



Analysis of variance (ANOVA)

Sebastian Jentschke



In today's lecture, I would like to give an introduction into the analysis of variance.



Overview

- introduction
- history and some mathematical background
- ANOVA with one factor in jamovi
- ANOVA with more than one factor in jamovi
- ANCOVA in jamovi
- ANOVA for repeated-measurements in jamovi
- for all: assumption checks (normality and variance homogeneity), effect sizes, and post-hoc tests



We will begin the lecture with contextualizing the ANalysis Of VAriance (ANOVA), using the distinction of categorical and continuous variables that I used before.

Afterward, we will turn to a bit of history of the ANOVA plus a step-by-step introduction into how an ANOVA is calculated manually. Even though it is a bit tedious, I highly recommend that you follow this introduction since it makes it easier to understand the principles (there is an accompanying spreadsheet with the calculations).

Afterwards, we will turn to how different kinds of ANOVAs are calculated in jamovi. We will begin with a simple ANOVA with one factor, and then extend it into an ANOVA with two factors (and their interaction).



Overview

- introduction
- history and some mathematical background
- ANOVA with one factor in jamovi
- ANOVA with more than one factor in jamovi
- ANCOVA in jamovi
- ANOVA for repeated-measurements in jamovi
- for all: assumption checks (normality and variance homogeneity), effect sizes, and post-hoc tests



Afterwards we will introduce the analysis of covariance, where we control for a continuous variable that may represent a nuisance variable (a typical example is age with e.g., accumulating knowledge or reaction times getting slower). The point is to remove the influence of that variable from our model.

Finally, we will turn to the ANOVA for repeated measurements. Prime examples where it is used are for comparing different conditions that were administered to one participant in the course of an experiment. Another common use case is an intervention with a pre-, post- (immediate after the intervention) and follow-up-measurements (some time after the intervention; for checking the stability of the effect).

There are three common, reoccurring themes that apply to all ANOVAs introduced here: Assumption checks, typically normality and homogeneity of variances, effect sizes, and post-hoc tests.



Introduction

In the very brief first part I would like to embed the analysis of variance within the same framework of categorical vs. continuous predictor / independent and outcome / dependent variables.

Related to this distinction is the one between difference and relation hypotheses.

Finally, for the ANOVA, we often encounter within-subject designs (e.g., when following a participant over time or when administering different conditions within an experiment to the same participant).



Categorical vs. continuous vars.

- categorical variables contain a limited number of steps (e.g., male – female, experimentally manipulated or not, level of education)
- continuous variables have a (theoretically unlimited) number of steps (e.g., body height, weight, IQ)
- ANOVA (this session) is for categorical predictors, Correlation and regression analyses (lecture given two weeks ago) is for continuous predictors



I will (yet another time) use the distinction between categorical and continuous variables to contextualize the analysis of variance.

Categorical variables encompass the variables from the two measurement levels nominal and ordinal (for an more extensive overview on measurement levels, see the crash course).

Continuous variables encompass the two variable levels interval and ratio.

For ANOVAs we have (mainly) categorical predictor (independent) variables and continuous outcome (dependent) variables. I used “mainly”, since it is possible to control for the influence of continuous predictors in the Analysis of Covariance (as special “flavour” of the ANOVA). For correlation and regression analyses we use exclusively continuous predictors.



Categorical vs. continuous vars.

		Dependent variable	
		Categorical	Continuous
Independent variable	Categorical	X ² test (chi-squared)	t-test ANOVA
	Continuous	Logistic regression	Correlation Linear regression



Most classes of multivariate methods (ANOVA as well as linear and logistic regression) are based upon the General Linear Model. ANOVA is covered quite extensively in this lecture, linear regression was in the previous lecture.

Both ANOVAs and regression analyses stand quite central in our methods repertoire. ANOVAs are typically used to analyse data from experiments (where we manipulated one or more factors, representing the categorical variables), whereas linear or logistic regression models are often used to analyse data acquired with questionnaires. They differ in that linear regression has a continuous variable as dependent variable (e.g., quality-of-life, job satisfaction), for logistic regression it is categorical (e.g., clinical vs. control group).

X² tests were introduced in the refresher session.



Difference vs. relation hypotheses

- difference hypotheses explore whether there is a difference between the steps of one (or more) independent and a dependent variable
- relation hypotheses explore whether there is a relation between one (or more) independent and a dependent variable
- → causality can only be inferred if the independent variable was manipulated before the outcome was measured



In the lecture on linear regression, we were dealing with hypotheses regarding relationships between variables and how they could be used to predict a certain behaviour of a person based upon those relations. For example, we could try to predict performance in the job based on job satisfaction, support from superiors, etc.

In today's lecture we turn to a method where we are most interested in evaluating differences. Typically, we use this method to analyze data from experiments. Here we are most interested to see whether an experimental manipulation gave rise to an effect that is statistically significant.

Given that we typically manipulate one or more independent variables *before* we measured the outcome when using experimental designs, such designs generally allow us to make claims about causality.



Within vs. between subject vars.

- within-subject variables are measures acquired from the same person (e.g., administering the same test before and after treatment; different experimental conditions; EEG / MRI data)
→ idea that the “performance” or “properties” that characterize the person stay the same
- between-subjects variables are variables that distinguish between individuals (e.g, male-female)



Another distinction that is important is whether one or several measurements per participant are assessed using the ANOVA.

For the three “classical” between-subject-types of ANOVA (ANOVA with one factor, ANOVA with several factors, ANCOVA), we are dealing with one dependent variable per participant (between-subject variables).

In contrast, what constitutes the dependent variables for the within-subject type of ANOVAs (for repeated measurements) are several measurements that are arranged into factors (e.g., pre, post, follow-up).

Such approach can reduce the amount of the variation we can't explain with our model by controlling for individual variation. This is done by considering the levels of the repeated-measurement factors as deviations from the mean of each individual. Thereby we can control for individual differences.



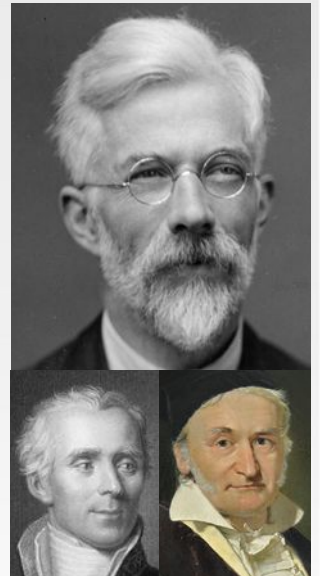
Principles and background

Let's start our journey with some definitions and history before learning a bit about the mathematical background of the Analysis of Variance.

Some history, definition

- introduced by Sir Ronald Fisher in 1921 based upon earlier ideas of Laplace and Gauss
- compare two (or more) means to see whether they differ from another
- evaluates the differences among means relative to the dispersion of the sampling distribution

$$H_0: Y_1 = Y_2 = \dots = Y_k \quad (\mu_1 = \mu_2 = \dots = \mu_k)$$

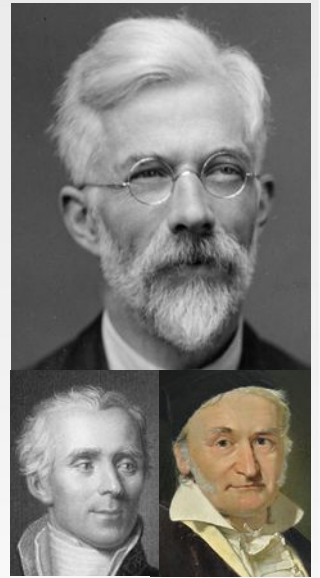


“The analysis of variance”, usually referred to as ANOVA, was first publicized as a method by Sir Ronald Fisher (top) in 1921. That said, there were antecedents that introduced concepts such as hypothesis testing, the partitioning of sums of squares, experimental techniques and the additive model as early as in the 18th century. For example performed Laplace (bottom left) hypothesis testing in the 1770s, and Laplace and Gauss (bottom right) developed the least-squares methods around 1800.

Some history, definition

- introduced by Sir Ronald Fisher in 1921 based upon earlier ideas of Laplace and Gauss
- compare two (or more) means to see whether they differ from another
- evaluates the differences among means relative to the dispersion of the sampling distribution

$$H_0: Y_1 = Y_2 = \dots = Y_k \quad (\mu_1 = \mu_2 = \dots = \mu_k)$$



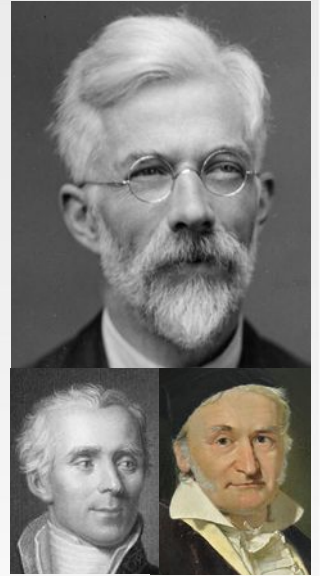
ANOVA is a form of statistical hypothesis testing heavily used in the analysis of experimental data. A test result (calculated from the null hypothesis and the sample) is called statistically significant if it is deemed unlikely to have occurred by chance if the null hypothesis were true.

In the typical application of an ANOVA, the null hypothesis is that all groups ($\mu_1 \dots \mu_k$) are random samples from the same population. For example, when studying the effect of different treatments on similar samples of patients, the null hypothesis would be that all treatments have no effect and therefore there is no difference between groups. Rejecting the null hypothesis is taken to mean that the differences in observed effects between treatment groups are unlikely to be due to random chance.

Some history, definition

- introduced by Sir Ronald Fisher in 1921 based upon earlier ideas of Laplace and Gauss
- compare two (or more) means to see whether they differ from another
- evaluates the differences among means relative to the dispersion of the sampling distribution

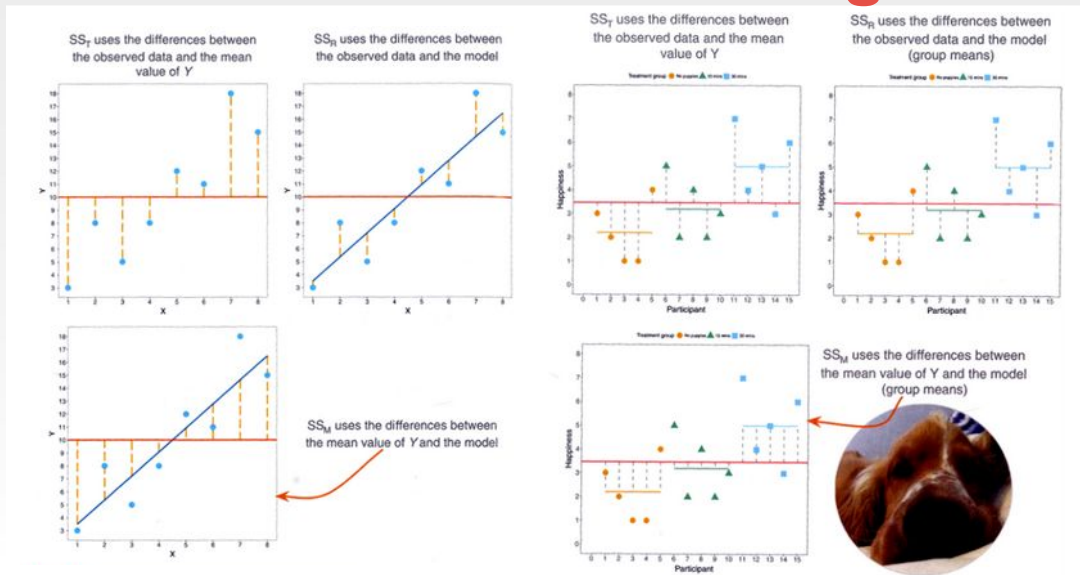
$$H_0: Y_1 = Y_2 = \dots = Y_k \quad (\mu_1 = \mu_2 = \dots = \mu_k)$$



What distinguishes the ANOVA from a t-test is that it can assess more than two groups at once (whereas one can't do more than one-by-one comparisons with a t-test) and it is also possible to assess more complex hypotheses encompassing several factors.



Some mathematical background



As I mentioned on one of the introductory slides, are both regression and the analysis of variance (ANOVA) based upon the same underlying model, the General Linear Model. In the case of a regression (left), we try to fit a regression line that represents how much change in the dependent variable goes along with a certain change in the independent variable (slope). The slope is adjusted such that the squared deviations (distance of the blue dots from the regression line) are minimized.

For the ANOVA (right), we use the means of the different groups in the model as predictors. Using those means “automatically” minimizes the squared deviations (as those deviations are always distances from the mean).

We then compare those squared deviations (as an indicator of what our model can't predict) to the variance that can be predicted by our means or the regression line



Some mathematical background

Two ways of calculating the variance:

$$\text{Var}(Y) = \frac{1}{N} \cdot \sum_{p=1}^N (Y_p - \bar{Y})^2 \quad \text{equals} \quad \text{Var}(Y) = \frac{1}{N} \cdot \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

Splitting the sum of squares:

$$SS_{\text{tot}} = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2 = SS_b = \sum_{k=1}^G N_k \cdot (\bar{Y}_k - \bar{Y})^2 + SS_w = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$$

$$\text{df:} \quad N - 1 \qquad G - 1 \qquad N - G$$

$$F = \frac{SS_b / df_b}{SS_w / df_w}$$



We start with an example data set (Clinicaltrial.omv) for some demonstration of the mathematical principles behind an ANOVA. The dataset contains two factors. The first factor (*drug*) compares the effect of a new antidepressant drug called Joyzepam to an existing drug called Anxifree and a placebo. The second factor (*therapy*) assesses whether the drug effects are modulated by providing CBT at the same time (vs. receiving no treatment) in combination with therapy (no vs. CBT). We will focus on the first factor (*drug*) for our demonstration. We start with the null hypothesis that all drugs have the same effect – $H_0: \mu_P = \mu_A = \mu_J$. Our alternative hypothesis says the opposite, the three levels of drugs differ with respect to their means – $H_1: \text{not } \mu_P = \mu_A = \mu_J$, maybe easier to understand in the form: $\mu_P \neq \mu_A$ OR $\mu_P \neq \mu_J$ OR $\mu_A \neq \mu_J$. “Analysis of variance” indicates that we are operating with variances when assessing these differences.



Some mathematical background

Two ways of calculating the variance:

$$\text{Var}(Y) = \frac{1}{N} \cdot \sum_{p=1}^N (Y_p - \bar{Y})^2 \quad \text{equals} \quad \text{Var}(Y) = \frac{1}{N} \cdot \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

Splitting the sum of squares:

$$SS_{tot} = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2 = SS_b = \sum_{k=1}^G N_k \cdot (\bar{Y}_k - \bar{Y})^2 + SS_w = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$$

df: $N - 1$ $G - 1$ $N - G$

$$F = \frac{SS_b / df_b}{SS_w / df_w}$$



Typically, the variance is obtained by calculating the squared distance of the score of each individual to the mean of the whole sample $(Y_p - \bar{Y})^2$. Although for some participants (with scores below the mean), the difference is negative, by squaring it, all values become positive. These squared distances are then summed them up over all participants.

