# QNM Lecture Notes

ANOVA, Regression, Introduction to MANOVA and GLM

Course on *Advanced Quantitative Methods for Human-Computer Interaction*

Utrecht University

**Björn van Zwol**

July 21, 2024

# Preface

## Recommended Resources

- *Learning statistics with R* [1]. Lecture Notes by Danielle Navarro, available for free online at https://learningstatisticswithr.com/

Lecture notes with many approachable explanations. They are aimed at (psychology) undergraduates, but are a useful reference for brushing up on basic statistics/additional background reading. They do not cover some of the slightly more advanced topics that our course does discuss, such as RBDs, repeated measures, ANCOVA and multivariate statistics.

- *Using Multivariate Statistics* [2]. Book by Barbara G. Tabachnick and Linda S. Fidell.

A more advanced textbook aimed at advanced undergraduates/graduate students. Very comprehensive and a great general reference.

- *NIST/SEMATECH e-Handbook of Statistical Methods* [3], available for free online at http://www.itl.nist.gov/div898/handbook/.

An engineering perspective. It provides nice and succinct explanations for technical concepts, in particular for ANOVA.

- *Statistical Inference* [4]. Book by George Casella and Roger L. Berger.

The most formal and rigorous in this list: a well-structured mathematics book of the more approachable sort, with many nice explanations.

- *Doing Better Statistics in Human-Computer Interaction* [5]. Book by Paul Cairns.
- *Statistics for HCI* [6]. Book by Alan Dix.

Two nice books that cover statistics from an explicit HCI perspective.

## Acknowledgements

These lecture notes are far from original. Paraphrasing David Tong who said it best[1]: my primary contribution has been to combine the best discussions and explanations I could find from the vast literature on the subject. I have relied especially on [2] and [1] as main sources.

I am also extremely thankful to my advisor Egon van den Broek for his general support and faith in my ability to teach statistics. I'm also hugely thankful to Julie Pivin-Bachler and Lukas Arts for their support and feedback. Finally, of course, a big thanks to all students for providing invaluable feedback!

For errata, please e-mail bjornvanzwol [at] gmail.com.

[1] http://www.damtp.cam.ac.uk/user/tong/teaching.html

# Contents

# Chapter 1

# Preliminaries

A key aspect of quantitative methods is knowing how to choose an appropriate statistical analysis for your situation. In this chapter we describe some important aspects to consider in making this choice.

## Further Reading for This Chapter

[2]: Ch. 1 (esp. 1.1.1, 1.2.4), 2, 3.1.

[1]: Ch. 1 and 2 (useful background reading).

### Types of Questions

Quantitative methods help answer specific research questions. For example:

- Do groups of $X$ have a significant difference in means on the measure $Y$?
- Can we predict $Y$ from $X$?
- Are $X$ and $Y$ independent?
- How strongly are $X$ and $Y$ related?

If this seems a bit abstract, we shall come with many examples later in these notes when we discuss specific techniques.

Key aspects for choosing an appropriate statistical analysis: (1) the *type of research question*.

### Dependent and Independent Variables

In the questions above, $X$ and $Y$ are the variables. In statistics, one talks about three basic types of variables: independent variables (IVs), dependent variables (DVs), and covariates (CVs). We discuss the first two and leave the third for later when it will be easier to understand.

Say we are measuring the IQ of Ravenclaw students and Gryffindor students, and want to know whether there is a significant difference. If true, the IQ *depends on* the house of the student. Hence, it's called the *dependent variable* of our experiment. This implies there is an *independent variable* (which does not depend on other variables): the house of the student.

Thus, the dependent variable is what you are interested in measuring. Hence, it is also called *measure* or *output*. The independent variable (IV) defines the groups that you study (in ANOVA-type), or predictors of the thing you want to predict (in regression-type). In this sense, they are a type of *input*. Hence, using typical math-notation, we use $X$ for IVs and $Y$ for DVs.

In our example, the IV (house of student) had two *levels*: Ravenclaw and Gryffindor. If we included Hufflepuff and Slytherin we would have 4 levels. Hence, the variable is *categorical* or *discrete*: they can only take certain specified values. In contrast, the DV was continuous variable: IQ can take any value. When a variable is continuous, one does not speak of levels.

This is important, because:

> Key aspects for choosing an appropriate statistical analysis: (2) the *variables*.
>
> Specifically:
>
> - the *role* of variables (DV, IV, CV or equal footing)
> - the *type* of variables (discrete or continuous),
> - the *number* of variables,
> - *the number of levels* for the variables (when discrete variables are involved)

To illustrate, we list some examples which will become clear later. If you have 1 DV, you have univariate statistics. For univariate statistics, if you have 1 IV, you do a t-test (if the IV has 2 levels) or ANOVA (if it has $> 2$ levels). If you have several IVs, you get factorial ANOVA. If you have more than one DV, you are doing multivariate statistics. And so on.

Does one always have to have dependent or independent variables? No. Sometimes you want to treat variables *on an equal footing*: e.g. when you want to know if two variables are independent ($\chi^2$ independence test) or to what extent they are related (correlation). For these cases, there is no 'input' and 'output'.

## Variance

The concept of *variance* is key to understanding ANOVA (*analysis of variance*) and more advanced methods, like ANCOVA (analys of *covariance*) and MANOVA (*multivariate* analysis of variance). Variance, also called *spread*, of a variable, simply means "how much it varies". Given a $N$ measurements of a variable $x$ giving values $x_i$, where $i = 1, 2, ..., N$ and a mean $\bar{x}$, the *sample variance* is[1]:

$$\mathrm{var}(x) = s^2(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{1.1}$$

Here, $s$ is the sample *standard deviation* (sometimes also denoted SD), which thus simply is the square root of the sample variance:

$$s(x) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{1.2}$$

Finally, the standard deviation *of the mean*, also called the *standard error* (sometimes denoted as SE or $\mathrm{SE}_M$), is:

$$\mathrm{SE} = \frac{s}{\sqrt{N}}$$

You might be wondering where this square root of $N$ comes from. There is a lot to say about this[2], but we shall need to keep it brief. Roughly speaking, we can understand it by the observation that *a mean becomes more accurate if you take more measurements* (larger sample sizes). Thus, the *distribution* of means will become more narrow. This is in contrast with the distribution of true values, for which the variance should be *independent* of sample size.

> **NOTE:** you will also see another formula for variance and standard deviation: $\mathrm{var}(x) = s^2(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})$ with $1/N$ instead of $1/(N-1)$. This is the formula that is used when you are

[1] Note that the notation used here is different from [1].
[2] See e.g. Ch. 10, esp. 10.3 in [1].

certain that you have the entire *population*, and you are not considering a *sample*.

