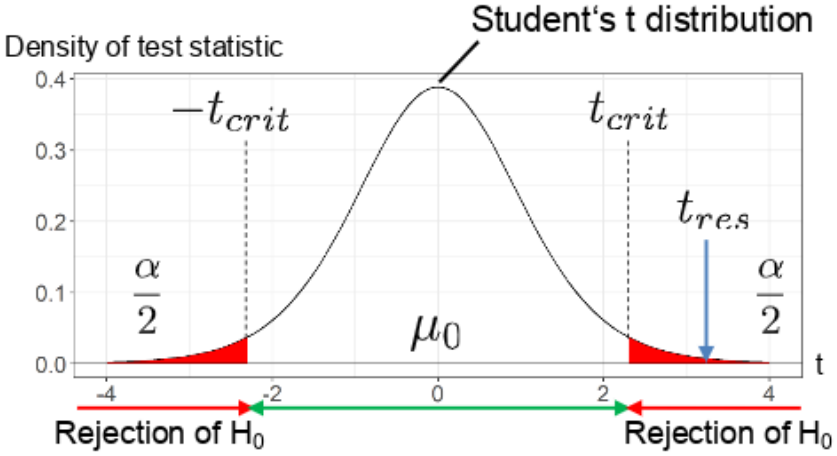
$$H_0: \quad x = \mu_0$$

The alternative hypothesis is defined that the two means are different from each other:

$$H_1: \quad x \neq \mu_0$$

Based on the data set you can then calculate the resulting t-value with the standardized test statistic:

$$t_{res} = \frac{\bar{x} - \mu_0}{\sigma} \sqrt{n}$$



According to the critical t-value that was set, we know whether to reject the null hypothesis or accept it.

$$|t_{res}| > t_{crit} \rightarrow Rejection\ of\ H_o$$

If the absolute value of the resulting t-value is above the critical t-value, then the null hypothesis is getting rejected.

You can then calculate the p-value from the concrete t-value: $p = 2 \cdot P(t > |t_{res}|)$

$$p = 2 * P\ (t > \ |t_{res}|)$$

Any statistical software (R, SPSS, Python, Excel, Matlab ,...) automatically returns the p-value.

Good practice

If you want to conduct a statistical test with the t-test it's a good idea to follow this following general procedure:

1. Design null hypothesis
2. Design alternative hypothesis
3. Decide on α and β and sample size
4. Data acquisition
5. Always report descriptive statistics and CI alongside with the p value

Often you see in papers in statistical test results that people are talking about a "significant difference" when the p-value is lower than 0.05 and a "highly significant difference" when the p-value is lower than 0.01.

Only talk about significant results in any mathematical connection when you actually did a statistical test. If that is not the case, it is better to say considerable difference.

Under the assumption of the null hypothesis H_0, p is the probability that the result happened by chance. If p is smaller than the significance level α the resulting deviation from H_0 could not happen just by chance (there probably is an effect).

- The p-value does **NOT** indicate the size of an effect
- The p-value does **NOT** give the probability of a hypothesis
- The p-value is **NOT** an indicator for replicability

Data dredging

randomness can have significant results.

A problem that is associated with just a pure p-value calculation is that oftentimes researchers need to report a significant effect on their study. Due to the necessity to publish relevant and significant results, researchers often look for a significant effect. That is called p-hacking. The problem with that is, that any data set with any degree of randomness is likely to contain some significant results!

With p-hacking, you discover data patterns exploratively and present results as statistically significant without specific hypotheses. To do this you collect or select the data until nonsignificant results become significant.

The problem is that every dataset with any degree of randomness is likely to contain some significant results. And as we already mentioned in a previous chapter, if our sample size is big enough, the probability to find significant differences increases. **So, always take a look at the effect size! This might be an indicator if p-hacking was used**.