We now change this formula slightly by assigning each participant to one of G levels of the factor that we would like to evaluate. For our factor *drug*, the first group ($k = 1$) would be those treated with Anxi-free, the second group ($k = 2$) would receive Joyzepam and the third group ($k = 3$) the placebo.

Within each group, the participants run from $i = 1$ to the number of participants in that group (N_k). For all participants, the squared differences are summed up, first within a group (the sum sign with the index i), later these sums are further summed up over groups (the sum sign with the index k).



Some mathematical background

Two ways of calculating the variance:

$$\text{Var}(Y) = \frac{1}{N} \cdot \sum_{p=1}^N (Y_p - \bar{Y})^2 \quad \text{equals} \quad \text{Var}(Y) = \frac{1}{N} \cdot \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

Splitting the sum of squares:

$$SS_{\text{tot}} = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2 = SS_b = \sum_{k=1}^G N_k \cdot (\bar{Y}_k - \bar{Y})^2 + SS_w = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$$

$$\text{df:} \quad N - 1 \qquad G - 1 \qquad N - G$$

$$F = \frac{SS_b / df_b}{SS_w / df_w}$$



What we just calculated is the variance. Variance denotes the average of the squared distance of each participant from the sample mean. It results from summing up the squared distances and then dividing it by the number of participants ($1 / N$).

For the ANOVA, we operate with a concept which is called sum of squares. It is very similar to the variance, except from that we just sum up the values (and not divide them by the number of participants). The variance is therefore equivalent to the whole variation denoted as total sum of squares (SS_{tot}).

This total sum of squares can now be “split” into the variation which is due to the group membership (called between groups – SS_b – because it “compares” the means of the different groups to the mean of the whole sample) and the variation within the group (SS_w – comparing the distance of each group member from the group mean).



Some mathematical background

Two ways of calculating the variance:

$$\text{Var}(Y) = \frac{1}{N} \cdot \sum_{p=1}^N (Y_p - \bar{Y})^2 \quad \text{equals} \quad \text{Var}(Y) = \frac{1}{N} \cdot \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

Splitting the sum of squares:

$$SS_{\text{tot}} = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2 = SS_b = \sum_{k=1}^G N_k \cdot (\bar{Y}_k - \bar{Y})^2 + SS_w = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$$

df: $N - 1$

$G - 1$

$N - G$

$$F = \frac{SS_b / df_b}{SS_w / df_w}$$



We assume that if the groups differ recognizable, SS_b would be relatively large in comparison to SS_w . Another way of describing the relation between SS_b and SS_w is that SS_b is the part of the variation that we can explain by our drugs having different effects on mood (leading to different means for each drug). In contrast is SS_w the part of the variation that is left over and that we can't explain ("the remaining variation within the group").



Some mathematical background

Two ways of calculating the variance:

$$\text{Var}(Y) = \frac{1}{N} \cdot \sum_{p=1}^N (Y_p - \bar{Y})^2 \quad \text{equals} \quad \text{Var}(Y) = \frac{1}{N} \cdot \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

Splitting the sum of squares:

$$SS_{\text{tot}} = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2 = SS_b = \sum_{k=1}^G N_k \cdot (\bar{Y}_k - \bar{Y})^2 + SS_w = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$$

$$\text{df: } N - 1$$

$$G - 1$$

$$N - G$$

$$F = \frac{SS_b / df_b}{SS_w / df_w}$$



To calculate the F-ratio (explained two slides ahead), we set these sums of squares into relation to (divide them by) their degrees of freedom.

Those are $N - 1$ for the total sum of squares, $G - 1$ for the sum of squares between groups and $N - G$ for the sum of squares within groups.

The degrees of freedom follow a certain rationale:

We start with that each participant in our sample provides a data point, i.e., one little bit of “freedom” that contributes to the variation.

In the whole sample, we “fixed” one parameter, the sample mean (\bar{Y}). This one parameter is our best estimate to describe the characteristics in our sample. This makes sense since all participants have – on average – the same distance from the mean. Therefore, in our whole sample, we end up with $N - 1$ degrees of freedom.



Some mathematical background

Two ways of calculating the variance:

$$\text{Var}(Y) = \frac{1}{N} \cdot \sum_{p=1}^N (Y_p - \bar{Y})^2 \quad \text{equals} \quad \text{Var}(Y) = \frac{1}{N} \cdot \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

Splitting the sum of squares:

$$SS_{\text{tot}} = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2 = SS_b = \sum_{k=1}^G N_k \cdot (\bar{Y}_k - \bar{Y})^2 + SS_w = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$$

$$\text{df: } N - 1$$

$$G - 1$$

$$N - G$$

$$F = \frac{SS_b / df_b}{SS_w / df_w}$$



For our group means and the sum of squares between groups. If we have G different means we consider, so can we describe them most “economically” as distance, e.g., from group 1 to group 2, group 1 to group 3, and so on. For 3 groups we need at least two such distances to describe the relation of the group means. More generally, we need $G - 1$ degrees of freedom.

Finally, we have the sum of squares within groups.

We start again with the number of our participants (since those provide the variation). However, we already have “used” two classes of means to describe our sample: (1) the mean of the whole sample ($df = 1$), and (2) the distances of means between our groups ($df = G - 1$). Therefore, our degrees of freedom are $df = N - (G - 1) - 1$. With resolving the parentheses, the $- 1$ inside the parentheses becomes positive: $df = N - G + 1 - 1 = N - G$ (as $+ 1$ and $- 1$ cancel each other out).



Some mathematical background

Two ways of calculating the variance:

$$\text{Var}(Y) = \frac{1}{N} \cdot \sum_{p=1}^N (Y_p - \bar{Y})^2 \quad \text{equals} \quad \text{Var}(Y) = \frac{1}{N} \cdot \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

Splitting the sum of squares:

$$SS_{\text{tot}} = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2 = SS_b = \sum_{k=1}^G N_k \cdot (\bar{Y}_k - \bar{Y})^2 + SS_w = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$$

$$\text{df:} \quad N - 1 \qquad G - 1 \qquad N - G$$

$$F = \frac{SS_b / df_b}{SS_w / df_w}$$



Now we are ready for our final bit, the F-ratio. To calculate it we set the sum of squares between groups divided by the degrees of freedom between groups ($G - 1$) into relation to the sum of squares within groups divided by the degrees of freedom within groups ($N - G$; set into relation means that we divide them through each other). Remember that the between groups variation was the part we can explain, the within groups variation that part we can't explain.



Some mathematical background

Analysis of variance in one glance:

	df	sum of squares	mean squares	F-statistic	p-value
between groups	$df_b = G - 1$	$SS_b = \sum_{k=1}^G N_k (\bar{Y}_k - \bar{Y})^2$	$MS_b = \frac{SS_b}{df_b}$	$F = \frac{MS_b}{MS_w}$	[complicated]
within groups	$df_w = N - G$	$SS_w = \sum_{k=1}^G \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$	$MS_w = \frac{SS_w}{df_w}$	-	-

see *Clinicaltrial – Step-by-step.xlsx* on MittUIB



On the current slide, all required pieces of information for calculating an ANOVA are collected. For the p-value denoted as “complicated” there exists an Excel-function as well as the opportunity to look the value up from a table with critical F-values.

For demonstrating the calculations, I created an Excel-file (*Clinicaltrial – Step-by-step.xlsx*; sheet “One-way ANOVA”) together with a video explaining it. Both can be found on MittUiB.



ANOVA with one factor in jamovi

For most of the following analyses we will use the clinicaltrials-data set that we already used for our step-by-step demonstration on the previous slides. It contains two predictor variables *drug* and *therapy* and one outcome variable *mood.gain*, and assesses which effect different pharmacological (*drug*) and psychological interventions (*therapy*) have on mood (*mood.gain*).



Equivalence of t-test and F-test

Independent Samples T-Test

Dependent Variables
mood.gain

Grouping Variable
therapy

Independent Samples T-Test

Independent Samples T-Test

	Statistic	df	p
mood.gain Student's t	1.307	16.000	0.2098

ANOVA

Dependent Variable
mood.gain

Fixed Factors
therapy

ANOVA

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p
therapy	0.467	1	0.467	1.708	0.2098
Residuals	4.378	16	0.274		

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 23

Before we really get into ANOVAs in jamovi, I would like to start with a brief demonstration of the equivalence of the t-test and the F-test. We start with clicking on “t-test” in the icon bar, selecting “Independent Samples t-test”, and assign *mood.gain* to “Dependent Variables” and *therapy* to “Grouping Variables”.

We continue by clicking on “ANOVA” in the icon bar, select “ANOVA”, assign *mood.gain* to “Dependent Variable” and *therapy* to “Fixed Factors”.

What we can see is that the p-values are identical to the fourth decimal. What we also can see is that t-value (1.307) and F-value (1.708) are related by a square (from t to F; $1.307^2 = 1.708$) or square root (from F to t; $\sqrt{1.708} = 1.307$) relation. Why this is the case becomes clear if we consider what t- and F-statistic are based upon: t is derived from the standard deviation (s), F from the variance (s^2). Standard deviations squared give the variance.



ANOVA with one factor in jamovi

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	η^2	ω^2
drug	3.453	2	1.727	18.611	<.0001	0.713	0.662
Residuals	1.392	15	0.093				

Homogeneity of Variances Test (Levene's)

F	df1	df2	p
1.450	2	15	0.2657

Normality Test (Shapiro-Wilk)

Statistic	p
0.960	0.6053

Standardized Residuals vs Theoretical Quantiles plot showing a linear trend.



However, when the number of levels of a factor we wish to assess is higher than two or if we have more than one factor, we have to use the ANOVA. We first begin with «replicating» the analysis we conducted «by hand» using LibreOffice Calc (Excel).

For doing this, we create a new ANOVA using the button «ANOVA» in the icon bar and assign *mood.gain* to «Dependent Variable» and *drug* to «Fixed factors». Furthermore, we tick η^2 and ω^2 under «Effect sizes».

Before we have a proper look at our results, it is wise to check whether all assumptions for running an ANOVA are fulfilled. We need to consider independence, normality, and homogeneity of variances. We therefore tick «Homogeneity test», «Normality test», and «Q-Q plot» from the drop-down-menu «Assumption checks».



ANOVA with one factor in jamovi

ANOVA

Dependent Variable: mood.gain

Fixed Factors: drug

Model Fit: Overall model test

Effect Size: η^2 partial η^2 ω^2

Assumption Checks:

Homogeneity test

Normality test

Q-Q Plot

+ independence

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	η^2	ω^2
drug	3.453	2	1.727	18.611	<.0001	0.713	0.662
Residuals	1.392	15	0.093				

Homogeneity of Variances Test (Levene's)

F	df1	df2	p
1.450	2	15	0.2657

Normality Test (Shapiro-Wilk)

Statistic	p
0.960	0.6053



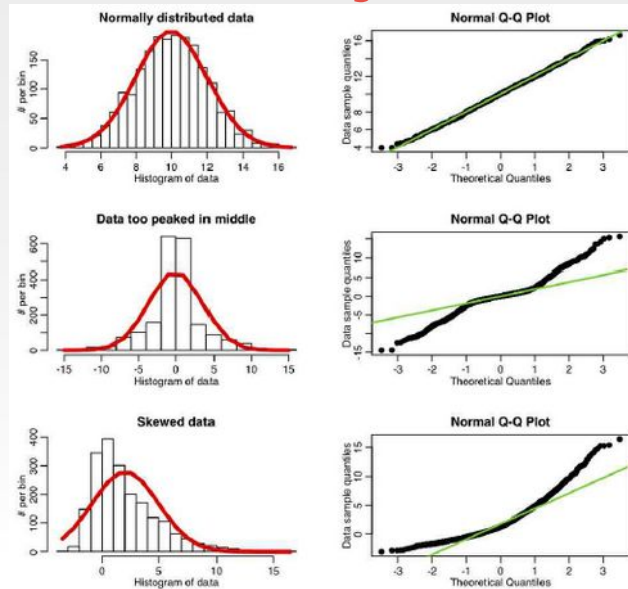
The ***independence*** assumption can't be tested but has to be ensured via experimental design. It demands that all measurements in our sample have been collected without any "relationship to" any of the other ones. Situations with clear violations of this demand, are e.g., such with experimenter effects: Let's assume that an experimenter had a certain behaviour or showed certain expectancies when collecting the data to which the participants reacted. As a consequence, the data collected by that experimenter would be related and would differ from data collected by another experimenter.

Another situation where the independence assumption is violated, this time by decision, is for repeated-measures designs, where each participant in appears in more than one condition. In such case, we need to use a special class of ANOVA, called repeated measures ANOVA (covered later in this lecture).



ANOVA with one factor in jamovi

- **assumptions I:**
normality and
possible causes for
normality violations



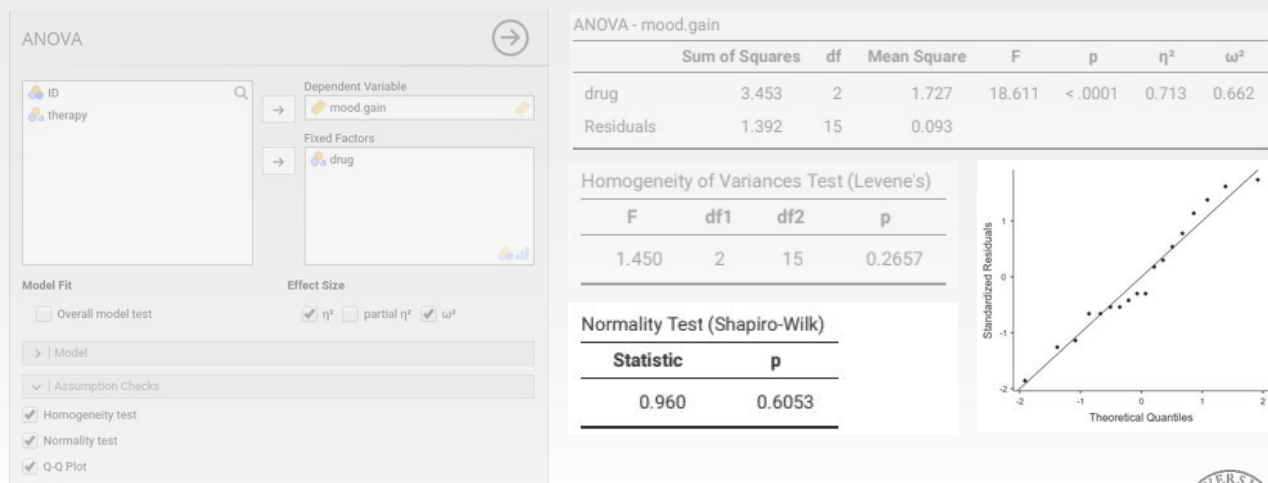
PAGE 26



Normality is a key requirement of any statistical analyses since any of the distributions that we use to test our models (t, F, etc.) rely on the assumption that the data that we use in our models follow a normal distribution.