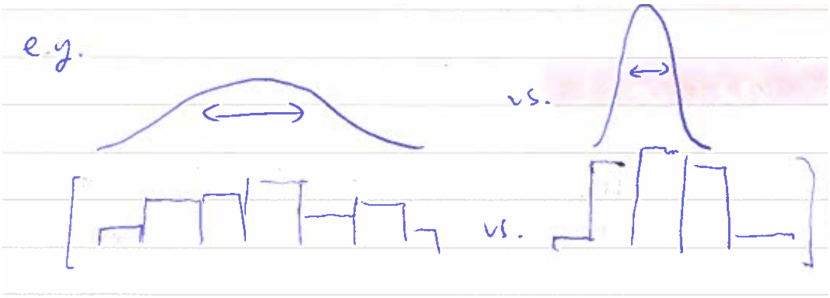



*Figure 1.1: Illustration of variance. Large variance (left) vs. small variance (right). The top row shows a theoretical normal distribution, and the bottom illustrates some collected data displayed as a histrogram.*

## Covariance

*Covariance* generalizes variance to two variables, and says how much two variable "covary". It is most easily understood using math. Given two samples $x_i$ and $y_i$ and means $\bar{x}$ and $\bar{y}$, the covariance is:

$$\text{cov}(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

Note again that we use the formula for a sample. How can we understand this? Consider some numbers:

- cov $> 0$ means that if $x$ goes up, so does $y$ on average, and vice versa.
- cov $< 0$ means that if $x$ goes up, $y$ goes down on average, and vice versa.
- cov $= 0$ means that if $x$ goes up, $y$ can do anything on average.

Also note that $\text{cov}(x, x) = \text{var}(x)$, i.e. the covariance of a variable with itself is just its variance.

## Correlation

If two variables are strongly *correlated*, that is a statistical way of saying that they 'seem related'. Of course, correlation does necessarily imply an *actual* or *meaningful* relation![3] It only means there is a *statistical* relationship.

Correlation is very close to covariance, but slightly different. There are also different types of correlation, but the one used overwhelmingly is *Pearson's correlation*, usually simply called correlation. In math:

$$r = r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}} = \frac{\text{cov}(x, y)}{s(x)s(y)}$$

I.e. a correlation is just a normalized (scale-independent) covariance. Thus, correlations take values between -1 and 1 instead of $-\infty$ and $+\infty$ like covariance. Note that for correlations, $X$ and $Y$ are not IVs and DVs, but treated on an equal footing.

Recall from Lecture 1 that t-tests had a measure of *effect size*. For the one-sample t-test one had Cohen's $d$ measure, which tells us how meaningful a difference between groups is. Similarly, correlations have a measure of effect size: $r^2$ is the *effect size* of a correlation. We discuss effect size and shared variance in more detail below.

[3] See https://www.tylervigen.com/spurious-correlations.

## Effect size

As an example, imagine we study the relation between people's income $I$ and the number of years in education $E$ they have experienced. We can draw the variance in these variables as circles, where the areas represent the amount of variance in these variables. Note that these areas are not meant to be interpreted quantitatively. Now, if the $E$ and $I$ are correlated, they have *shared variance*. This is illustrated below.

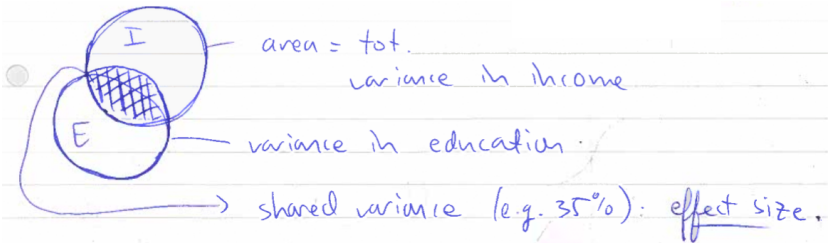



*Figure 1.2: Illustration of shared variance/effect size.*

Now if we would consider instead of correlation e.g. a t-test or regression, we can view $I$ as the DV and $E$ as the IV. In this view, we can say that the variance in income can be *explained by* the variance in education. Statements like this are often found when statistics is used, and thus has a precise meaning: *how much* variance is explained is quantified by the effect size.

Now imagine we include a second IV, job prestige $P$. This variable can be added to the previous figure, such that we might get something like Fig. 1.3. However, this figure is unrealistic, since job prestige and number of years of education are likely to be correlated as well. This leads us to *covariates*.

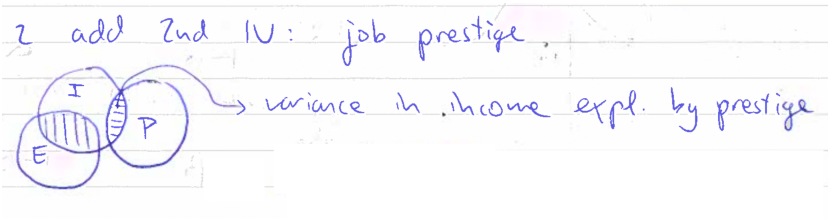



*Figure 1.3: Illustration of shared variance with 2 IVs and 1 DV. Unrealistic, since job prestige and number of years of education are likely to be correlated.*

## Covariates

An updated illustration is given in Fig. 1.4. This is more realistic, but also makes the analysis more complicated. It requires a decision on how to deal with the variance *between IVs*. For instance, one might want to study the effect of $E$ on $I$ while *controlling/adjusting* for $P$. This means we look exclusively at the effect of one IV, and consider $P$ as a *covariate* (CV) instead. How to do this correctly, depends on the research question:

> Key aspects for choosing an appropriate statistical analysis: (3) *presence of covariates and choices on how to deal with them.*

Options include *randomized block designs, ANCOVA* and *sequential regression*. These will be discussed in later lectures.

## NHST revisited[4]

Null hypothesis significance testing (NHST) can seem a strange or abstract when doing it for the first time. What are we actually *doing*? One way of looking at it is that we are effectively comparing two

[4]This section and the next are an extended version of sections 3.1.1 and 3.1.2 in [2].

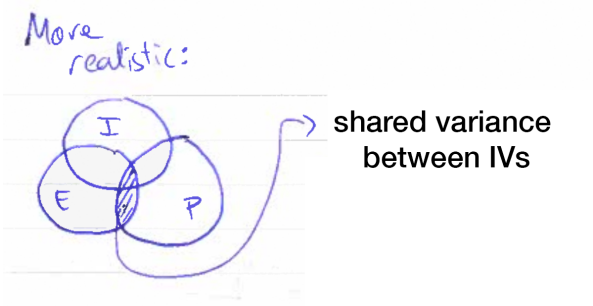



*Figure 1.4: More realistic illustration of shared variance. Shared variance between IVs requires deciding on how to deal with this variance.*

'alternate realities'. The null hypothesis $H_0$, represents the reality we assume we live in. $H_a$ is an 'alternate reality'. The p-value that we get from NHST is the *probability* that we live in $H_0$. If this probability is very small, we are more likely to live in the $H_a$ reality instead.

The distribution under the null hypothesis shown in the top row in Fig. 1.5. It represents the *probability* of some *mean IQ* being found under the null hypothesis.[5] E.g., if we find a mean IQ of 120, the *p-value* would be the area of the curve to the right of 120.