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) **The Extent and Consequences of P-Hacking in Science**. PLoS Biol 13(3): e1002106. https://doi.org/10.1371/journal.pbio.1002106

Bonferroni correction

The statistical methods we presented so far are limited because they can only compare one group to another. But what if we need to compare multiple groups? Performing several t-tests can increase the risk of false positives and this means a reduced statistical power! The probability for at least one false positive decision is accumulated:

$$\alpha' = 1 - (1 - \alpha)^k$$

If you do multiple hypothesis tests in the same data set, you need to prevent the accumulation of the α-value. To do that, you can use the Bonferroni correction. The Bonferroni correction is compensating for the increased likelihood of performing a Type I error.

- Correction decision level: $\alpha' = \frac{\alpha}{m}$
- Rejection of H_0 for each: $p_i < \frac{\alpha}{m}$

α: original α-value
m: number of tested hypotheses

Even though we can compensate the risk of false positives with the Bonferroni in parts, we should avoid using several t-tests, since we have tests like ANOVA that account for that problem by performing only one test.

Analysis of variance (ANOVA)

Analysis of variance.

We use ANOVA if we want to compare more than two means, which you will often face in research. This test is more general than the two-sample t-test. Usually, we apply this test when we have a discrete independent variable, also called factor, and we want to evaluate whether there is a measurable effect of this factor on a continuous dependent variable. This factor can also have different factor levels, representing different conditions.

discrete independent variable "factor" → evaluate effect on continuous dependent variable.

We could use it for example to compare:

- Three different kinds of keyboards
- Four different visualization techniques
- Five different VR scenarios

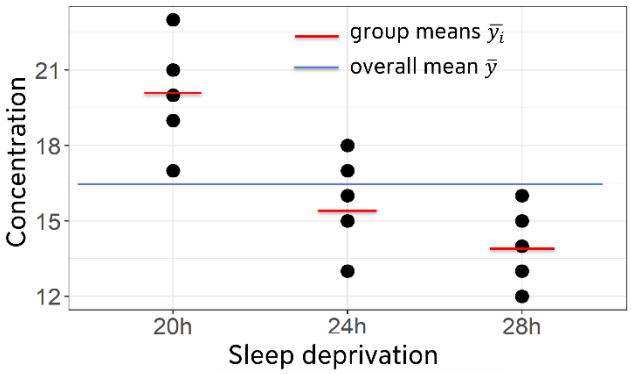
The null hypothesis for the ANOVA is

$$H_0: \overline{x_1} = \overline{x_2} = \ldots = \overline{x_m}$$

With $\bar{x}_i$ representing the mean of the independent variables.

Let us have a closer look at the ANOVA with an example: Imagine, you conducted a study among students, where you tested their concentration and asked them about their sleep deprivation. You split up the total variance into the variance due to group differences and differences within the groups. So you have a number of s groups with n_i samples each:

$$n = \sum n_i$$



To calculate your total variance, first, you need to calculate the sum of squares for your group differences (SQ_{group}) and the sum of squares for the differences within the groups (SQ_{within}):

$$SQ_{total} = SQ_{group} + SQ_{within}$$

$$SQ_{total} = \sum_{i=1}^{s}\sum_{k=1}^{n_i}(y_{ik} - \bar{y})^2$$

y_i → group. unit ($\bar{y}$ = Global mg)

$$SQ_{group} = \sum_{i=1}^{s}(y_i - \bar{y})^2$$ → Group (or) bunch as a

$$SQ_{within} = \sum_{i=1}^{s}\sum_{k=1}^{n_i}(y_{ik} - \bar{y}_i)^2$$

The test statistics then tells us, how much variance is due to group differences as compared to the variance within groups:

$$F = \frac{n-s}{s-1} * \frac{SQ_{group}}{SQ_{within}}$$

If $|F| > F_{crit}$ we reject our null hypothesis H_0. F_{crit} is depending on α, n and s. If you want to make it a bit easier for yourself, you can use statistical software, like R, SPSS, Python, … to calculate the ANOVA. Since the result of ANOVA only tells us, whether a difference is present or not, you might need to conduct follow-up tests to evaluate pairwise differences for example with a t-test.