ANOVA with one factor in jamovi



To assess **normality** it is also sensible to carry out a further analysis **before** our ANOVA. Using «Descriptives», it should be checked whether our dependent variable *mood.gain* is normally distributed. Use the Shapiro-Wilk test and the Q-Q-plot for assessing that.

Within our ANOVA, we can assess the table «Normality test (Shapiro-Wilk)» and the Q-Q plot. These tests give no indication for concern either that residuals may deviate from a normal distribution: The Shapiro-Wilk test is not significant ($p = 0.605$), and the points in the Q-Q-plot don't deviate visibly from the diagonal line for the residuals (bottom right).

The results for the dependent variable carried out with the Descriptive statistics analysis before the ANOVA revealed the same pattern (no significance for Shapiro-Wilk, normal Q-Q-plot) but it is not shown here because it did not give indication for concern.



ANOVA with one factor in jamovi

- **assumptions II:**
homogeneity of variance =
homoscedasticity

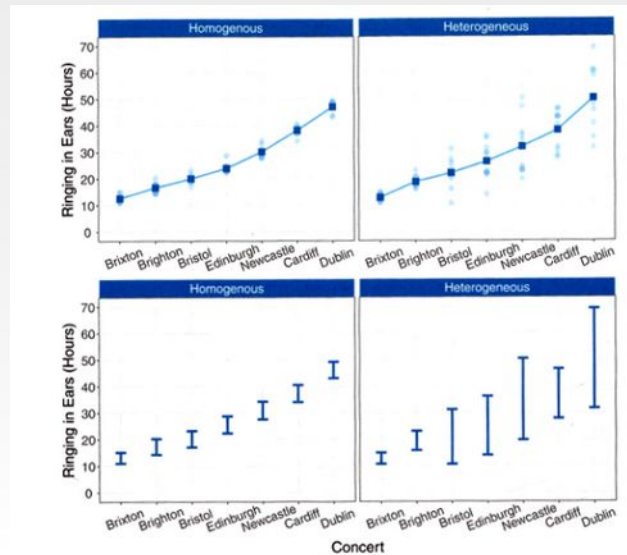


Figure 6.7 Graphs illustrating data with homogeneous (left) and heterogeneous (right) variances



Homogeneity of variance requires that the variances at each level of the factors we use in our model are similar. If they are not, as shown on the right-hand side, this would affect our within-groups square sum which again becomes part of the F-ratio (see slide 19). This F-ratio is used to test our model for significance (and thereby would variance inhomogeneity likely render our statements about the model's significance invalid).



ANOVA with one factor in jamovi

ANOVA

Dependent Variable: mood.gain

Fixed Factors: drug

Model Fit: Overall model test

Effect Size: η^2 partial η^2 ω^2

Assumption Checks:

- Homogeneity test
- Normality test
- Q-Q Plot

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	η^2	ω^2
drug	3.453	2	1.727	18.611	<.0001	0.713	0.662
Residuals	1.392	15	0.093				

Homogeneity of Variances Test (Levene's)

F	df1	df2	p
1.450	2	15	0.2657

Normality Test (Shapiro-Wilk)

Statistic	p
0.960	0.6053



To assess the **homogeneity of variance** (sometimes also called homoscedasticity), we open the drop-down-menu «Assumption checks» another time and tick «Homogeneity test». It is assumed that we've got only one value for the population standard deviation (i.e., σ), rather than separate values for each group (i.e., σ_k). That is, the different values for σ_k are assumed to be similar to σ .

The Levene's test assessing those differences and whether the assumption of homogeneity is violated, is not significant ($p = 0.266$).



ANOVA with one factor in jamovi

ANOVA

Dependent Variable: mood.gain

Fixed Factors: drug

Model Fit: Overall model test

Assumption Checks: Homogeneity test, Normality test, Q-Q Plot

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	η^2	ω^2
drug	3.453	2	1.727	18.611	<.0001	0.713	0.662
Residuals	1.392	15	0.093				

Homogeneity of Variances Test (Levene's)

F	df1	df2	p
1.450	2	15	0.2657

Normality Test (Shapiro-Wilk)

Statistic	p
0.960	0.6053



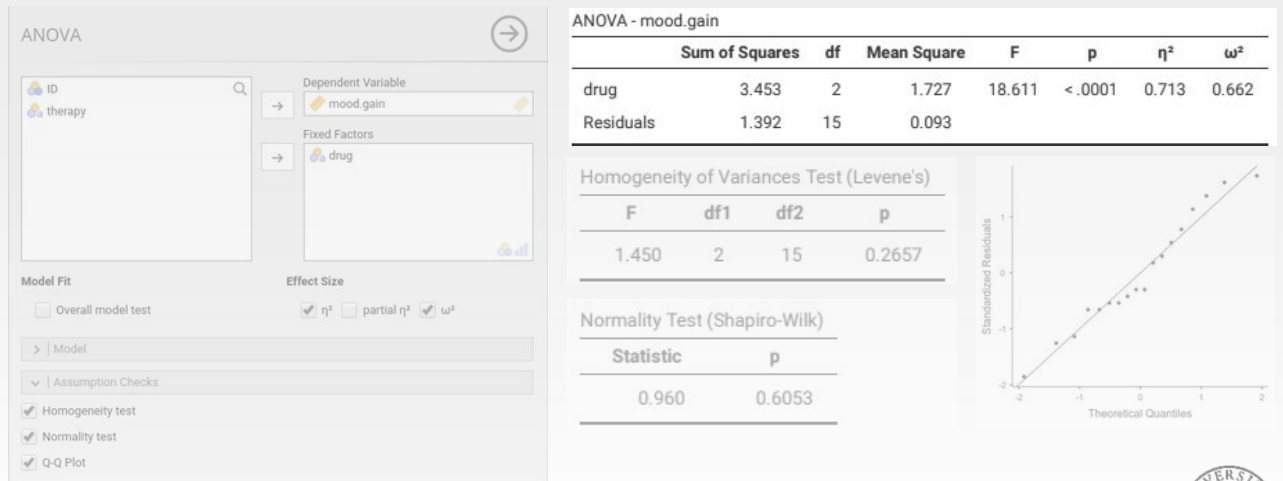
There is one last caveat with assumption checking: If the sample size is large then a significant Shapiro-Wilk or Levene's test may in fact be a false positive. With large numbers, the assumed measurement error on which our statistics is based becomes quite small. As a consequence, relatively small deviations might cause a significant result.

For normality, we have the opportunity to double-check this using the Q-Q-plot. If the points don't deviate much from the diagonal line, we can report this and use this as a justification why we "disregarded" a significant Shapiro-Wilk test.

For the homogeneity of variances one option might be to break down our design so that it just contains one factor and to test this factor using the One-way ANOVA with Welch's correction (described later). For either of the two assumption violations there is also the opportunity to use the non-parametric alternative Kruskal-Wallis (also described later).



ANOVA with one factor in jamovi



Finally, we are sure that our assumptions are met and we can have a closer look at our main results. They reveal that the effect of drug is statistically highly significant ($p < 0.001$).

Apart from significance, we are interested in whether the effect also has practical significance by checking the effect sizes.

η^2 is a measure of what proportion of the variance in the outcome variable (*mood.gain*) we can explain using our model, i.e., in terms of the predictor (*drug*): $\eta^2 = SS_b / SS_{tot} = 3.45 / 4.85 = 0.713$

The η^2 -value is very closely related to the concept of R^2 that we discussed in linear regression, and has an equivalent interpretation: A value of $\eta^2 = 0$ means no relationship at all between the two, whereas a value of $\eta^2 = 1$ means that the relationship is perfect.



ANOVA with one factor in jamovi

ANOVA

Dependent Variable: mood.gain

Fixed Factors: drug

Model Fit: Overall model test

Effect Size: η^2 partial η^2 ω^2

Assumption Checks:

- Homogeneity test
- Normality test
- Q-Q Plot

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	η^2	ω^2
drug	3.453	2	1.727	18.611	<.0001	0.713	0.662
Residuals	1.392	15	0.093				

Homogeneity of Variances Test (Levene's)

F	df1	df2	p
1.450	2	15	0.2657

Normality Test (Shapiro-Wilk)

Statistic	p
0.960	0.6053



There are recommendations to use ω^2 instead because that measure is less biased when having small sample sizes. You can get ω^2 by using the respective tick box in jamovi.

The calculation of ω^2 is a little more complicated:

$$\begin{aligned}\omega^2 &= (SS_b - df_b \cdot MS_w) / (SS_{tot} + MS_w) \\ &= (3.45 - 2 \cdot 0.09) / (4.85 + 0.09) \\ &= 3.27 / 4.94 = 0.662\end{aligned}$$

Whereas η^2 is more intuitive, ω^2 corrects for small sample sizes and the bias that we collected measurements from a sample while we try to make statements about a certain relationship within the population. However, still, the interpretation follows the same rationale as R^2 in the linear regression.

Cohen (1992) regards correlation coefficients of 0.5 as large effect ($r/R = 0.50 \rightarrow r^2/R^2 = 0.25$). Both $\eta^2 = 0.713$ and $\omega^2 = 0.662$ are far above that and represent very substantial effect sizes and high practical relevance in addition to significance.



ANOVA with one factor in jamovi

Comparison with our step-by-step calculations:

	Sum of Squares	df	Mean Square	F	p	η^2	ω^2
drug	3.453	2	1.727	18.611	<.0001	0.713	0.662
Residuals	1.392	15	0.093				

	SS	df	MS	F	p	η^2	ω^2
between	3.453	2	1.727	18.611	0.0001	0.713	0.662
within	1.392	15	0.093				
total	4.845						



If we compare the output we obtain to the one we earlier calculated in LibreOffice, we see that the values are identical to the last decimal. I obviously put in enough concentration when creating those calculations (and some time for finding the mistakes... ;-).

One difference though is the header of line of the tables. It is denoted as “between” groups variance in the spreadsheet and corresponds to the effect that the independent variable *drug* has on the dependent variable *mood.gain* in jamovi. What is called the within groups variance corresponds to the “leftover” or unexplained variability called the Residuals.



ANOVA with one factor in jamovi

post-hoc tests:

variance is pooled, not possible to determine which pair difference caused significance → post-hoc-tests

possibility:	is $\mu_P = \mu_A$?	is $\mu_P = \mu_J$?	is $\mu_A = \mu_J$?	which hypothesis?
1	✓	✓	✓	null
2	✓	✓		alternative
3	✓		✓	alternative
4	✓			alternative
5		✓	✓	alternative
6		✓		alternative
7			✓	alternative
8				alternative



In our analysis, we pooled the variance. That is, we can't determine which difference in mean scores for the different drugs was decisive for being able to reject the H_0 for our model.

Our null hypothesis contained of three parts:

$$H_0: \mu_P = \mu_A = \mu_J$$

We could instead also write it like:

$$H_0: \mu_P = \mu_A \text{ AND } \mu_P = \mu_J \text{ AND } \mu_A = \mu_J$$

This results in seven possible options why the test became significant. We often have interest in determining which of these seven options was responsible. To get a clearer answer, it might help to run some tests, denoted as post-hoc tests.



ANOVA with one factor in jamovi

post-hoc tests:

Post Hoc Tests

→ drug

Correction

No correction

Tukey

Scheffe

Bonferroni

Holm

Effect Size

Cohen's d

Post Hoc Comparisons - drug

Comparison		Mean Difference	SE	df	t
anxifree	- joyzepam	-0.767	0.176	15.000	-4.360
	- placebo	0.267	0.176	15.000	1.516
joyzepam	- placebo	1.033	0.176	15.000	5.876

Post Hoc Comparisons - drug

Comparison		p	Ptukey	Pscheffe	Pbonferroni	Pholm
anxifree	- joyzepam	0.0006	0.0015	0.0022	0.0017	0.0011
	- placebo	0.1502	0.3115	0.3431	0.4506	0.1502
joyzepam	- placebo	< .0001	< .0001	0.0001	< .0001	< .0001

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 35



Basically, what we do is to run three separate t-tests for each pairs of means: *placebo* vs. *Anxifree*, *placebo* vs. *Joyzepam*, *Anxifree* vs. *Joyzepam*. To do this in jamovi, go to the drop-down-menu «Post Hoc Tests», move the variable *drug* across from the left into the variable box on the right, and then tick the ‘No correction’, “Tukey”, “Scheffe”, “Bonferroni” and “Holm” checkboxes.

This will produce a neat table showing all the pairwise t-test comparisons among the three levels of the drug variable (please note that I divided the table so it looks different from how it appears in jamovi).

Of the correction options, “Bonferroni” and “Holm” are easier accessible, therefore I will explain them, whereas “Tukey”, and “Scheffe” are more complicated to calculate and I will just say what they are good for.



ANOVA with one factor in jamovi

post-hoc tests:

Post Hoc Tests

→ drug

Correction

No correction

Tukey

Scheffe

Bonferroni

Holm

Effect Size

Cohen's d

Post Hoc Comparisons - drug

Comparison		Mean Difference	SE	df	t
anxifree	- joyzepam	-0.767	0.176	15.000	-4.360
	- placebo	0.267	0.176	15.000	1.516
joyzepam	- placebo	1.033	0.176	15.000	5.876

Post Hoc Comparisons - drug

Comparison		p	Ptukey	Pscheffe	Pbonferroni	Pholm
anxifree	- joyzepam	0.0006	0.0015	0.0022	0.0017	0.0011
	- placebo	0.1502	0.3115	0.3431	0.4506	0.1502
joyzepam	- placebo	< .0001	< .0001	0.0001	< .0001	< .0001

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 36



When running post-hoc-analyses, a lot of care is required: Each individual t-test is designed to have a 5% Type I error rate (i.e., $\alpha = 0.05$) and our analysis included three of these tests. For 10 different groups / levels of the variable, there would have been 45 “post hoc” t-tests (and you’d expect 2 or 3 of them to come up significant by chance alone; $45 \cdot 0.05 = 2.25$). In addition, the experiment-wise α that we would have to accept, would raise to about 90%. I put this calculation into Clinical trial – Step-by-step.xlsx (sheet: “Alpha error inflation”).

The solution is to introduce an adjustment to the p-value, which aims to control the total error rate across all tests we conduct.