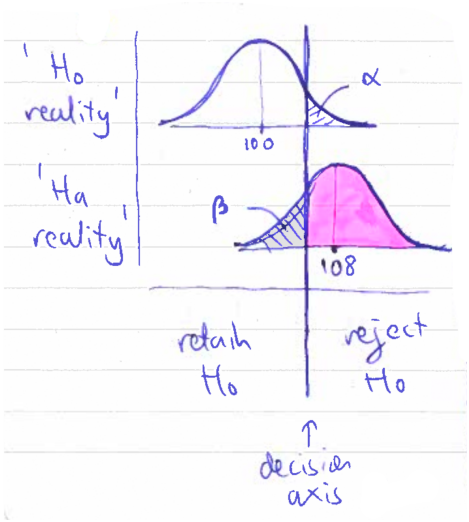



*Figure 1.5: Alternate realities in NHST.*

Now, with this p-value, we need to decide whether it is 'small enough' to reject $H_0$. For this we use the *significance level* $\alpha$. (Recall that is simply a threshold which we *define* to be what we mean by 'small enough'. Typically, for historical reasons, we use a 1/20 probability, i.e. 0.05) Thus, if the p-value (the probability of finding our calculated mean assuming a normal distribution) is smaller than 0.05, we *reject* $H_0$ and say we have a significant result.

Now, you might think: we never calculated areas under a curve in previous lectures, did we? That's true. Since calculating these probabilities by hand is a lot of work, *critical value tables* provide a shortcut[6]. Instead of calculating a p-value for your calculated mean each time, we reverse the process: for several probabilities $\alpha$, these tables show *how large your mean at least needs to be*, in order to be smaller than that probability. Well... almost: since it is not convenient to have a different table for each different mean, we first *subtract* the expected mean from our calculated mean, such that our

[5] An important distinction here is between the *sampling distribution* and the *sampling distribution of means*. Hypotheses are always tested about means, not about individual scores, which is why we are always using the latter. Sampling distributions of means have smaller standard deviations, since one divides by $\sqrt{N}$ to obtain the standard deviation of the mean. See section 3.1.1 in [2] and Chapter 10 in [1].

[6] At least for calculations using pen and paper – if a computer is used, you will be given a p-value automatically.

distribution is centered around zero. We also adjust for the standard deviation. This gives the *test statistic*, and critical value tables show how large your test statistic should at least be (i.e. *critical values*) for a given significance level.

## Error types and statistical power

Look again at Fig. 1.5. The top represents the '$H_0$ reality', with mean 100. The bottom row is the '$H_a$ reality', with a distribution of IQs, shifted 8 IQ points to the right, centered instead around $\mu = 108$. We can see how the critical value, which was determined by $\alpha$, acts as a type of *decision axis*. If we calculate our test statistic to be larger than the critical value, we are on the right side of the axis, and we reject $H_0$. And if smaller, we are on the left, and we retain $H_0$.

We have continued the decision axis vertically downwards to the distribution in the $H_a$ reality. We can see how this axis splits up this distribution in two. We call the area on the left of this axis $\beta$. The probability on the right, then, is $1 - \beta$. We see that we have obtained 4 'quadrants'. These are shown again in the table in Fig. 1.6: they represent the four different things can happen with NHST, with corresponding probabilities. In the top left, $H_0$ is true and we retain it: this is a correct decision. In the bottom right, we reject $H_0$ when it is false: again a correct decision. In the other quadrants, we have made an *error*! A *type I error* means we reject $H_0$ even though it is actually true, which happens with probability $\alpha$. A *type II error* means we retain $H_0$ although it is false; this happens with probability $\beta$. These two error types are foundational to statistical practice.

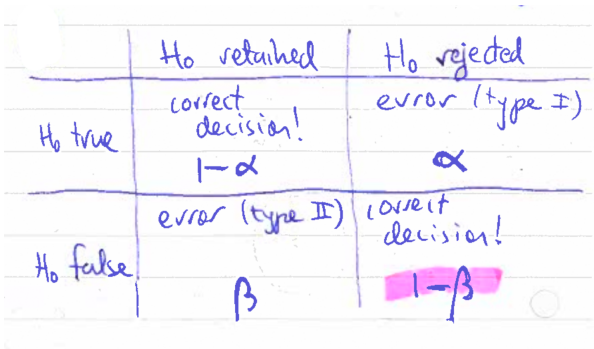



*Figure 1.6: Four possible outcomes of NHST.*

Clearly, we prefer not making errors. Hence, we would like to maximize the probabilities of the top left and bottom right quadrants, i.e. $1 - \alpha$ and $1 - \beta$. The former is mostly out of our control, since it is determined by convention and journal requirements. In contrast, $1 - \beta$ we have some control over. It is called the *statistical power*, which we would like to be large.

How could we increase this statistical power? Looking at Fig. 1.5 we see that $1 - \beta$ would increase if we made the distribution more *narrow*. This can be done in general by *increasing one's test statistic*. For our simple examples, one could try to *decrease the variance of the sample* somehow (which might be hard to do however without changing the research question), or by *increasing the sample size*. We close by quoting an important note from [2] (emphasis ours):

> There is occasionally the danger of *too much power*. The null hypothesis is probably never exactly true and any sample is likely to be slightly different from the population value. With a large enough sample, rejection of $H_0$ is virtually certain. Hence, a *"minimal meaningful difference"* and acceptable effect size should guide the selection of sample size (Kirk, 1995). The sample size should be large enough to be likely to reveal a minimal meaningful difference.

## Assumptions

As a final preliminary note we emphasize again that statistical analyses rely on assumptions, which should be checked using appropriate tests.