ANOVA procedure

1. Define hypotheses
2. Define decision rule (α)
3. Collect your data
4. Perform ANOVA
5. If(!) ANOVA is significant: Perform multiple comparisons using t-tests to detect specific group differences → **This step depends on the defined hypotheses!**

Classification of tests

When you want to statistically analyze your data and are looking for a statistic test you can base your decision on the following aspects:

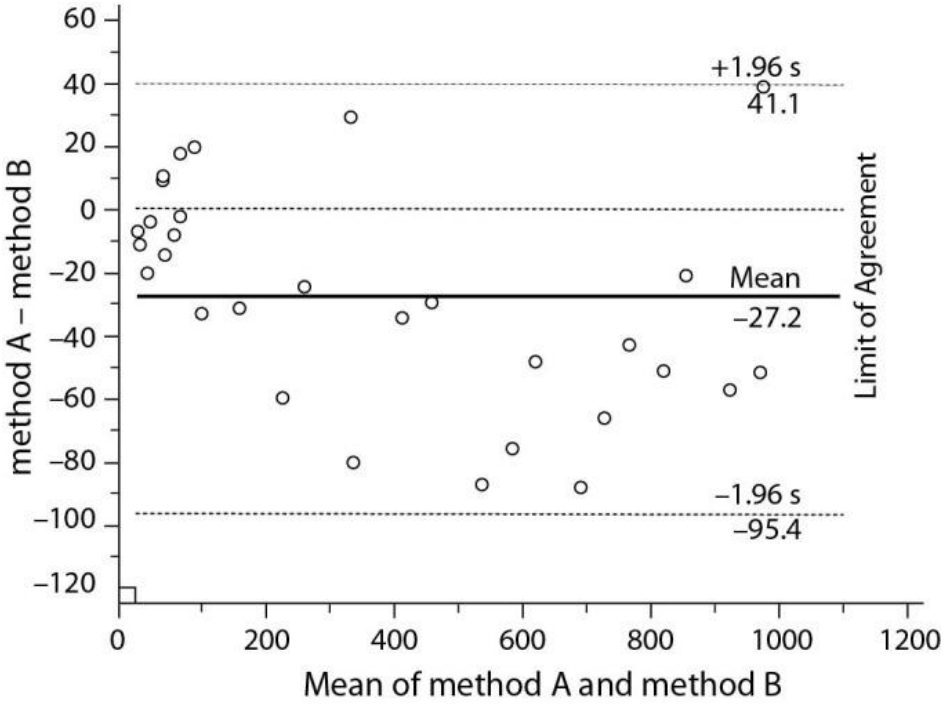
- Number of groups
- Distribution of data: Parametric / non-parametric tests
- Type of variables
- Hypotheses: Comparison of means / variances
- One-sided / two-sided tests
- Experimental method

Different statistics test books or guidelines can help you additionally to find the correct statistical test.