ANOVA with one factor in jamovi

post-hoc tests:

Post Hoc Tests

→ drug

Correction

No correction

Tukey

Scheffe

Bonferroni

Holm

Effect Size

Cohen's d

Post Hoc Comparisons - drug

Comparison					
drug	drug	Mean Difference	SE	df	t
anxifree	- joyzepam	-0.767	0.176	15.000	-4.360
	- placebo	0.267	0.176	15.000	1.516
joyzepam	- placebo	1.033	0.176	15.000	5.876

Post Hoc Comparisons - drug

Comparison		p	Ptukey	Pscheffe	Pbonferroni	Pholm
anxifree	- joyzepam	0.0006	0.0015	0.0022	0.0017	0.0011
	- placebo	0.1502	0.3115	0.3431	0.4506	0.1502
joyzepam	- placebo	< .0001	< .0001	0.0001	< .0001	< .0001

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 37



The simplest of these adjustments is called the Bonferroni correction. All raw p-values are multiplied by the number of comparisons we conducted (3 in the current analysis; cf. the 3 lines in the Table “Post-Hoc Comparisons – drug”). For example, for the comparison between «anxifree» and «joyzepam» the original (uncorrected) p-value $p = 0.0006$ is multiplied by 3 resulting in $p = 0.0017$ ($0.0006 \cdot 3 = 0.0018$ – the difference is due to some rounding error).



ANOVA with one factor in jamovi

post-hoc tests:

Post Hoc Tests

→ drug

Correction

No correction

Tukey

Scheffe

Bonferroni

Holm

Effect Size

Cohen's d

Post Hoc Comparisons - drug

Comparison		Mean Difference	SE	df	t
anxifree	- joyzepam	-0.767	0.176	15.000	-4.360
	- placebo	0.267	0.176	15.000	1.516
joyzepam	- placebo	1.033	0.176	15.000	5.876

Post Hoc Comparisons - drug

Comparison		p	Ptukey	Pscheffe	Pbonferroni	Pholm
anxifree	- joyzepam	0.0006	0.0015	0.0022	0.0017	0.0011
	- placebo	0.1502	0.3115	0.3431	0.4506	0.1502
joyzepam	- placebo	< .0001	< .0001	0.0001	< .0001	< .0001

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 38



Often, the Holm correction is used instead. The idea behind the Holm correction is to assume that the tests are done sequentially. We would start with the comparison where we obtained the smallest p-value and multiply that with the total number of comparisons – 3 – then proceed to the next smaller p-value and multiply that with 2 and multiply the largest p-value with 1 (i.e., leave it as it is).

For our analysis, the smallest p-value comes from the comparison of *joyzepam* with the *placebo*: it is $p < 0.0001$ and it is so low that even multiplying it with 3 results in the same $p < 0.0001$ after correction. The next larger value comes from the comparison of *anxifree* and *joyzepam*: the original $p = 0.0006$ is multiplied with 2 resulting in $p = 0.0011$. Finally, the p-value from the comparison of *anxifree* to the *placebo* $p = 0.1502$ is multiplied by 1 (i.e., not corrected) and remains $p = 0.1502$.



ANOVA with one factor in jamovi

post-hoc tests:

Post Hoc Tests

→ drug

Correction

No correction

Tukey

Scheffe

Bonferroni

Holm

Effect Size

Cohen's d

Post Hoc Comparisons - drug

Comparison					
drug	drug	Mean Difference	SE	df	t
anxifree	- joyzepam	-0.767	0.176	15.000	-4.360
	- placebo	0.267	0.176	15.000	1.516
joyzepam	- placebo	1.033	0.176	15.000	5.876

Post Hoc Comparisons - drug

Comparison						
drug	drug	p	Ptukey	Pscheffe	Pbonferroni	Pholm
anxifree	- joyzepam	0.0006	0.0015	0.0022	0.0017	0.0011
	- placebo	0.1502	0.3115	0.3431	0.4506	0.1502
joyzepam	- placebo	< .0001	< .0001	0.0001	< .0001	< .0001

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 39



Compared to Bonferroni, the Holm correction is a little harder to calculate (but only a little). In exchange, it has a lower Type II error rate. As counter-intuitive as it might seem, it has the same Type I error rate. As a consequence, there's no reason to use the simpler Bonferroni correction since it is always outperformed by the slightly more elaborate Holm correction.



ANOVA with one factor in jamovi

post-hoc tests:

Post Hoc Tests

→ drug

Correction

No correction

Tukey

Scheffe

Bonferroni

Holm

Effect Size

Cohen's d

Post Hoc Comparisons - drug

Comparison		Mean Difference	SE	df	t
anxifree	- joyzepam	-0.767	0.176	15.000	-4.360
	- placebo	0.267	0.176	15.000	1.516
joyzepam	- placebo	1.033	0.176	15.000	5.876

Post Hoc Comparisons - drug

Comparison		p	Ptukey	Pscheffe	Pbonferroni	Pholm
anxifree	- joyzepam	0.0006	0.0015	0.0022	0.0017	0.0011
	- placebo	0.1502	0.3115	0.3431	0.4506	0.1502
joyzepam	- placebo	< .0001	< .0001	0.0001	< .0001	< .0001

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 40



Tukey and Scheffé are a little more complex to calculate. As a general rule, Tukey is more appropriate for equal groups sizes, Scheffé if group sizes are unequal (in our analysis with $N = 6$ for all groups, Tukey would be more appropriate).

When reporting results from a post-hoc-analysis, you might write up the results like this: "Post hoc tests (using the Holm correction to adjust p) indicated that Joyzepam produced a significantly larger mood change than both Anxifree ($p = 0.001$) and the placebo ($p < 0.001$). In contrast, there was no evidence that Anxifree performed better than the placebo ($p = 0.150$)."

Don't be confused, the four decimals in the output are rounded to three decimals in the sentence.



If assumptions are not met...

(1) One-way ANOVA (and Welch's correction):

The screenshot shows the Jamovi One-Way ANOVA interface. The dependent variable is 'mood.gain' and the grouping variable is 'drug'. The 'Variances' section has 'Don't assume equal (Welch's)' and 'Assume equal (Fisher's)' checked. The 'Post-Hoc Tests' section has 'Games-Howell (unequal variances)' selected. The results table shows the following data:

One-Way ANOVA					
		F	df1	df2	p
mood.gain	Welch's	26.322	2	9.493	0.0001
	Fisher's	18.611	2	15	<.0001

Post Hoc Tests

Games-Howell Post-Hoc Test – mood.gain

		anxifree	joyzepam	placebo
anxifree	Mean difference	–	-0.767	0.267
	p-value	–	0.0080	0.4027
joyzepam	Mean difference		–	1.033
	p-value		–	0.0001
placebo	Mean difference			–
	p-value			–

SLIDE 41



If the assumption of homogeneity of variances were violated, it would have been an option to choose “ANOVA” from the icon bar and then select “One-way ANOVA” instead “ANOVA” (which we just used). The One-way ANOVA implements a procedure to handle unequal variances on the different stages of a factor by using Welch’s correction. The post-hoc tests within jamovi’s One-way ANOVA can also handle inhomogeneous variances (Games-Howell) as well as homogeneous ones (Tukey).

Those of you who are very attentive and / or knowledgeable might remember the name Welch from the t-test where the same correction method exists. The results we obtain are similar / identical to those obtained with the “normal” ANOVA. With Welch’s correction, the effect of drug is (still) highly significant. Post-hoc tests show that Joyzepam leads to larger mood changes as both Anxifree and the placebo while these two don’t differ from one another.



If assumptions are not met...

(2) Non-param. one-way ANOVA (Kruskal-Wallis):

One-Way ANOVA (Non-parametric)

Dependent Variables: mood.gain

Grouping Variable: drug

Effect size
 DSCF pairwise comparisons

Kruskal-Wallis			
	χ^2	df	p
mood.gain	12.076	2	0.0024

Dwass-Steel-Critchlow-Fligner pairwise comparisons

Pairwise comparisons - mood.gain			
		W	p
anxifree	joyzepam	4.091	0.0107
anxifree	placebo	-1.704	0.4504
joyzepam	placebo	-4.098	0.0105

In either case of violation (normality and homogeneity of variances), there is yet another option: Using a non-parametric equivalent of the ANOVA with one factor, called the Kruskal-Wallis rank sum test. This analysis can be found under ANOVA → Non-parametric → One-way ANOVA (Kruskal-Wallis). Other than its parametric equivalent which is based upon variances, the Kruskal-Wallis test sorts the values of the dependent variable (*mood.gain*) and assigns ranks to each value. Those ranks are then summed up per group and then compared between the groups. A detailed description of the mathematics behind the test can be found on p. 349 – 352 of the jamovi-book).



If assumptions are not met...

(2) Non-param. one-way ANOVA (Kruskal-Wallis):

One-Way ANOVA (Non-parametric)

Dependent Variables: mood.gain

Grouping Variable: drug

Effect size
 DSCF pairwise comparisons

Kruskal-Wallis

	χ^2	df	p
mood.gain	12.076	2	0.0024

Dwass-Steel-Critchlow-Fligner pairwise comparisons

Pairwise comparisons - mood.gain

		W	p
anxifree	joyzepam	4.091	0.0107
anxifree	placebo	-1.704	0.4504
joyzepam	placebo	-4.098	0.0105

It reveals results that are quite similar to the parametric equivalent: The p-value for the whole model is $p = 0.0024$.

The Kruskal-Wallis test even includes something similar to the post-hoc-tests in the ANOVA, called “Dwass-Steel-Critchlow-Fligner pairwise comparisons” (activated by ticking «DSCF pairwise comparisons»). Here we also obtain results that largely match those obtained with the first ANOVA: *Joyzepam* significantly differs from both *Anxifree* and the *placebo*, whereas *Anxifree* and the *placebo* don’t significantly differ from one another.

Please note that both the One-Way ANOVA with Welch’s correction as well as the non-parametric Kruskal-Wallis test are only available for designs with one factor. You would have to drop further factors from your analysis and to concentrate on the factor of most interest if you want to use them.



ANOVA with more than one factor in jamovi

The framework of the ANOVA can be extended to encompass multiple predictors. For instance, suppose we were interested in using a reading comprehension test to measure student achievements in three different schools, and we suspect that girls and boys are developing at different rates.

Each student is classified in two different ways: on the basis of their gender and on the basis of their school. What we'd like to do is analyse the reading comprehension scores in terms of both of these grouping variables. Factorial ANOVA is the tool for answering such questions. Dependent on the number of grouping variables (factors), we could also refer to the analysis as a two-way ANOVA.



Some background

	no therapy	CBT	total
placebo	μ_{11}	μ_{12}	$\mu_{1.}$
anxifree	μ_{21}	μ_{22}	$\mu_{2.}$
joyzepam	μ_{31}	μ_{32}	$\mu_{3.}$
total	$\mu_{.1}$	$\mu_{.2}$	$\mu_{..}$

	no therapy	CBT	total
placebo	0.30	0.60	0.45
anxifree	0.40	1.03	0.72
joyzepam	1.47	1.50	1.48
total	0.72	1.04	0.88

Contingency Tables

drug	therapy		Total
	CBT	no.therapy	
anxifree	3	3	6
joyzepam	3	3	6
placebo	3	3	6
Total	9	9	18



For such two-way ANOVA we can continue with the clinical trial data set that we already used for our one-way ANOVA. In addition to looking at the effect of different *drugs* on the *mood.gain* experienced by each person, we could further look whether there was an effect of *therapy*. We used *therapy* before (in order to demonstrate the equivalence of t-test and ANOVA) but without significant result ($p = 0.2098$).

From the contingency table, we can see that the design is completely crossed (i.e., there exist all possible combinations of the two factors) and even balanced (with an equal number of people in each group).



Some background

	no therapy	CBT	total
placebo	μ_{11}	μ_{12}	$\mu_{1.}$
anxifree	μ_{21}	μ_{22}	$\mu_{2.}$
jozepam	μ_{31}	μ_{32}	$\mu_{3.}$
total	$\mu_{.1}$	$\mu_{.2}$	$\mu_{..}$

	no therapy	CBT	total
placebo	0.30	0.60	0.45
anxifree	0.40	1.03	0.72
jozepam	1.47	1.50	1.48
total	0.72	1.04	0.88

first pair of hypotheses
(for the effect of *drug*)

$$H_0: \mu_{1.} = \mu_{2.} = \mu_{3.}$$

$$H_1: \mu_{1.} \neq \mu_{2.} \text{ OR } \mu_{1.} \neq \mu_{3.} \\ \text{OR } \mu_{2.} \neq \mu_{3.}$$



Like the one-way ANOVA, the factorial ANOVA is a tool for testing hypotheses about population means. For the first pair of hypotheses (for the effect of *drug*), the null hypothesis (H_0) claims that all row means are the same: $\mu_{1.} = \mu_{2.} = \mu_{3.}$ Our alternative hypothesis (H_1) claims that at least one row mean is different, so either is $\mu_{1.} \neq \mu_{2.}$ OR $\mu_{1.} \neq \mu_{3.}$ OR $\mu_{2.} \neq \mu_{3.}$ Note, though, that the indices have changed. Whereas we denoted them with the kind of drug earlier (μ_p, μ_a, μ_j) so are they now changed to denote the means of the rows in the table above: $\mu_{1.}, \mu_{2.}, \mu_{3.}$ Please note the tiny dot in the index: for the rows it is in the second position (indicating an average over columns) and for the columns it is in the first position. This is done to increase flexibility as we could easily increase the number of factors, e.g., $\mu_{1..}$ (even though then we couldn't summarize them so handy in a table any more).



Some background

	no therapy	CBT	total
placebo	μ_{11}	μ_{12}	$\mu_{1.}$
anxifree	μ_{21}	μ_{22}	$\mu_{2.}$
joyzepam	μ_{31}	μ_{32}	$\mu_{3.}$
total	$\mu_{.1}$	$\mu_{.2}$	$\mu_{..}$

second pair of hypotheses (for the effect of *therapy*)

$$H_0: \mu_{.1} = \mu_{.2}$$

$$H_1: \mu_{.1} \neq \mu_{.2}$$

	no therapy	CBT	total
placebo	0.30	0.60	0.45
anxifree	0.40	1.03	0.72
joyzepam	1.47	1.50	1.48
total	0.72	1.04	0.88



The second pair of hypotheses (for the effect of *therapy*) is much easier (as it only contains two levels). Our null hypothesis (H_0) assumes that our **column means** are the same: $\mu_{.1} = \mu_{.2}$. Our alternative hypothesis (H_1) claims instead that our column means are different: $\mu_{.1} \neq \mu_{.2}$.

For this analysis each person is cross-classified by the drug they were given (a factor with 3 levels) and what therapy they received (a factor with 2 levels). We refer to this as a 3×2 factorial design.

Our hypotheses are exactly the same as for earlier analyses. Even the sum of squares, degrees of freedoms, and mean squares stay the same.

However, it often is better to run a single analysis that includes both *drug* and *therapy* as predictors. This has to do with the residuals and is discussed later.



Factorial ANOVA in jamovi (Main eff.)