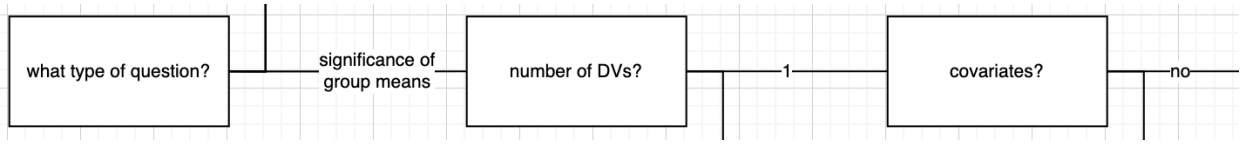



*Figure 1.7: Start of path towards the one-sample t-test (left to right).*

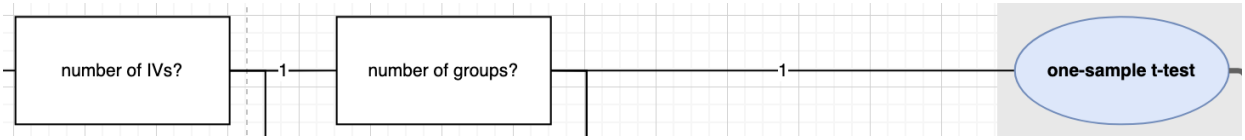



*Figure 1.8: Continued path towards one-sample t-test (left to right).*

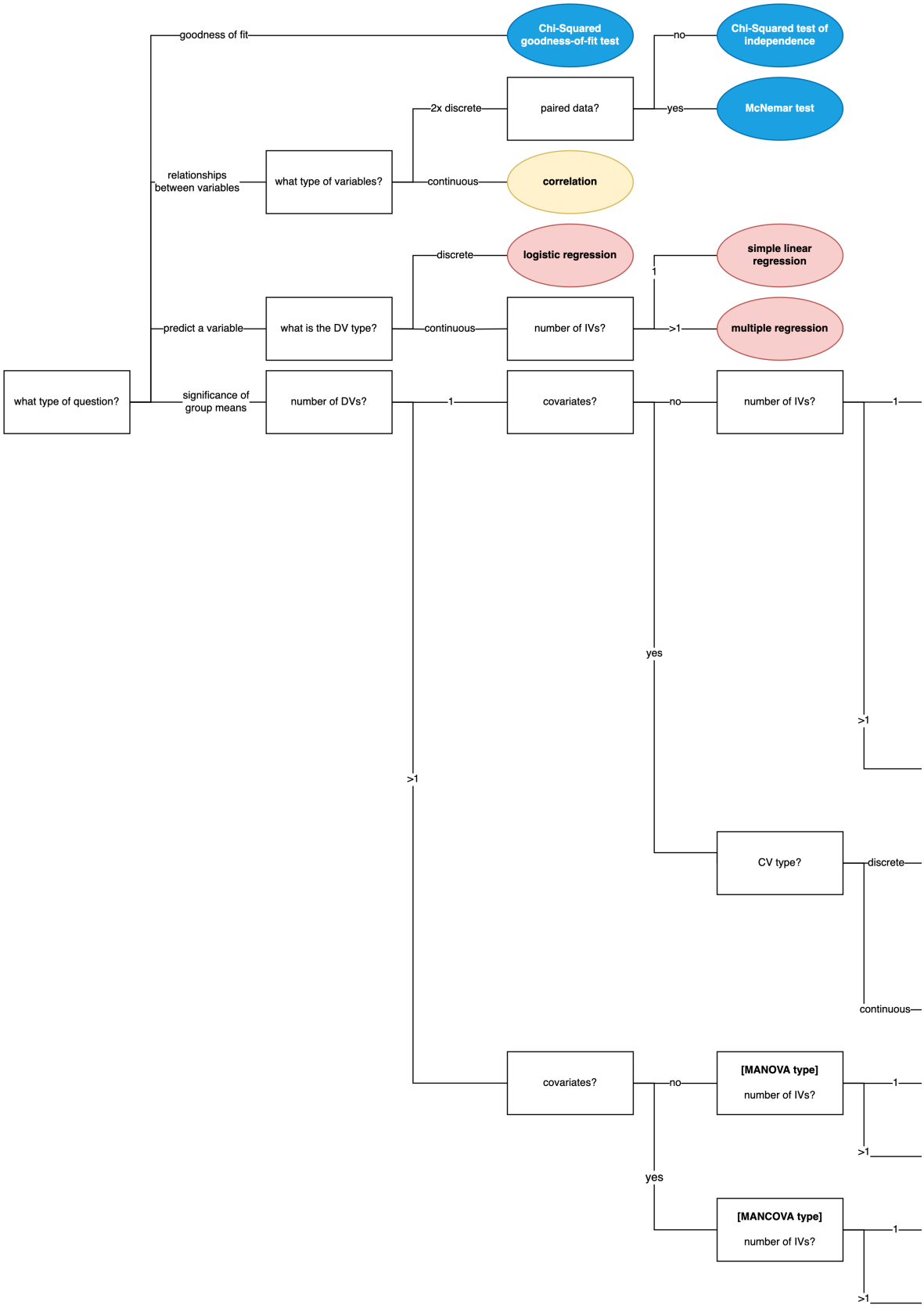
Key aspects for choosing an appropriate statistical analysis: (5) *assumptions*.

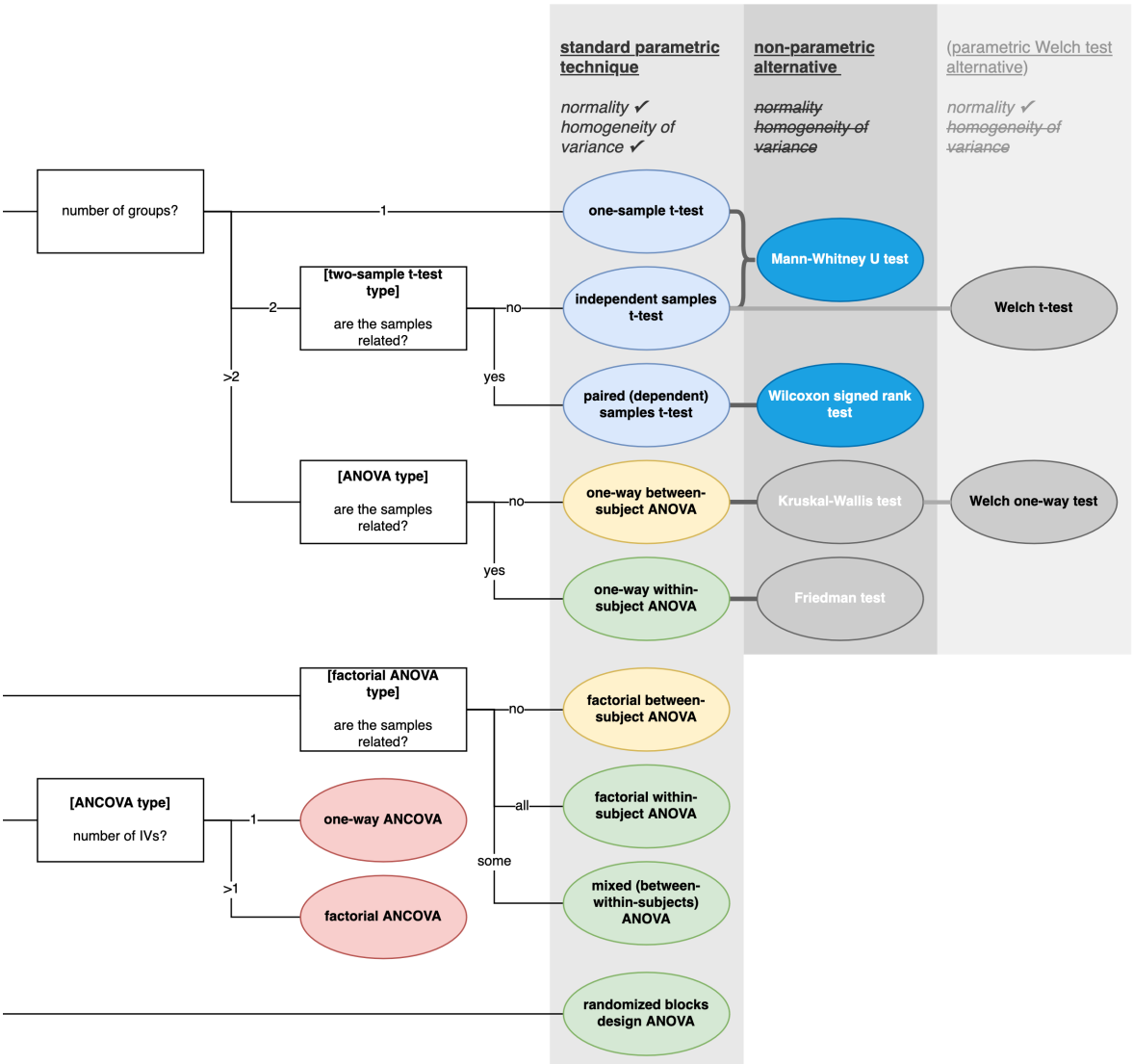
E.g. if normality does not hold, one could consider a non-parametric test over a standard parametric technique.