https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/

Bland-Altman plot

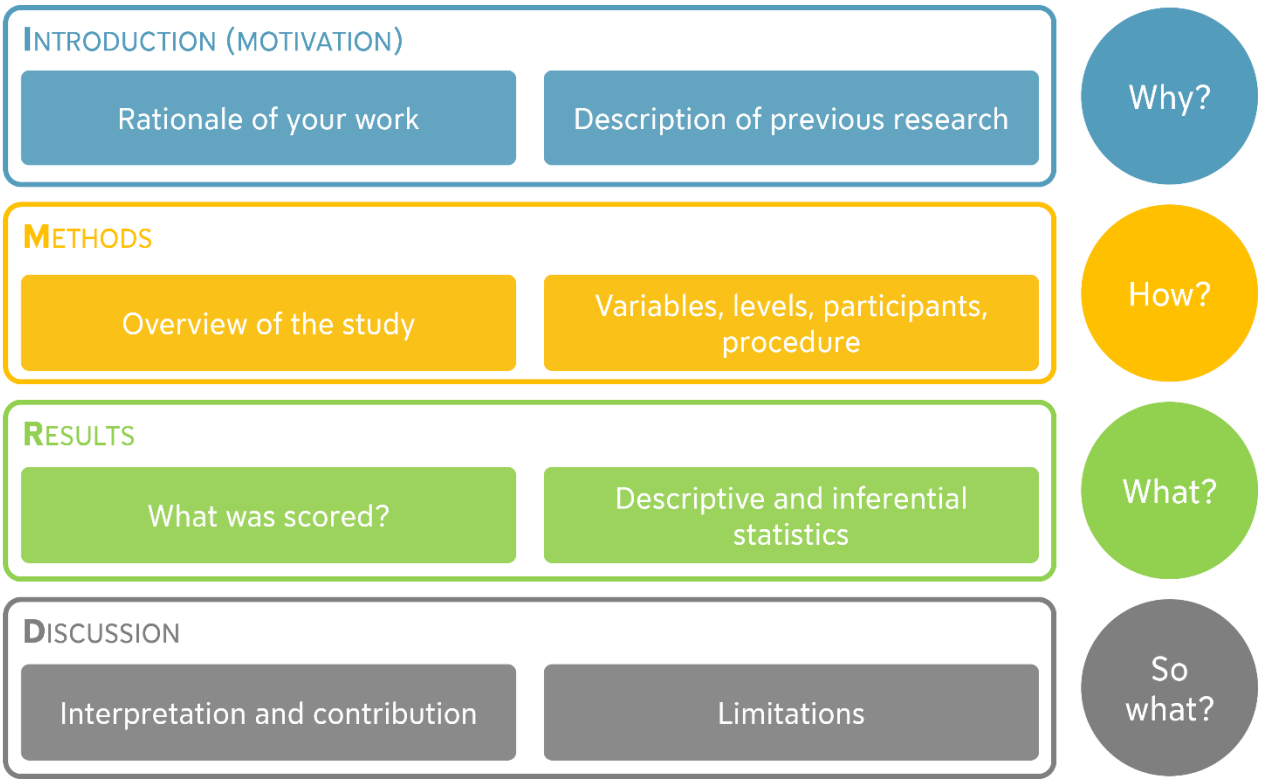
This plot can be used when you want to compare two different measuring methods graphically. In clinical research this is often used to compare new intervention methods against a clinical gold standard. To generate a Bland-Altman plot you need two series of measurements, one for each method. In a special scatter plot, the differences of the two measurement methods or alternatively the ratio against the mean of the two methods are plotted, as the example plot below shows.



Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb). 2015;25(2):141-151. Published 2015 Jun 5. doi:10.11613/BM.2015.015

Reporting study results

An important part of research is not only analysing the data, but also making them public to the scientific community. When it comes to writing a scientific paper you usually follow the IMRaD structure: the main body consists of an **I**ntroduction, the **M**ethods you used, the **R**esults you obtained **a**nd a **D**iscussion. In these parts you basically answer 4 Questions:



These parts are in the end framed by the title and an abstract in the beginning and a Conclusion, your References, and optional appendices in the end.

Summary

- Inferential statistics allow us to compare two (t-test,...) or more samples (ANOVA, ...) to show if our manipulation had an effect
- Always take a closer look when interpreting the p value and keep in mind our guidance on good practice
- Inferential statistics tests
  - Calculate the probability that two samples are from the same population
  - If $p < \alpha$ (typically 0.05), we conclude there is a significant difference
- Examples: t-test and ANOVA

References

1 Alan Dix, Janet Finlay, Gregory Abowd and Russell Beale. (1998) Human Computer, Interaction (second edition), Prentice Hall, ISBN 0132398648 (new Edition announced for October 2003)

2 Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The Extent and Consequences of P-Hacking in Science. PLoS Biol 13(3): e1002106. https://doi.org/10.1371/journal.pbio.1002106

3 Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb). 2015;25(2):141-151. Published 2015 Jun 5. doi:10.11613/BM.2015.015

4 Schiefer, H., & Schiefer, F. (2021). Statistics for Engineers. https://doi.org/10.1007/978-3-658-32397-4

5 https://medium.com/nerd-for-tech/p-hacking-explained-45d4980abf11