	Sum of Squares	df	Mean Square	F	p
drug	3.453	2	1.727	26.149	<.0001
therapy	0.467	1	0.467	7.076	0.0187
Residuals	0.924	14	0.066		

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 48



When assembling this analysis, we create a new ANOVA with ANOVA → ANOVA, assign *mood.gain* to «Dependent Variable» and *drug* and *therapy* to fixed factors. In order to keep the model simple, and to concentrate on the main effects of *drug* and *therapy* first, we have to open the drop-down-menu “Model” and remove the interaction *drug* * *therapy* from the variable list “Model Terms” (use the arrow to the left for removing it).

The ANOVA table for this more complex factorial ANOVA can be read and interpreted as the table for the simpler one way ANOVA. The factorial ANOVA for our 3×2 design found a significant main effect of *drug*: $F_{(2,14)} = 26.15$, $p < 0.001$; as well as a significant main effect of *therapy*: $F_{(1,14)} = 7.08$, $p = 0.019$. This shows that the basic logic and structure behind factorial ANOVA is the same as that which underpins one way ANOVA.



Factorial ANOVA in jamovi (Main eff.)

ANOVA

Search

Dependent Variable
mood.gain

Fixed Factors
drug
therapy

Model Fit
 Overall model test

Effect Size
 η^2 partial η^2 ω^2

Model

Components
drug
therapy

Model Terms
drug
therapy
drug * therapy

Sum of squares Type 3

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p
drug	3.453	2	1.727	26.149	<.0001
therapy	0.467	1	0.467	7.076	0.0187
Residuals	0.924	14	0.066		

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 49



From the F-ratio of the first analysis by hand where we started with $F = MS_b / MS_w$, we replace the MS_w with MS_R since this mean square (MS) is denoted in the jamovi-table with Residuals (hence the index R). We extend or refine the F-ratio for the first factor (*drug*) into $F_D = MS_D / MS_R$. An equivalent formula exists for the second factor (*therapy*): $F_T = MS_T / MS_R$. In comparison to the previous ANOVA (only using *drug* as a independent variable), we can see that SS_R and MS_R went down. This means that *therapy* was suitable to account for variance that wasn't explained by first factor *drug*, which led to the SS_R (i.e., the sum of squares of the residuals, the variation which couldn't be explained by the model) went down and with it MS_R . MS_R is used to calculate F_D and F_T . F_D is therefore higher than in the previous analysis (with one factor) and the p-value get smaller (even though it is already so small that this is not directly visible).



Factorial ANOVA in jamovi (Main eff.)

ANOVA

Dependent Variable: mood.gain

Fixed Factors: drug, therapy

Model Fit: Overall model test

Effect Size: η^2 partial η^2 ω^2

Model Terms: drug, therapy, drug * therapy

Sum of squares: Type 3

	Sum of Squares	df	Mean Square	F	p
drug	3.453	2	1.727	26.149	<.0001
therapy	0.467	1	0.467	7.076	0.0187
Residuals	0.924	14	0.066		

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 50



Mathematically, SS_A and SS_B are also calculated by quite similar (unfortunately a little more complex) formulas as those used in the one way ANOVA:

$$SS_A = (N \cdot C) \cdot \sum_{r=1}^R (\bar{Y}_{r.} - \bar{Y}_{..})^2 \quad SS_B = (N \cdot R) \cdot \sum_{c=1}^C (\bar{Y}_{.c} - \bar{Y}_{..})^2$$

Please note that to keep the formulas general, SS_D is SS_A here and SS_T is SS_B . This generic form also means that the formulas are arranged in rows (for factor A) and columns (for factor B), hence the indices r or c .

This is also the form used in an external file where I describe how these calculations are carried out by hand. You find the file "Analysis of Variance - Factorial by hand.pdf". Don't be afraid to have a look, it is easier and less frightening than you may believe. The calculations for the main effects are on the first pages and that document.



Factorial ANOVA in jamovi (Main eff.)

ANOVA

Search

Dependent Variable
mood.gain

Fixed Factors
drug
therapy

Model Fit
 Overall model test

Effect Size
 η^2 partial η^2 ω^2

Model

Components
drug
therapy

Model Terms
drug
therapy
drug * therapy

Sum of squares Type 3

ANOVA - mood.gain					
	Sum of Squares	df	Mean Square	F	p
drug	3.453	2	1.727	26.149	<.0001
therapy	0.467	1	0.467	7.076	0.0187
Residuals	0.924	14	0.066		

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 51



When comparing the results from the two individual ANOVAs that we did with *drug* and *therapy* with the current analysis, we can see that the sum of squares, degrees of freedom and mean square for each factor are identical. The null and alternative hypotheses tested by the one-way ANOVAs are also identical to the hypotheses tested by the factorial ANOVA.

However, the results (F- and p-statistics) are different. This becomes most clearly when considering the one-way ANOVA for *therapy* that we did to demonstrate the equivalence of t-test and F-test. There, we didn't find a significant effect (the p-value was 0.2098), whereas the main effect of *therapy* within the context of the two-way ANOVA is significant ($p = 0.019$). Why do the results from the two analyses differ so much?



Factorial ANOVA in jamovi (Main eff.)

ANOVA

Dependent Variable: mood.gain

Fixed Factors: drug, therapy

Model Fit: Overall model test η^2 partial η^2 ω^2

Effect Size: η^2 partial η^2 ω^2

Model Terms: drug, therapy, drug * therapy

Sum of squares Type 3

	Sum of Squares	df	Mean Square	F	p
drug	3.453	2	1.727	26.149	<.0001
therapy	0.467	1	0.467	7.076	0.0187
Residuals	0.924	14	0.066		

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 52



The reason lies in how the residuals are calculated.

The idea behind an F-test is to compare the variability that can be attributed to a particular factor with the variability that cannot be accounted for by the whole model (the residuals). The one-way ANOVA for *therapy* “ignored” the effect of *drug*, and vice versa ignored the one-way ANOVA for *drug* the effect of *therapy*.

Variability induced by the “ignored” effects ends up in the residuals. As a consequence, the data appear to include more “noise” or unexplained variance. If we ignore a factor that actually matters (e.g., *drug*) when trying to assess the contribution of something else (e.g., *therapy*), our analysis is distorted.

That is, you should be carefully consider (already when designing an experiment) which variables might make a difference to explain variation in order to prevent that they end up in the residuals.



Factorial ANOVA in jamovi (Main eff.)

ANOVA

Search

Dependent Variable
mood.gain

Fixed Factors
drug
therapy

Model Fit
 Overall model test

Effect Size
 η^2 partial η^2 ω^2

Model

Components
drug
therapy

Model Terms
drug
therapy
drug * therapy

Sum of squares Type 3

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p
drug	3.453	2	1.727	26.149	< .0001
therapy	0.467	1	0.467	7.076	0.0187
Residuals	0.924	14	0.066		

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 53



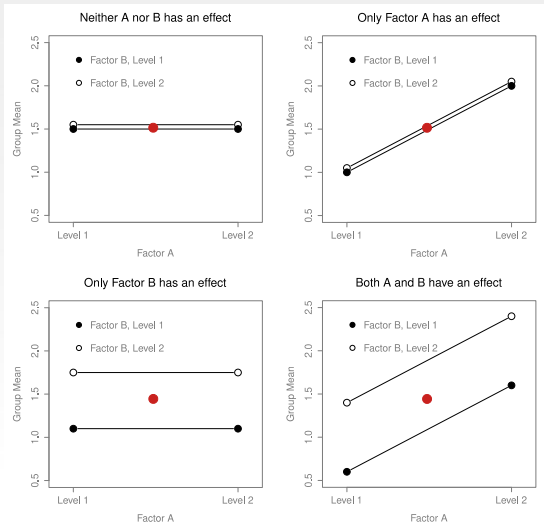
However, this is not an argument to indiscriminately add factors to our models. Factors that we include have to be genuinely relevant to the phenomenon of interest. And strictly speaking, we should only include factors that we ***hypothesized*** to have an influence when designing the experiment, not only we thought they might be a nice addition while analysing our data.

If an additional factor that we considered turned out to be non-significant in a three-way ANOVA, it is perfectly fine to disregard it and just report the simpler two-way ANOVA. Often doing that makes the model easier to understand and to report.

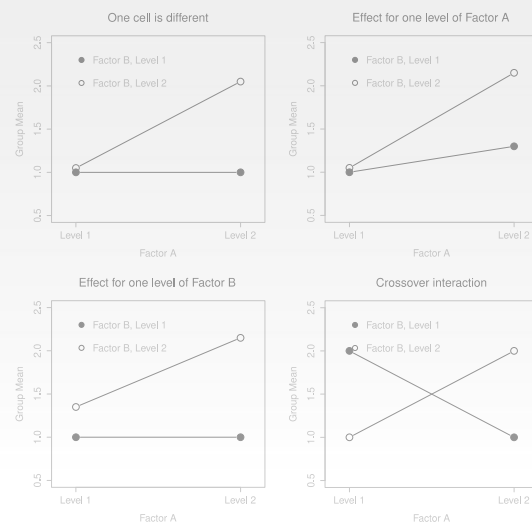


Visualization of effects

Main effects



Interactions



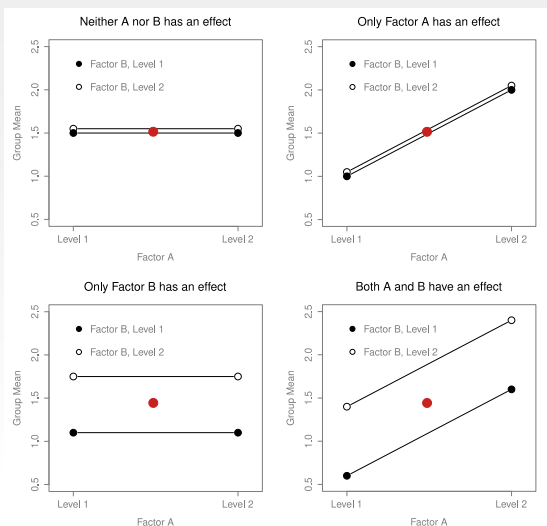
The ANOVA model that we've used so far covers a range of different patterns that we might observe in our data. These are called **main effects** and are shown in the **figure on the left**. As a general rule, when we deal with a main effect, the lines in the plot are parallel.

In a two-way ANOVA design we have four possible main effects: neither Factor A nor Factor B matters (top-left quadrant), only Factor A matters (top-right quadrant), only Factor B matters (bottom-left quadrant), and both Factor A and Factor B matter (bottom-right quadrant).

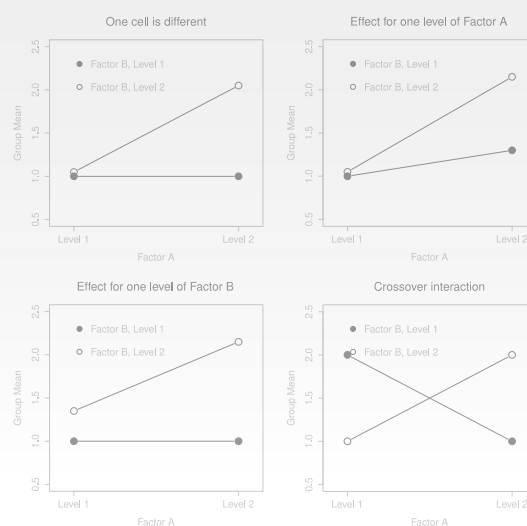


Visualization of effects

Main effects



Interactions

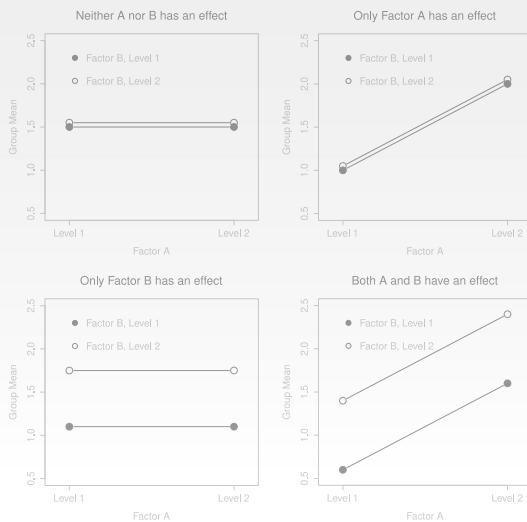


I used a little read dot to make clear where the mean of the whole sample would have been. For the example top-left all dots representing group means (black and white) are on the same level as the red dot. For the example top-right the means are lower left and higher right. For the example bottom-left one line lies above, and one below. The last quadrant it is a combination of the two previous ones. This is the situation, that we had in the ANOVA for the clinicaltrials data set where we found significant main effects for both drug and therapy. The plots are for demonstrating a simple general case. If we wanted to adapt them to our clinicaltrials-ANOVA, the x-axis would have to contain a third level (and the factors would have to be adapted (Factor A – drug, Factor B – therapy)).

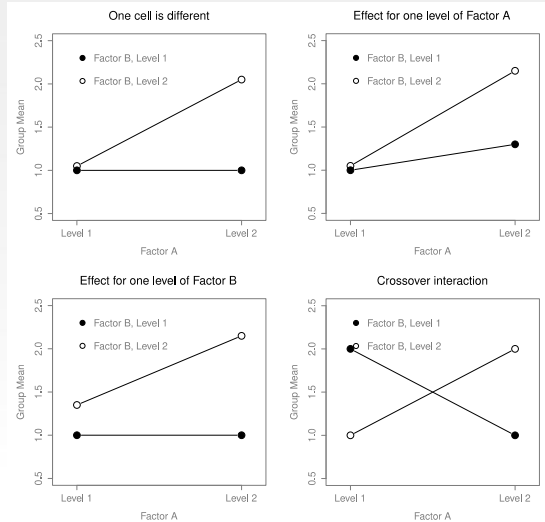


Visualization of effects

Main effects



Interactions



In addition to the main effects that we covered so far, there is another class of effects called ***interactions***. These are shown in the ***figure on the right-hand side***. An interaction between Factor A and Factor B occurs whenever the effect of Factor A is different, depending on which level of Factor B we're considering.

This is maybe a bit abstract to understand. A concrete example for an interaction using our current data set is to suppose that the operation of Anxifree and Joyzepam is governed by different physiological mechanisms. A consequence is that Joyzepam has a similar effect on mood regardless of whether one is in therapy or not. In contrast is Anxifree much more effective when administered in conjunction with CBT. That is, the effect of *drug* is different in dependence of which level of *therapy* we are on (this is shown in the quadrant top-left).



Factorial ANOVA in jamovi (Interact.)