## Test overview

The next page shows an overview of most of the techniques (and some additional ones) covered in the course, structured as a tree. Techniques in the tree that we have already are the blue ellipses: the t-tests, their non-parametric alternatives (Mann-Whitney U test and Wilcoxon signed rank test) and the Chi-Squared tests. This Chapter, 2.1, 2.2 discuss the yellow ellipses: correlation, one-way ANOVA and factorial ANOVA. Section 2.3 and Section 2.4.1 discuss randomized blocks designs (RBDs), repeated measures/within-subjects ANOVA, and combiantionsl: the green ellipses. Chapter **??** then discusses regression, including ANCOVA. Finally, purple ellipses discuss MANOVA, covered in Chapter 4.

As an example, Figs. 1.7 and 1.8 show the path in the tree to reach the one-sample t-test. It can be seen that the *aspects* for determining the appropriate statistical analysis are what define the path in the tree.


what type of question?
goodness of fit
Chi-Squared goodness-of-fit test
relationships between variables
what type of variables?
2x discrete
paired data?
no
Chi-Squared test of independence
yes
McNemar test
continuous
correlation
predict a variable
what is the DV type?
discrete
logistic regression
continuous
number of IVs?
1
simple linear regression
>1
multiple regression
significance of group means
number of DVs?
1
covariates?
no
number of IVs?
1
yes
>1
CV type?
discrete
continuous
>1
covariates?
no
[MANOVA type]
number of IVs?
1
>1
yes
[MANCOVA type]
number of IVs?
1
>1




LEGEND
- The colors correspond to different lectures
- Black letters: parametric technique. White letters: non-parametric technique.
- Greyed out ellipses represent topics that are conceptually similar to the connected ellipses, but not discussed in our course.

NOTES
- Note that several techniques, especially more advanced ones, are not included in this tree. See e.g. Ch. 2 in Tabachnick & Fidell for a more complete tree (including e.g. mixed models, multilevel models, canonical correlation).
- 'Repeated measures' (RM) is 'within-subject' when the measures are taken on one subject. The same techniques can also be used in a '(matched-)randomized blocks design' (RBD).
- 'Factorial' is also called 'two-way' when there are 2 IVs, N-way when there are N IVs.
- 'Mixed' designs are also called 'split-plot' ANOVA. Note that mixed designs are not the same as mixed models, which refer to the combination of fixed and random effects (not covered in our course).
- Note that it is often true that several techniques can be applied in a single situation. This tree should not be taken as the only source in determining your analysis.

# Chapter 2

# ANOVA

### Further reading for this chapter

[1]: Ch. 14 (one-way ANOVA), Ch. 16 (factorial ANOVA).

[2]: 3.2 (up to and including 3.2.5).

[3]: 3.2.3, 7.4.3 (one-way, two-way ANOVA), 7.4.4 (fixed/random effects)

## What is ANOVA?

ANOVA stands for *analysis of variance*. If you understand t-tests, we can give an important intuition for ANOVA as follows:

> ANOVA generalizes the t-test to any number of groups/levels.

To understand this, consider Fig. 2.1.

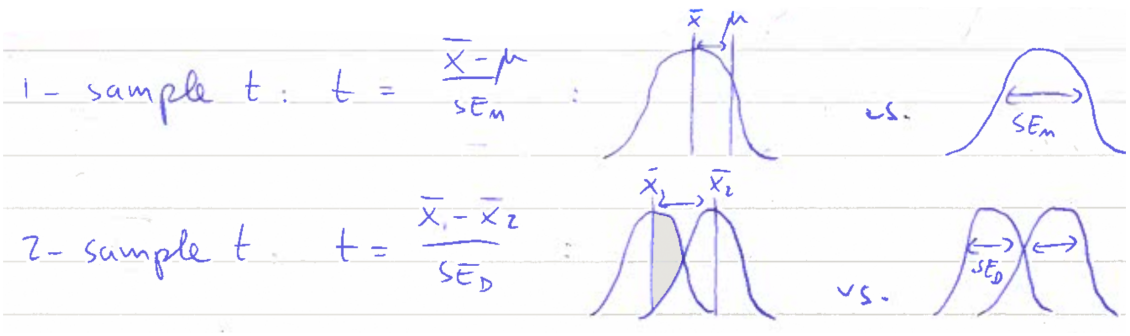



*Figure 2.1: Illustrated summary of the one-sample (top) and two-sample (bottom) t-tests. The t-test statistics are shown, along with an illustration of what the tests 'implicitly' compare: differences in means (i.e. variation* between groups*) and standard deviations of samples (i.e. variation* within groups*).*

For the one-sample t-test, one computes the difference between calculated mean and expected mean. This is compared to the standard error of the sample, i.e. the variation within the group. For the two-sample t-test one has the difference in the two calculated means (variation between groups), which is compared to the *combined* standard error of the two samples (variation within groups). Then, Fig. 2.2 shows how when $t$ is large, the two groups means should be 'far apart' compared to the width of the distribution, i.e. the variation within the group. This is when we expect the difference in means to be

*significant*. In contrast, when $t$ is small the difference between means is smaller than (or approximately of the same size) as the variation within the groups.

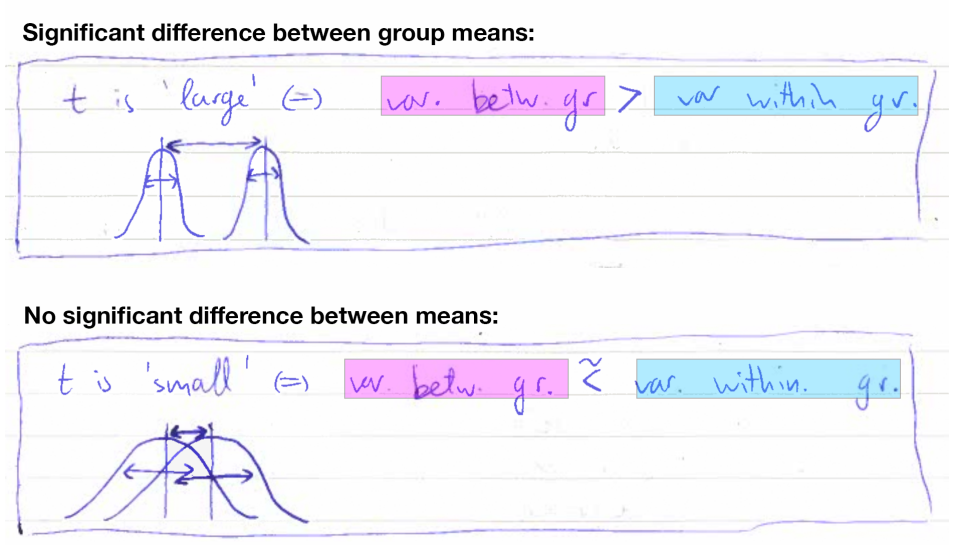



*Figure 2.2: A large t-test statistic means variation between groups is larger than variation within groups, and vice versa.*

In other words, we have made clear that:

T-tests calculate

$$\frac{\text{variation between groups}}{\text{variation within groups}} \tag{2.1}$$

This is also the essence of ANOVA!

Fig. 2.3 illustrates the main idea. Three groups are drawn, but any number is allowed.

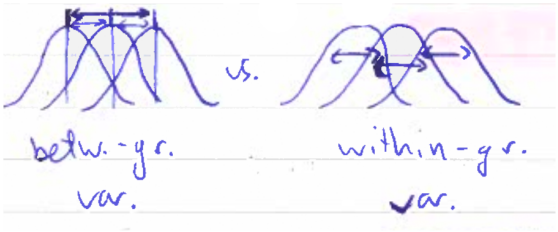



*Figure 2.3: Illustration capturing an important intuition behind ANOVA.*

## 2.1 One-way ANOVA

The simplest ANOVA is called *one-way ANOVA*. One-way here means we have one IV. If we have two IVs we get two-way ANOVA, which is a form of the general *factorial ANOVA*, which deals with any number of IVs. That is also called *N-way ANOVA*. These are discussed later.

### 2.1.1 Basics

The easiest way to understand ANOVA is using a concrete example. We borrow from [1]. Consider a medical trial where we compare the medications Joyzepam (J) vs. Anxifree (A) vs. a placebo (P). These are the groups/levels in our IV, medication. The variable we are interested in, the DV, is the 'mood gain', which we are presuming we have found using some appropriate measure. This gives a table of results, of which a couple rows are shown in Fig. 2.4.

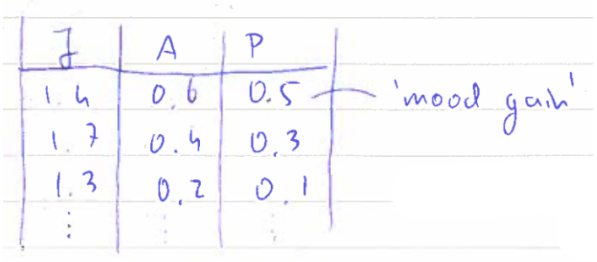



*Figure 2.4: First rows of a data table for the medical trial example.*

Our first research question might be: *is there a significant difference in group means between J, A and P?* Our hypotheses we can write as:

$$\begin{aligned} H_0: &\quad \mu_A = \mu_J = \mu_P \\ H_1: &\quad \text{not}(\mu_A = \mu_J = \mu_P) \end{aligned} \tag{2.2}$$

(note that we will mix our notation for the alternative hypothesis, $H_a$ and $H_1$). Now, given a concrete example, we can proceed with the general theory for one-way ANOVA. To do so, we need some notation.

- number of groups $N_G$ (3 in our example)
- number of subjects per group $N_k$ (assume 6 for our example)
- number of subjects $N$ (3*6=18 in our example)

Then, for the DV we use $Y$ in general, and $Y_{ik}$ is the value for subject $i$ in group $k$. We then get two types of means:

- $\bar{Y}_k$ is the mean in group $k$
- $\bar{Y}$ is the mean of all subjects, i.e. the *grand mean*

To understand ANOVA we will look at so-called *sums of squares*. To understand this, consider first the formula for *total variance*:

$$\text{var}(y) = \frac{1}{N-1} \sum_{k=1}^{N_G} \sum_{i=1}^{N_k} \left(Y_{ik} - \bar{Y}\right)^2$$

Notice that this is simply Eq. 1.2 written differently: since we have assigned our variable $Y$ to groups $k$, we obtain an additional index and sum compared to Eq. 1.2. Now, the *total sum of squares* is the same, but without the factor $1/(N-1)$:

$$\text{SS}_{\text{tot}} = \sum_k \sum_i \left(Y_{ik} - \bar{Y}\right)^2$$

where we simplify the summation notation for convenience. We now define the *within-group sum of squares* as

$$\mathrm{SS}_w = \sum_k \sum_i \left(Y_{ik} - \bar{Y}_k\right)^2 .$$

Notice in this equation that we subtract *values for individual subjects* from the *group means*: i.e. we look *within* the group. Then, the *between-group sum of squares* is:

$$\begin{aligned}\mathrm{SS}_b &= \sum_k \sum_i \left(\bar{Y}_k - \bar{Y}\right)^2 \\ &= \sum_k N_k \left(\bar{Y}_k - \bar{Y}\right)^2 .\end{aligned}$$

Here, observe that we subtract the *group means* from the *grand mean*: we look at differences between groups. The second equality in the equation holds because the sum over $i$ simply gives the number of subject in that cell. Notice how these quantify the sizes of the arrows in Fig. 2.3, i.e. the variation in mood gain. One can show that[1]:

$$\mathrm{SS}_w + \mathrm{SS}_b = \mathrm{SS}_{\mathrm{tot}}$$

which is sensible: the total variation in mood gain is a sum of the variation between groups and the variation within all the individual groups. Now, as mentioned, we would like to compare these using something like 2.1. Can we just divide $\mathrm{SS}_b$ by $\mathrm{SS}_w$, similar to the t-test? No. We need to take into account the *degrees of freedom* (DoF).

### 2.1.2 The F-test

We can understand the F-test statistic as follows:

> *ANOVA* generalizes the *t-test* (name of the technique).
>
> In the same way, the *F-test statistic* generalizes the *t-test statistic* (name of the test statistic).

Recall that for the two-sample t-test, the DoF were:

$$df = (n_1 - 1) + (n_2 - 1),$$

where $n_1, n_2$ are the two sample sizes. For the F-test, one has *three types* of DoF:

$$\begin{aligned}df_{\mathrm{tot}} &= N - 1 \\ df_b &= N_G - 1 \\ df_w &= N - N_G .\end{aligned}$$

Notice that $df_{\mathrm{tot}} = df_b + df_w$. From these and the sums of squares, one defines *mean squares*:

$$\begin{aligned}\mathrm{MS}_b &= \frac{\mathrm{SS}_b}{df_b} \\ \mathrm{MS}_w &= \frac{\mathrm{SS}_w}{df_w}\end{aligned}$$

Finally, the F-test statistic is calculated as:

$$\boxed{F = \frac{\mathrm{MS}_b}{\mathrm{MS}_w}}$$

I.e. we see that the F-test statistic indeed has the form of Eq. 2.1 but with DoF included. Using the F-test works the same way as with the t-test: one calculates it using equations above, and compares the result to the *F-distribution*. Values of the F-distribution can be found in critical value tables online,

[1] Try this!

which provide values of the F-distribution for different significance levels and DoF's. Note that an F-test table will look slightly different to t-test tables, since one has two DoF's instead of one DoF to keep track of. Sometimes one finds tables with $df_1$ and $df_2$ instead of $df_b$ and $df_w$. Usually, 1 refers to the numerator (i.e. $df_b$) and 2 to the denominator ($df_w$), but make sure to read the instructions of the table in order to use it in the right way. These tables are not symmetric across the diagonal!

The F-distribution is illustrated in Fig. 2.5. It is worth mentioning that in contrast to the t-distribution and the z-distribution (the normal distribution), one has that $F \geq 0$ (i.e. the distribution is only defined for positive values). This means that one is always doing a one-tailed test when using the F-distribution.

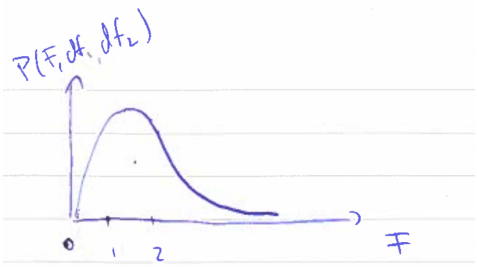



*Figure 2.5: Sketch of the F-distribution.*

Of course, for larger experiments one uses a computer to perform the test, which gives you a corresponding p-value directly. However, for pedagogical purposes it is useful to it by hand. Consider the continuation of the example of a medical trial, Fig. 2.6 shows an outline of the calculations necessary.