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p
drug	3.453	2	1.727	31.714	< .0001
therapy	0.467	1	0.467	8.582	0.0126
drug * therapy	0.271	2	0.136	2.490	0.1246
Residuals	0.653	12	0.054		

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 57



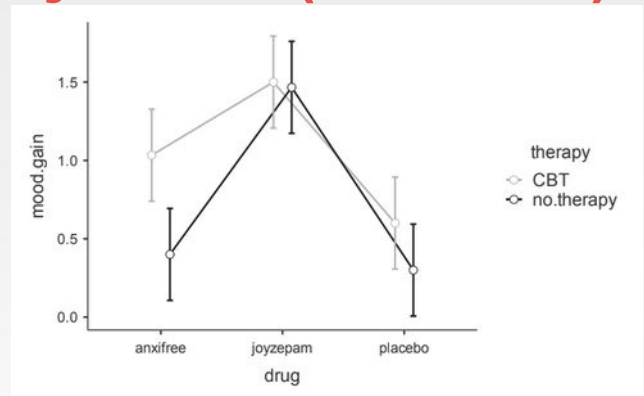
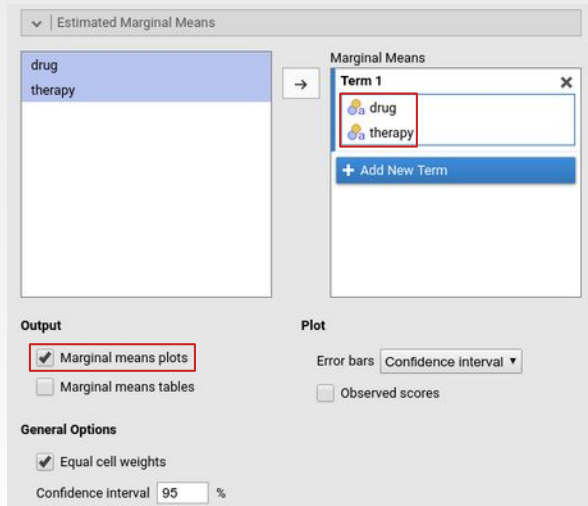
The ANOVA that we used so far doesn't capture such an interaction (actually we removed it to have a didactic step-by-step approach).

Therefore we have now to extend our ANOVA again to include the interaction. We do this by opening the drop-down-box «Model» again, selecting the two variables in «Components» and assigning them to «Model Terms».

This also extends our main table «ANOVA – mood.gain». It allows us to assess for the interaction (*drug * therapy*) whether the effect is real, i.e. not just random variation due to chance. However, while the two main effects for *drug* and *therapy* are significant (as we already saw in the previous model), the interaction is not significant ($p = 0.125$).



Factorial ANOVA in jamovi (Interact.)



Often it is easier to assess main effects and interactions visually (especially if we have complex model with three or more factors and their respective interactions). For the visualization, we open the drop-down-menu «Estimated Marginal Means» and move *drug* and *therapy* into the variable list «Marginal Means» within 'Term 1'. If there were only main effects, the two lines should be (more or less) parallel (if you were to remove the interaction we just added under «Model» this in fact would happen). Even though the interaction was not statistically significant ($p = 0.125$), the effect of CBT (i.e., the distance between the black and the grey line) varies: when the drug is Joyzepam (middle) it appears to be near zero (the black and grey circle are at about the same level), it is a little larger when a placebo is used (right), however, when Anxifree is administered, the effect of CBT is largest (left).



Factorial ANOVA in jamovi (Interact.)

Estimated Marginal Means

drug
therapy

Marginal Means

Term 1

drug
therapy

+ Add New Term

Output

Marginal means plots

Marginal means tables

General Options

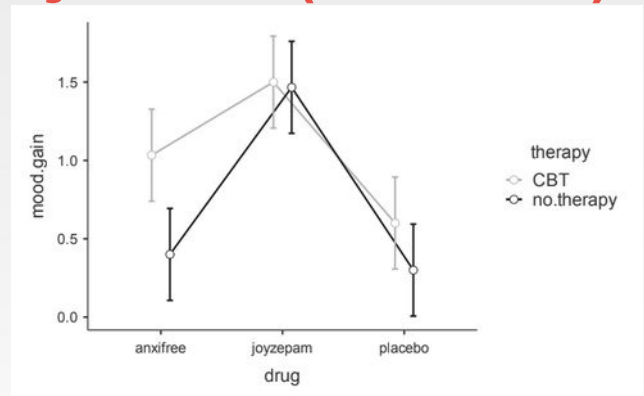
Equal cell weights

Confidence interval 95 %

Plot

Error bars Confidence interval

Observed scores



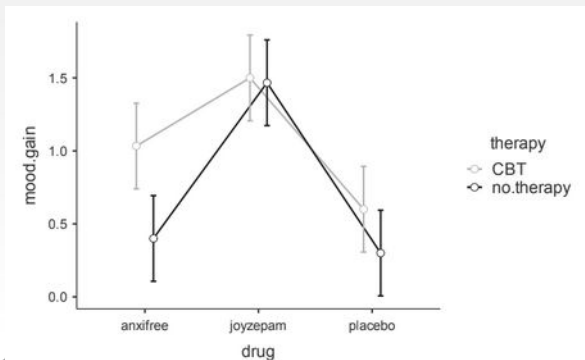
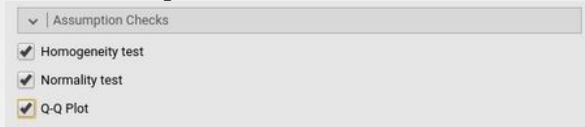
We obtained two significant main effects whereas the interaction wasn't significant. The opposite case may also happen. Then, we would have obtained a significant interaction effect but no corresponding main effects.

A prototypical example is the crossover interaction shown three slides ago (bottom right corner). This is a bit difficult to interpret. When adjusting our dataset a bit to be a 2×2 design, we could imagine comparing the combination of two pharmacological interventions (e.g., Anxifree vs Joyzepam) and two different treatments for phobias (e.g., systematic desensitisation vs flooding). If we found that Anxifree had no effect when desensitisation was used, and Joyzepam had no effect when flooding was the treatment, this would represent a classic crossover interaction. We'd find that there is no main effect of drug or therapy: they cancel out each other as a consequence of the interaction.



Factorial ANOVA in jamovi (Interact.)

assumption checks:

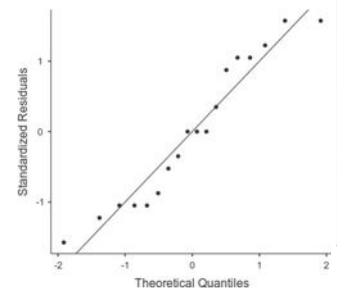


Homogeneity of Variances Test (Levene's)

F	df1	df2	p
0.206	5	12	0.9538

Normality Test (Shapiro-Wilk)

Statistic	p
0.929	0.1851



SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 60

The key assumptions of factorial ANOVA are homogeneity of variance, normality of the residuals, and independence of the observations. The latter, we have to ensure with our choice of research design, the first two assumptions can we check.

As in the case of the one-way ANOVA, normality of the residuals is checked by using the Shapiro-Wilk test and the Q-Q-plot of the residuals, the homogeneity of variances with Levene's test (just tick "Normality test" for Shapiro-Wilk, "Q-Q-plot", and "Homogeneity test" for Levene's).

Both the Shapiro-Wilk and Levene's test are not significant (i.e., our assumptions are not violated) and the residuals in the Q-Q-plot (right) don't deviate much from the main diagonal.

For the homogeneity of variances, there is also a visual way of assessing the assumption. In the Estimated marginal means plot (bottom-left), all error bars should have about the same length.



Factorial ANOVA in jamovi (Interact.)

effect sizes:

ANOVA - mood.gain								
	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p	ω^2
drug	3.453	2	1.727	31.714	<.0001	0.713	0.841	0.683
therapy	0.467	1	0.467	8.582	0.0126	0.096	0.417	0.084
drug * therapy	0.271	2	0.136	2.490	0.1246	0.056	0.293	0.033
Residuals	0.653	12	0.054					



We had a brief introduction into effect sizes earlier.

Within the ANOVA, we can ask for three effect sizes: η^2 (eta squared), η_p^2 (partial eta squared) and ω^2 (omega squared). Partial eta squared is not recommended. Why is explained on the next slide.

η^2 is possibly the most intuitive one, given that it can be interpreted in much the same way as R^2 in regression. It denotes the proportion of variance, the factor accounts for in relation to the total variance (or sum of squares which is variance multiplied by the number of participants): $\eta^2_{\text{drug}} = SS_{\text{drug}} / SS_{\text{total}}$. It ranges from 0 (no effect at all) to 1 (accounts for all of the variability in the outcome).

For the current analysis, we said earlier that $\eta^2_{\text{drug}} = 0.713$ is fairly substantial and practically relevant, $\eta^2_{\text{therapy}} = 0.096$ is a medium sized effect ($r = 0.3 \rightarrow r^2 = 0.09 \sim \eta^2_{\text{therapy}} = 0.096$).



Factorial ANOVA in jamovi (Interact.)

effect sizes:

ANOVA - mood.gain								
	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p	ω^2
drug	3.453	2	1.727	31.714	<.0001	0.713	0.841	0.683
therapy	0.467	1	0.467	8.582	0.0126	0.096	0.417	0.084
drug * therapy	0.271	2	0.136	2.490	0.1246	0.056	0.293	0.033
Residuals	0.653	12	0.054					



A main idea with partial η^2 (or η_p^2) is that, when measuring the effect size for a particular factor, to deliberately ignore the other effects in the model. η_p^2 «pretends» that the effect of all these other contributions is zero, and then calculates what the η^2 value would have been: $\eta_p^2 = SS_{\text{drug}} / (SS_{\text{drug}} + SS_R)$. The absolute value of η_p^2 always will give a larger number than η^2 since $SS_{\text{drug}} + SS_R$ will always be smaller than SS_{tot} . Like η^2 , partial η^2 varies between 0 and 1. It is fairly popular, likely due to the easy interpretation, the larger absolute value compared to η^2 , and the fact that SPSS provides it as (the only) effect size output.

A clear disadvantage is that partial η^2 “scales” with how many factors we include in our model (and hence, how much variance is left for the residuals). Thus, you might end up reporting an outdated value after a change to the model.



Factorial ANOVA in jamovi (Interact.)

effect sizes:

ANOVA - mood.gain								
	Sum of Squares	df	Mean Square	F	p	η^2	η^2_p	ω^2
drug	3.453	2	1.727	31.714	<.0001	0.713	0.841	0.683
therapy	0.467	1	0.467	8.582	0.0126	0.096	0.417	0.084
drug * therapy	0.271	2	0.136	2.490	0.1246	0.056	0.293	0.033
Residuals	0.653	12	0.054					



The total variance used for η^2 , in contrast, is a property of the data collected and remains unchanged (regardless of whether the model is changed). Finally, in the output, η^2 for all effects is $0.713 + 0.096 + 0.056 = 0.865$, whereas η_p^2 is $0.841 + 0.417 + 0.293 = 1.551$.

This has two undesirable consequences: (1) When using η_p^2 the different contributions (main effects and interactions) can add up to more than 1 which is rather counter-intuitive; it is definitely easier to imagine which proportion of variance is explained by a value that adds up to maximally 1 (= 100%). (2) η_p^2 also “blows up” small effects (from looking at η_p^2 for therapy, the value is rather misleading – 0.417 – appears as about half the size of η_p^2 for drug – 0.841 – which doesn’t reflect the contributions in terms of sum of squares: 0.467 vs. 3.453).

Taken together, the advice is to rather not use η_p^2 .



Factorial ANOVA in jamovi (Interact.)

effect sizes:

ANOVA

Search

ID

Dependent Variable

mood.gain

Fixed Factors

drug

therapy

Model Fit

Overall model test

Effect Size

η^2 partial η^2 ω^2


ANOVA - mood.gain								
	Sum of Squares	df	Mean Square	F	p	η^2	η^2p	ω^2
drug	3.453	2	1.727	31.714	<.0001	0.713	0.841	0.683
therapy	0.467	1	0.467	8.582	0.0126	0.096	0.417	0.084
drug * therapy	0.271	2	0.136	2.490	0.1246	0.056	0.293	0.033
Residuals	0.653	12	0.054					



η^2 also has a drawback because it is affected by the sample size and even though it is accurate for the sample variance, it overestimates the proportion of population variance explained.

For small samples, it is therefore better to choose an unbiased effect size measure such as ω^2 (omega squared). ω^2 has the same basic interpretation (proportion of the variance explained). It is an unbiased estimate of the population variance (and, given the corrections, always smaller than η^2).

This becomes especially obvious the combination of both a small sample size and a small effect size (where we are particularly likely to make an error in the estimation): Compared to η^2 for *drug * therapy* (0.056), ω^2 is about half-size (0.033). Please note that this argument was only for illustration given that we shouldn't assess the effect size when the *drug * therapy* interaction isn't significant.



Analysis of covariance (ANCOVA) in jamovi

Another “flavour” of ANOVA is used when you have a continuous variable that you believe might be related to the dependent variable (and that you wish to control for). This additional variable can be added to the analysis as a covariate, in the aptly named analysis of covariance (ANCOVA).

In an ANCOVA, the values of the dependent variable are “adjusted” for the influence of the covariate, and then the “adjusted” score means are tested between groups in the usual way. This technique can increase the precision of a model, and provide a more “powerful” test of the equality of group means in the dependent variable.

It is, however, required that the covariate is not confounded with any of our categorical variables (factors). If it were, we would diminish or cancel out effects in those categorical variables. An example is if we used age as covariate when the experimental groups differ in mean age.



ANCOVA in jamovi

ANCOVA

Dependent Variable: happiness

Fixed Factors: stress, commute

Covariates: age

Model Fit: Overall model test

Effect Size: η^2 partial η^2 ω^2

Assumption Checks: Homogeneity test, Normality test, Q-Q Plot

ANCOVA - happiness

	Sum of Squares	df	Mean Square	F	p	η^2	ω^2
stress	2751.517	1	2751.517	52.614	< .0001	0.403	0.393
commute	2213.928	1	2213.928	42.334	< .0001	0.324	0.314
age	334.351	1	334.351	6.393	0.0232	0.049	0.041
stress * commute	740.123	1	740.123	14.152	0.0019	0.108	0.100
Residuals	784.449	15	52.297				

Homogeneity of Variances Test (Levene's)

F	df1	df2	p
0.155	3	16	0.9249

Normality Test (Shapiro-Wilk)

Statistic	p
0.969	0.7354

Standardized Residuals vs Theoretical Quantiles plot

SEBASTIAN.JENTSCHKE@UIB.NO SLIDE 66

To conduct an ANCOVA in jamovi, we open the data set *ancova.omv* which contains four variables: *happiness*, our dependent variable, as well as two categorical (*stress – low vs. high*, *commute – cycling vs. not cycling*) and one continuous (*age*) independent variables.

We start the analysis with clicking on «ANOVA» in the icon bar and then choosing «ANCOVA». In the input window, we assign *happiness* to «Dependent variable», *stress* and *commute* to «Fixed factors» and *age* to «Covariates». We then open the drop-down-menu “Assumption checks” and check all three options.

They reveal that neither Homogeneity of Variances (Levene’s test) nor the Shapiro-Wilk test of Normality became significant, so we must not be concerned that our assumptions might be violated. The Q-Q-plot also gives no reason for concern.



ANCOVA in jamovi

assumption check II – interactions with covariate:

ANCOVA - happiness

	Sum of Squares	df	Mean Square	F	p	η^2	ω^2
stress	1.446	1	1.446	0.025	0.8766	0.001	-0.048
commute	110.932	1	110.932	1.931	0.1899	0.099	0.046
age	144.672	1	144.672	2.519	0.1385	0.130	0.074
stress * commute	11.433	1	11.433	0.199	0.6634	0.010	-0.039
stress * age	69.911	1	69.911	1.217	0.2916	0.063	0.011
commute * age	37.056	1	37.056	0.645	0.4375	0.033	-0.017
stress * commute * age	50.788	1	50.788	0.884	0.3656	0.046	-0.006
Residuals	689.310	12	57.442				

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 67

Finally, there is an additional assumption when we would like to include a covariate: the relationship between the covariate and the dependent variable should be similar for all levels of the independent variables / fixed factors.

This can be checked by opening the drop-down-menu “Model” selecting all variables in “Components” simultaneously and assigning them to “Model Terms” (using the upper arrow).

This makes our “ANCOVA – happiness” table twice as long as before. We go through the table and assess all interactions involving *age*. It reveals that neither interaction is significant: *stress * age* ($p = 0.292$), *commute * age* ($p = 0.438$), *stress * commute * age* ($p = 0.366$).

We can therefore remove these interactions from the model terms (but leave the main effect of *age* included).



ANCOVA in jamovi

ANCOVA

Dependent Variable: happiness

Fixed Factors: stress, commute

Covariates: age

Model Fit: Overall model test

Effect Size: η^2 partial η^2 ω^2

Assumption Checks:

- Homogeneity test
- Normality test
- Q-Q Plot

ANCOVA - happiness

	Sum of Squares	df	Mean Square	F	p	η^2	ω^2
stress	2751.517	1	2751.517	52.614	< .0001	0.403	0.393
commute	2213.928	1	2213.928	42.334	< .0001	0.324	0.314
age	334.351	1	334.351	6.393	0.0232	0.049	0.041
stress * commute	740.123	1	740.123	14.152	0.0019	0.108	0.100
Residuals	784.449	15	52.297				

Homogeneity of Variances Test (Levene's)

F	df1	df2	p
0.155	3	16	0.9249

Normality Test (Shapiro-Wilk)

Statistic	p
0.969	0.7354

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 68

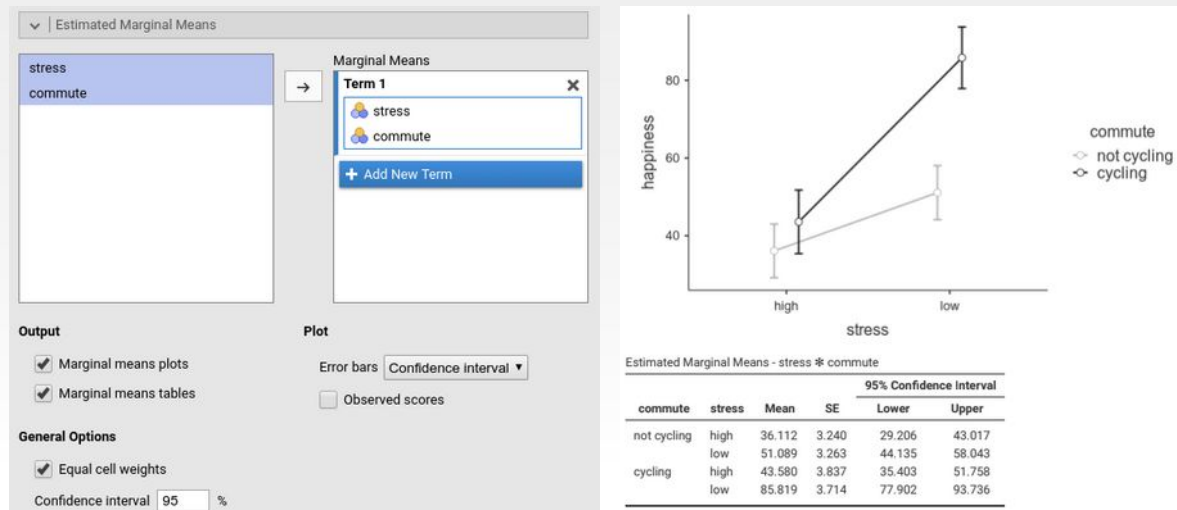


After we checked all these assumptions, we can turn to the table “ANCOVA – happiness” and assess the results. It appears as if the covariate *age* makes an contribution to predicting the dependent variable *happiness* ($F_{(1, 15)} = 6.39$; $p = 0.023$). The other main effects and their interaction are significant as well: *stress* ($F_{(1, 15)} = 52.61$; $p < 0.001$), *commute* ($F_{(1, 15)} = 42.33$; $p < 0.001$), *stress* \times *commute* ($F_{(1, 15)} = 14.15$; $p = 0.002$).



ANCOVA in jamovi

estimated marginal means:



We can use “Estimated Marginal Means” and tick both “Marginal means plots” and “Marginal means tables”. This allows us two things: First, we can visually assess the main effects and interactions of *stress* and *commute*.

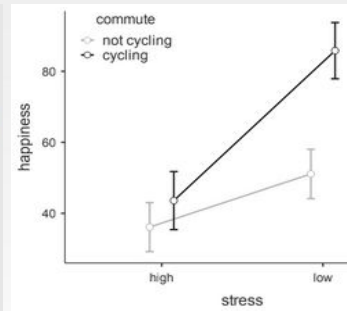
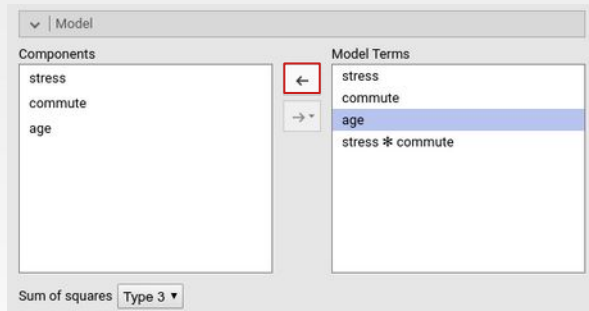
The significant interaction effect we obtained is reflected in almost no difference in happiness regardless of whether they are driving or cycling for people experiencing high levels of stress, whereas the difference (i.e., the advantage of cycling over driving) is large for people experiencing low stress levels.

In addition, there are main effects for both *stress* – people with low stress are, on average, happier than those with high stress – and for *commuting* behaviour – people who cycle are happier than those who drive to work.

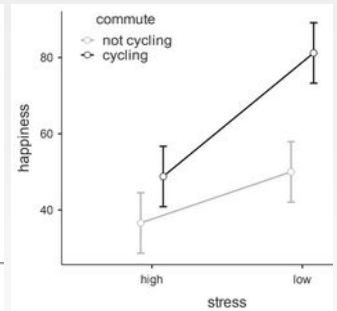


ANCOVA in jamovi

estimated marginal means II:



with age



without age



We can use the estimated marginal means also for assessing consequences of including the covariate in our ANOVA: The easiest way of adding or removing it is by opening the drop-down-menu “Model”, and moving age forth and back between “Model Terms” and “Components”.

When can then look at how the estimated marginal mean scores in the table “Estimated Marginal Means - commute * stress” are adjusted depending on whether the covariate *age* is included in or removed from “Model Terms”.

Alternatively, we could use the estimated marginal means plot. We need a bit of a keen eye to see that the happiness values for high stress differ less between cycling or not cycling when age is included while the gap widens if not.



ANCOVA in jamovi

estimated marginal means II:

Model

Components

- stress
- commute
- age

Model Terms

- stress
- commute
- age
- stress * commute

Sum of squares Type 3

Normality Test (Shapiro-Wilk)

Statistic	p
0.872	0.0126


ANCOVA - happiness							
	Sum of Squares	df	Mean Square	F	p	η^2	ω^2
stress	2751.517	1	2751.517	52.614	<.0001	0.403	0.393
commute	2213.928	1	2213.928	42.334	<.0001	0.324	0.314
age	334.351	1	334.351	6.393	0.0232	0.049	0.041
stress * commute	740.123	1	740.123	14.152	0.0019	0.108	0.100
Residuals	784.449	15	52.297				

ANCOVA - happiness							
	Sum of Squares	df	Mean Square	F	p	η^2	ω^2
stress	2622.050	1	2622.050	37.498	<.0001	0.401	0.386
commute	2354.450	1	2354.450	33.671	<.0001	0.360	0.345
stress * commute	451.250	1	451.250	6.453	0.0218	0.069	0.058
Residuals	1118.800	16	69.925				



We can also see this in the table with the main results “ANCOVA – happiness” where the interaction effect (*stress* × *commute*) gets stronger if *age* is included as covariate (*age* included: $F_{(1, 15)} = 14.15$; $p = 0.002$; *age* not included: $F_{(1, 16)} = 6.45$; $p = 0.022$).

There is another aspect to argue for including the covariate *age*: If it is not included, the Shapiro-Wilk test becomes significant, that is we have to conclude that our assumption of normality would be violated if we don't include *age* in our model.



Analysis of variance for repeated measurements (rmANOVA) in jamovi

Let's turn to our final bit. Quite often, we acquire multiple measurements from within one person. Two typical cases are: (1) that different experimental conditions are used within the same participant (usually as trials belonging to different conditions and presented in random order) or (2) that we have a pre-measurement followed by some treatment or intervention, a post-measurement and another measurement after some time to assess the stability of the intervention.

The repeated-measures ANOVA test is a statistical method of testing for significant differences between three or more levels of a factor applied to the same participant (or a participant that is closely matched with participants on another level of the experimental conditions). As a consequence, there should always be an equal number of measurements (data points) on each level.



rmANOVA in jamovi

Repeated Measures ANOVA

Participant

Repeated Measures Factors
Language skills
Speech
Conceptual
Syntax
Level 4

Repeated Measures Cells
Speech
Conceptual
Syntax

Between Subject Factors

Covariates

Effect Size
 Generalised η^2 η^2 Partial η^2

Dependent Variable Label
Scores

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2
Language skills	24.778	2	12.389	6.925	0.0130	0.414
Residual	17.889	10	1.789			

Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2
Residual	17.111	5	3.422			

Note. Type 3 Sums of Squares

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 73



This type of design and analysis can also be called a «within-subjects» design. Generally, the logic behind a repeated measures ANOVA is very similar to that of an independent ANOVA (sometimes called an independent ANOVA or «between-subjects» design).

Like in the standard or independent ANOVA, the total variability is still partitioned into between-groups variability (SS_b) and within-groups variability (SS_w), and after each is divided by the respective degrees of freedom to give MS_b and MS_w from which the F-ratio is calculated: $F = MS_b / MS_w$.



rmANOVA in jamovi

Repeated Measures ANOVA

Participant

Repeated Measures Factors
Language skills
Speech
Conceptual
Syntax
Level 4

Repeated Measures Cells
Speech
Conceptual
Syntax

Between Subject Factors

Covariates

Effect Size
 Generalised η^2 η^2 Partial η^2

Dependent Variable Label
Scores

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2
Language skills	24.778	2	12.389	6.925	0.0130	0.414
Residual	17.889	10	1.789			

Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2
Residual	17.111	5	3.422			

Note. Type 3 Sums of Squares

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 74



In a repeated measures ANOVA, the SS_w is divided into two parts: One part, the variability due to the individual differences between subjects (referred to as SS_{subjects}) is controlled for because we are using the same or a matched subject for each level of our independent variable / factor.

Only the other part of SS_w which is due to random variation, measurement errors, etc. remains. Using each participant as its own control and thereby removing or controlling for one source of variation in SS_w typically leads to the repeated-measurement ANOVA being more powerful. However, this does depend on whether the reduction in SS_w compensates for the reduction in degrees of freedom for the error term as they go from $(n - k)$ to $(n - 1) \cdot (k - 1)$.



rmANOVA in jamovi

Repeated Measures ANOVA

Participant

Repeated Measures Factors

Language skills

Speech
Conceptual
Syntax

Level 4

Repeated Measures Cells

Speech
Conceptual
Syntax

Between Subject Factors

Covariates

Effect Size

Generalised η^2 η^2 Partial η^2

Dependent Variable Label

Scores

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2
Language skills	24.778	2	12.389	6.925	0.0130	0.414
Residual	17.889	10	1.789			

Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2
Residual	17.111	5	3.422			

Note. Type 3 Sums of Squares

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 75



We use a dataset from a sample of patients with Broca's aphasia. For each patient, communication difficulties within three domains were assessed: «Speech» required patients to repeat single words read out aloud by the researcher. «Conceptual» required matching a series of pictures with their correct name. «Syntax» required bringing words within sentences into a syntactically correct order. To conduct the analysis, we open ANOVA → Repeated Measures ANOVA. Then, we assign a name to the factor denoted as RM Factor 1, e.g., «Language skills». Afterwards, we assign names to the different levels – it can be the same as our variable names («Speech», «Conceptual», «Syntax»). Finally, we assign our variables to the variable list called «Repeated Measurement Cells». Ensure that the variables are in the right place!



rmANOVA in jamovi

assumption checks:

Assumption Checks

Sphericity tests

Sphericity corrections

None Greenhouse-Geisser Huynh-Feldt

Homogeneity test

Tests of Sphericity		signif.?		GG $\epsilon > 0.75$?	
	Mauchly's W	p	Greenhouse-Geisser ϵ	Huynh-Feldt ϵ	
Language skills	0.849	0.7201	0.868	1.000	

Within Subjects Effects							
	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η^2
Language skills	None	24.778	2	12.389	6.925	0.0130	0.414
	Huynh-Feldt	24.778	2,000	12.389	6.925	0.0130	0.414
Residual	None	17.889	10	1.789			
	Huynh-Feldt	17.889	10,000	1.789			

Note. Type 3 Sums of Squares



In addition, we open the dropdown-menu «Assumption checks» and tick «Sphericity tests». In the output, we first assess Mauchly's Test of Sphericity. Conceptually, it is equivalent to the homogeneity of variances in the "classical" ANOVA (i.e., the one without repeated measurements). It tests the hypothesis that the variances on the different factor levels (i.e., «Speech», «Conceptual», and «Syntax») are equal. In our analysis, Mauchly's test is not significant ($p = 0.720$), and we can conclude that the variances are not significantly different. If they were, i.e., if Mauchly's test had been significant ($p < 0.05$) and there were differences in the variances at the different levels, we should apply a correction to the F-value.



rmANOVA in jamovi

assumption checks:

Assumption Checks

Sphericity tests

Sphericity corrections

None Greenhouse-Geisser Huynh-Feldt

Homogeneity test

Tests of Sphericity		signif.?		GG $\epsilon > 0.75?$		
	Mauchly's W	p	Greenhouse-Geisser ϵ	Huynh-Feldt ϵ		
Language skills	0.849	0.7201	0.868	1.000		

Within Subjects Effects							
	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η^2
Language skills	None	24.778	2	12.389	6.925	0.0130	0.414
	Huynh-Feldt	24.778	2,000	12.389	6.925	0.0130	0.414
Residual	None	17.889	10	1.789			
	Huynh-Feldt	17.889	10,000	1.789			

Note. Type 3 Sums of Squares



We could do this by ticking either «Greenhouse-Geisser» or «Huynh-Feldt» under «Sphericity correction» in «Assumption checks».