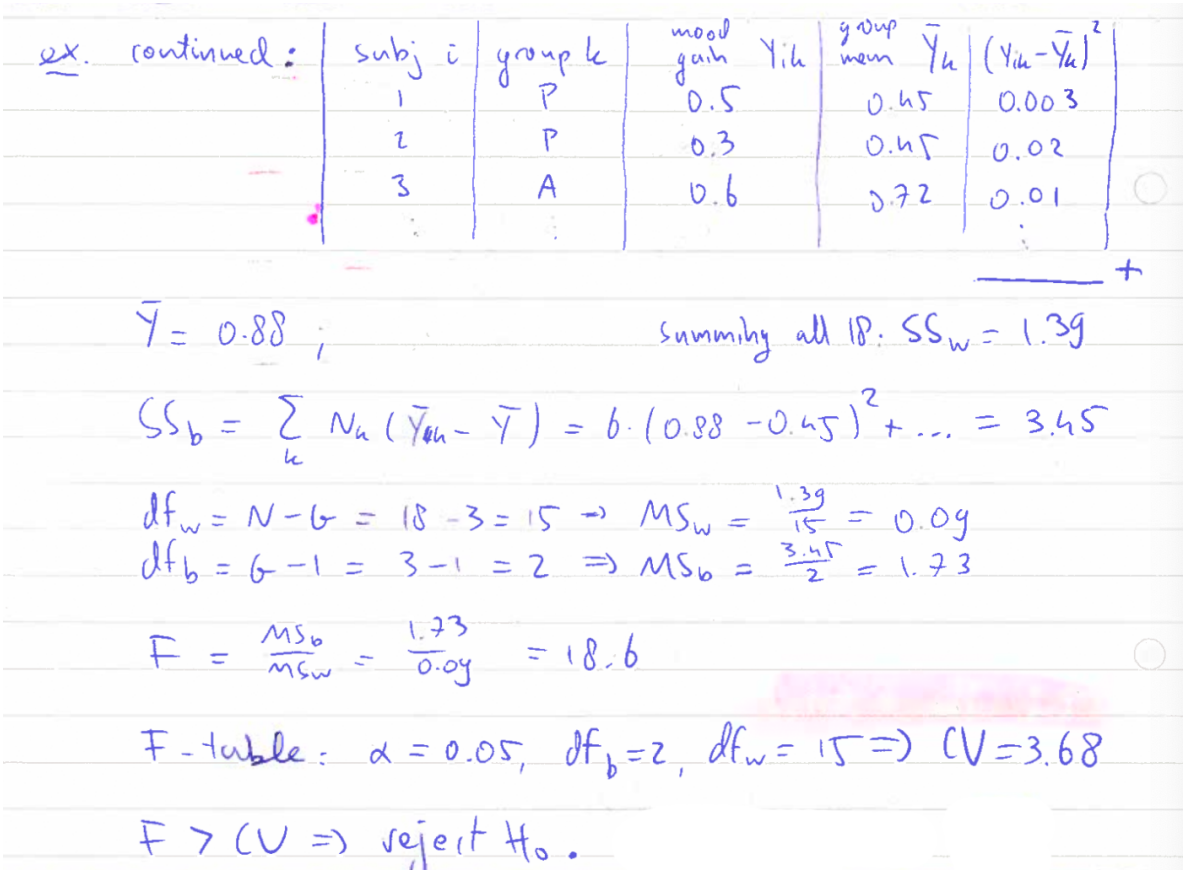



*Figure 2.6: Calculations for the medical trial example.*

From a table of results, one calculates different means, sums of squares and DoF's. Then, one calculates

mean squares and the F-statistic. Finally, one uses an F-table to find a critical value. From the last line, one has that $F$ is larger than the critical value, so we reject $H_0$. The results can be reported as follows:

> One-way ANOVA showed a significant effect of drug on mood gain ($F(2, 15) = 18.6$, $p < .001$).

Great! We know that drug has an effect. The logical next question is: is Joyzepam, Anxifree, or the placebo better? Note that the result to this question is not given by ANOVA! It only tells us there is *a* significant difference over all means. To find out which drug is better, we need to do a *post hoc analysis/multiple comparisons*. We discuss this after some important remarks on the F-test.

### 2.1.3 Miscellaneous important things

**Effect size**

Like with all other significant tests, F-tests have a corresponding effect size: *eta squared* ($\eta^2$). This is defined as:

$$\boxed{\eta^2 = \frac{\mathrm{SS}_b}{\mathrm{SS}_{\mathrm{tot}}}}$$

For our example, we have $3.45/4.84 = 0.71$. This means that *medication explains 71% of the variance in mood gain*!

**'The model'**

One sometimes finds the following description/definition of ANOVA:

$$\begin{aligned} H_0: &\quad Y_{ik} = \mu + \epsilon_{ik} \\ H_1: &\quad Y_{ik} = \mu_k + \epsilon_{ik} \end{aligned}$$

These equations say the following. The null hypothesis states that the values of the DV, $Y_{ik}$, are a single mean $\mu$ plus a random *error*, or *residual*, $\epsilon_{ik}$ – which does not depend on group. In contrast, $H_1$ is saying that the DV values depend on a *group-dependent* mean $\mu_k$, plus a random error. If true, this is equivalent to saying there is a (significant) difference in means. As such, it is a general way of writing down the one-way ANOVA hypotheses. Alternatively, one can say it *defines* the *one-way ANOVA 'model'*.[2] This can also be written as:

$$Y_{ik} = \mu + \alpha_k + \epsilon_{ik} \tag{2.3}$$

where $\alpha_k$ is the *effect* of group $k$ (the difference between the mean of group $k$ and the grand mean). This way of writing things becomes useful when we start considering more other more advanced models.

The word *model* should remind you that our analysis relies on assumptions which justify the use of the particular statistical test. These are discussed in detail later, but a first one to mention here is that the *errors $\epsilon_{ik}$ should be normally distributed*.

### 2.1.4 Post hoc analysis/multiple comparisons

Say we now want to know *which* medication actually works best. For this, we can just use *pairwise t-tests* between all the different groups: $A$ vs. $J$, $A$ vs. $P$, and $P$ vs. $J$. **HOWEVER:** there is a problem here! *If we do many of tests, a number of these tests will give significant results by chance only!* In other words, we get wat is known as an *inflated type I error rate*. For instance, say we have 10 groups, and do all 45 possible tests, we would get 2-3 that are significant by chance only. As such, *one cannot simply do all pairwise t-tests*, but one needs to *correct for multiple comparisons*.

The simplest way of doing this is called *Bonferroni corrections*. Say we do $m$ tests, and obtain a result $p_i$ for test $i = 1, 2, ..., m$. We define a total type I error as $\alpha$. Then, the *corrected/adjusted/honest* p-value is

$$p_i' = mp_i$$

[2] Here, the word model should remind you that the statistical analysis relies on assumptions, to be discussed in a bit.

and we reject $H_0$ if $p_i' < \alpha$. Another method is *Holm's method*, which is overall better but more complex. See [1] for more.