«Huynh-Feldt» is chosen if the Greenhouse-Geisser value in the table «Tests of Sphericity» is larger than 0.75, otherwise «Greenhouse-Geisser» has to be used.

If we were to apply a correction for our dataset, we would have to tick «Huynh-Feldt» since the «Greenhouse-Geisser» value in the table «Tests of Sphericity» is 0.868 (which is larger than 0.75).

Since the p-value for Mauchly's test was relatively high (i.e., far from significance), we can see that the applied corrections would have been so minor that they are hardly detectable.

However, this was just for demonstration: The Mauchly's Test of Sphericity was not significant ($p = 0.720$): No correction is required and we can leave the tick at «Sphericity corrections» → «None».



rmANOVA in jamovi

Repeated Measures ANOVA

Participant

Repeated Measures Factors
Language skills

Speech
Conceptual
Syntax
Level 4

Repeated Measures Cells

Speech
Conceptual
Syntax

Between Subject Factors

Covariates

Effect Size
 Generalised η^2 η^2 Partial η^2

Dependent Variable Label
Scores

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2
Language skills	24.778	2	12.389	6.925	0.0130	0.414
Residual	17.889	10	1.789			

Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2
Residual	17.111	5	3.422			

Note. Type 3 Sums of Squares

ANALYSIS OF VARIANCE

SEBASTIAN.JENTSCHKE@UIB.NO

SLIDE 78



Finally, we can assess our results. We obtain a $F_{(2,10)} = 6.93$ and an $p = 0.013$ for the repeated measure «*Language skills*» and can conclude that the performance in each language task did vary significantly.



rmANOVA in jamovi

effect sizes:

Effect Size **Dependent Variable Label**

Generalised η^2 η^2 Partial η^2

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2
Language skills	24.778	2	12.389	6.925	0.0130	0.414
Residual	17.889	10	1.789			

Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2
Residual	17.111	5	3.422			

Note. Type 3 Sums of Squares

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G	η^2	η^2_P
Charisma	23233.600	2	11616.800	77.525	<.0001	0.593	0.363	0.803
Residual	5694.178	38	149.847					
Attractivity	20779.633	2	10389.817	81.796	<.0001	0.566	0.325	0.812
Residual	4826.811	38	127.021					
Charisma * Attractivity	4055.267	4	1013.817	16.526	<.0001	0.203	0.063	0.465
Residual	4662.289	76	61.346					

Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G	η^2	η^2_P
Residual	760.422	19	40.022					

Note. Type 3 Sums of Squares



For the repeated measurement ANOVA, we have the choice of three effect sizes. I had to use a different data set on the right hand side because for the case of only one repeated-measurements factor (language skills in our current data set) all three reveal identical results.

The three effect sizes differ though in how they are calculated, more specifically by which value the square sum of our effect of interest is divided. I will go through each of them and briefly explain their calculation. As I argued more comprehensively on an earlier slide, η^2 is the most solid among the choices.

η^2 , at the same time, is also the one that is lowest. In exchange, it is the one to be understood most easily and the least flawed one. It is calculated by dividing the sum of squares for the effect by the total sum of squares.



rmANOVA in jamovi

effect sizes:

Effect Size Dependent Variable Label

Generalised η^2
 η^2
 Partial η^2
Scores

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2
Language skills	24.778	2	12.389	6.925	0.0130	0.414
Residual	17.889	10	1.789			

Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2
Residual	17.111	5	3.422			

Note. Type 3 Sums of Squares

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G	η^2	η^2_P
Charisma	23233.600	2	11616.800	77.525	<.0001	0.593	0.363	0.803
Residual	5694.178	38	149.847					
Attractivity	20779.633	2	10389.817	81.796	<.0001	0.566	0.325	0.812
Residual	4826.811	38	127.021					
Charisma * Attractivity	4055.267	4	1013.817	16.526	<.0001	0.203	0.063	0.465
Residual	4662.289	76	61.346					

Note. Type 3 Sums of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G	η^2	η^2_P
Residual	760.422	19	40.022					

Note. Type 3 Sums of Squares



The generalized η^2 (η^2_G) divides the sum of squares for the effect by the sums of squares for this effect plus all residuals. As a consequence, the individual effect sizes can add up to a value larger than one, so it is not very intuitive.



rmANOVA in jamovi

effect sizes:

Effect Size		Dependent Variable Label	
<input type="checkbox"/> Generalised η^2	<input checked="" type="checkbox"/> η^2	<input type="checkbox"/> Partial η^2	Scores

	Sum of Squares	df	Mean Square	F	p	η^2_G	η^2	η^2_P
Charisma	23233.600	2	11616.800	77.525	<.0001	0.593	0.363	0.803
Residual	5694.178	38	149.847					
Attractivity	20779.633	2	10389.817	81.796	<.0001	0.566	0.325	0.812
Residual	4826.811	38	127.021					
Charisma * Attractivity	4055.267	4	1013.817	16.526	<.0001	0.203	0.063	0.465
Residual	4662.289	76	61.346					

Note. Type 3 Sums of Squares

	Sum of Squares	df	Mean Square	F	p	η^2_G	η^2	η^2_P
Residual	760.422	19	40.022					

Note. Type 3 Sums of Squares



Partial η^2 is actually the worst among the three as it only considers the residuals for the effect itself. It is calculated by dividing the square sum of the effect by the square sums of the effect plus the residuals of the effect. Among the three effect size measures, it is the largest (hence it's popularity). Like the generalized η^2 , it can add up to values larger than one which is difficult to interpret.

All three effect sizes are the same with regards to the numerator (i.e., the part above the fraction bar):

This is always the sum of squares for the effect.

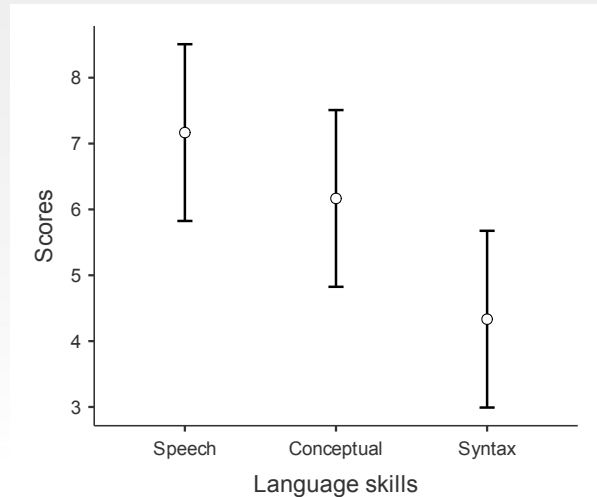
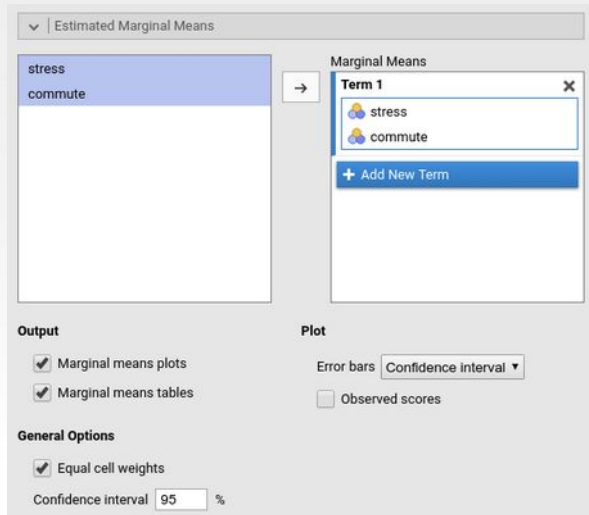
They differ, though, with respect to the denominator (the part below the fraction bar): For η^2 it is the total variance (i.e., the total amount of variation; therefore, the added effect sizes can be maximally 1).

Generalized η^2 takes into consideration the sum of squares of the effect plus all variation that we can't explain with our model. Partial η^2 only considers the sum of squares plus the residuals of that effect.



rmANOVA in jamovi

estimated marginal means:



Using the drop-down-menu «Estimated Marginal Means», we can ask for either descriptive statistics (in tables) or a plot showing the performance in the different tasks.

The plot demonstrates that the performance is highest for repeating the word, lower for matching the names of the words to the pictures and lowest for arranging words in accordance with their syntactically correct order in a sentence.



rmANOVA in jamovi

post-hoc tests:

Post Hoc Tests

Language skills

Corrections

- No correction
- Tukey
- Scheffe
- Bonferroni
- Holm

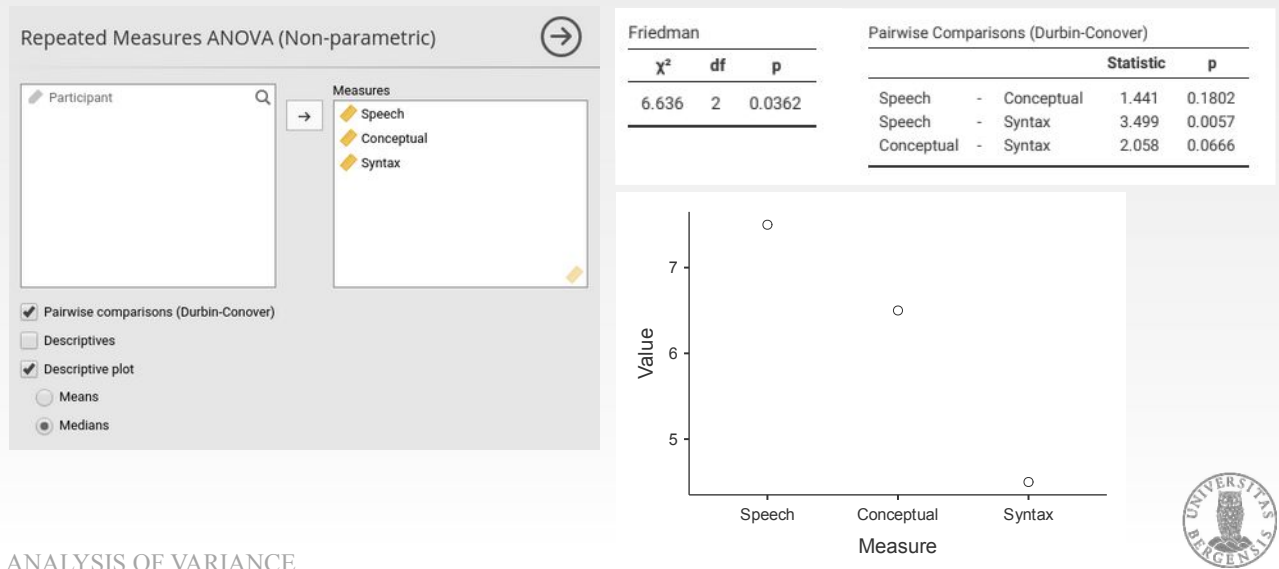


Like for the independent ANOVA, we can use post-hoc tests to assess the differences between the different levels. We open the drop-down-menu «Post-hoc tests», assign «Language skills» to the variable box and tick «Holm» as correction for multiple comparisons («Tukey» is already ticked as default).

We obtain no significant differences. This indicates that whereas the pooled variance over all three stages of our factor is significant, the one-by-one differences between those stages are not. As we saw on the previous slide, there is a linear trend with a falling performance from the first (speech) to the last stage (syntax) of our factor, that drives the significance for the whole factor whereas the error bars on those stages are still overlapping.



Non-param. rmANOVA in jamovi



In case one or more of our assumptions are not met, there is a non-parametric equivalent to the repeated-measures ANOVA (with one factor), called Friedman test.

It reveals the same results as its parametric counterpart: The whole model ($\chi^2 = 6.64$; $df = 2$; $p = 0.036$) being significant, mainly driven by the difference in performance between «Speech» and «Syntax» (Durbin-Conover value = 3.499, $p = 0.006$) with the two other comparisons not being significant. This is a difference to the parametric ANOVA before where this difference was close to but did not reach significance.

It can, however, not be used with more complex designs (i.e., two or more factors). In case, the factors have to be broken down and be tested individually.



Summary and literature

You (almost) made it through the lecture!



Summary

- introduction
- history and some mathematical background
- ANOVA with one factor in jamovi
- ANOVA with more than one factor in jamovi
- ANCOVA in jamovi
- ANOVA for repeated-measurements in jamovi
- for all: assumption checks (normality and variance homogeneity), effect sizes, and post-hoc tests



Let's briefly summarize it. We started with embedding the ANOVA within the context of other methods (using categorical vs. continuous variables, relation vs. difference hypotheses, and within-subject vs. between-subjects type of variables and which type of ANOVA to use for each).

That was followed by a bit of history and a by-hand calculation of an ANOVA with one factor.

Afterwards we turned to ANOVAs with one or several factors, the ANCOVA and the ANOVA for repeated measurements. If there were non-parametric alternatives (as for the independent ANOVA and the repeated measures ANOVA with one factor), we also introduced those. Each of the four parts also contained a discussion of assumption checks, effect sizes and post-hoc tests.



Literature

Navarro, D. J., & Foxcroft, D. R. (2022). *Learning statistics with jamovi*.
<https://doi.org/10.24384/hgc3-7p15> (Ch. 13 & 14,
p. 327 – 418) – **Essential**

Aron, A., Coups, E. J., & Aron, E. (2013). *Statistics for psychology* (6th ed). Pearson. (Ch. 9, 10, 16,
17; p. 349 – 486, 672 - 716) – **Recommended**



The literature that was covered in the lecture was chapter 13 and 14 from the jamovi-book (Navarro & Foxcroft, 2022).

If you wish an alternative account or would like a stronger focus on SPSS, you can have a look at the chapters 9, 10, 16 and 17 in Aron, Aron & Coups (2013).

Finally, there are books that really take you to the depth of what can be explored using ANOVAs and the mathematical background. One recommendation, available on Oria, is:

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed). Sage Publications.



**Thanks for bearing
with me...**



UNIVERSITY OF BERGEN