### 2.1.5 Assumptions

ANOVA has three key assumptions: *normality*, *homogeneity of variance*, and *independence*. These are summarized in Table 2.1. It is worth studying these in detail. In particular, it is valuable to try the Shapiro-Wilk, Kruskal-Wallis, Levene, and Brown-Forsythe test in SPSS/R/Python.

*Table 2.1: ANOVA assumptions. An example for A2 is e.g. $\sigma_A = \sigma_J = \sigma_P$.*

| | Assumption | Description | Checking | Removing |
|---|---|---|---|---|
| A1 | **Normality** | $\epsilon_{ik}$ normal | QQ plot/ Shapiro-Wilk test | Non-parametric test: 2 groups: Wilcoxon; $>2$ groups: Kruskal-Wallis test |
| A2 | **Homogeneity of variance** | $\sigma_1 = \sigma_2 = ... = \sigma$ | Levene/ Brown-Forsythe test | Welch one-way test |
| A3 | **Independence** | Observations unrelated | *Tricky...* | Probably: Repeated measures |

### 2.1.6 ANOVA table

Table 2.2 summarizes equations for one-way ANOVA. Also, Fig. 2.7 shows the *sum decomposition* of sums of squares and degrees of freedom, which we will extend in later sections.

*Table 2.2: ANOVA table for one-way design. Calculation of the mean squares is not included since this is always the SS divided by the corresponding df.*

| Source | Sum of Squares | DoF | $F$ |
|---|---|---|---|
| Between-group (effect) | $\mathrm{SS}_b = \sum_k N_k \left(\bar{Y}_k - \bar{Y}\right)^2$ | $N_G - 1$ | $\mathrm{MS}_b/\mathrm{MS}_w$ |
| Within-group (residual) | $\mathrm{SS}_w = \sum_k \sum_i \left(Y_{ik} - \bar{Y}_k\right)^2$ | $N - N_G$ | |
| Total | $\mathrm{SS}_{\mathrm{tot}} = \sum_k \sum_i \left(Y_{ik} - \bar{Y}\right)^2$ | $N - 1$ | |

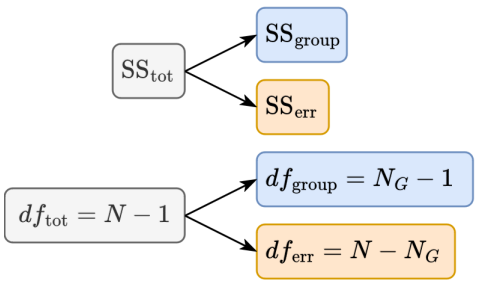



*Figure 2.7: Sum decompositions for one-way ANOVA. Sums of squares and degrees of freedom are shown.*

## 2.2 Factorial ANOVA

Consider again the example of the medical trial, where we were trying to find out whether medication had any effect on mood gain (ANOVA), as well as which of the medications had an effect (post hoc analysis). Now, imagine that our knowledge of psychology and neurobiology suggests that this medication should work even better when combined with therapy. To test this hypothesis, we will do the same experiment as before, but now two times with different subjects: one time with therapy, and one time without. As such, we add a *second IV*, meaning we enter the domain of *factorial ANOVA*.[3]

### 2.2.1 What changes?

What changes when we introduce another IV? We go through these in turn.

#### The 'design' and the 'design table'

Table 2.3 shows a '*design table*' for one-way ANOVA, which is a slightly more abstract version of Fig. 2.4. Then, Table 2.4 shows such a table for two-way factorial ANOVA. We see that these tables are a useful way of illustrating the *design* of our experiment: we see that for our two-way experiment, we get six different *cells* for which to find subjects. Drawing design tables is also a useful way of understanding differences between different methods, such as randomized blocks designs, repeated measures or mixed designs. Hence, we shall see many design tables when discussing different statistical methods.[4]

*Table 2.3: Design table for one-way ANOVA, with three levels for variable A.*

| $a_1$ | $a_2$ | $a_3$ |
|---|---|---|
| $s_1$<br>$s_2$ | $s_3$<br>$s_4$ | |

*Table 2.4: Design table for two-way ANOVA, with three levels for variable A and two levels for variable B.*

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $b_1$ | $s_1$<br>$s_2$ | $s_5$<br>$s_6$ | |
| $b_2$ | $s_3$<br>$s_4$ | $s_7$<br>$s_8$ | |

The tables shows numbered *subjects* $s_i$ with $i = 1, 2, ..., N$ in different cells (for simplicity, we shall sometimes leave out some subjects, if the pattern is clear from the given numbers). In Table 2.3, the columns are different levels of a single variable, here labeled by $a_j$ with $j = 1, 2, 3$, for some general variable $A$. Following our example, the table has three columns for the three levels (Joyzepam, Anxifree and placebo). Since we have only one variable, we also have three cells. Recall that the number of subjects per cell is determined or estimated based on the desired statistical power.

In Table 2.4, we have added rows for the two levels of our second IV: therapy/no therapy, labeled by $t_1$ and $t_2$. Since we have two IVs, this table corresponds to *two-way ANOVA*. Since we have 3 levels for IV1, and 2 levels for IV2, we see now that we get $3 \times 2 = 6$ cells for which to find subjects. As such, we could refer to the *design* of our experiment as a $(3 \times 2)$-*design*.

Now we can understand that with $N$ IVs, we get *N-way ANOVA*. If the IVs numbered by 1, 2,..., $N$, we get a:

$$(\text{levels}_{IV1}) \times (\text{levels}_{IV2}) \times ... \times (\text{levels}_{IVN})$$

-design. The number of cells in the design is simply equal to this multiplication. For instance, with 3 IVs with 2, 4, 3 levels respectively, one has 24 cells. Notice that when we have more than two IVs,

[3] What does the word factorial mean? It comes from the term *factor*, which in statistics is simply another word for a *discrete variable*. By requirement, IVs in (factorial) ANOVA are indeed discrete – hence this term. In these notes we shall usually simply to discrete variables.

[4] Note that, despite their usefulness, design table is not a standard term in the statistical literature.

one can no longer visualize the cells in the same nice way as in the figure, since one is restricted by the two dimensions of paper. One might imagine going to a three-dimensional cube with cells when adding the third variable.

### Sums of squares

We now return the the example with 2 IVs, i.e. two-way ANOVA. Importantly, another change from one-way ANOVA are the sums of squares. In particular, it is now ambiguous what should meant by *between-group* sum of squares: which groups? Do we mean the groups with/without therapy, or the groups with different medications?

Indeed, when we add IVs the term $\text{SS}_b = \text{SS}_{\text{between}}$ is *decomposed*. For two IVs $A$ and $B$, it decomposes into two terms for *main effects*: $\text{SS}_A$ and $\text{SS}_B$; and a term for the *interaction effect*, $\text{SS}_{AB}$. The sum of these, which we called $\text{SS}_{\text{between}}$ before, we now will write as $\text{SS}_{\text{group}}$. Furthermore, *within-group* is now also an ambiguous term. Hence, for factorial ANOVA and more advanced methods, we will instead refer to this term as the *residual* sum of squares, since it is the variation that is left over once we have subtracted all the *effects* (main and interaction) from the total variation. This is illustrated in the figure below.

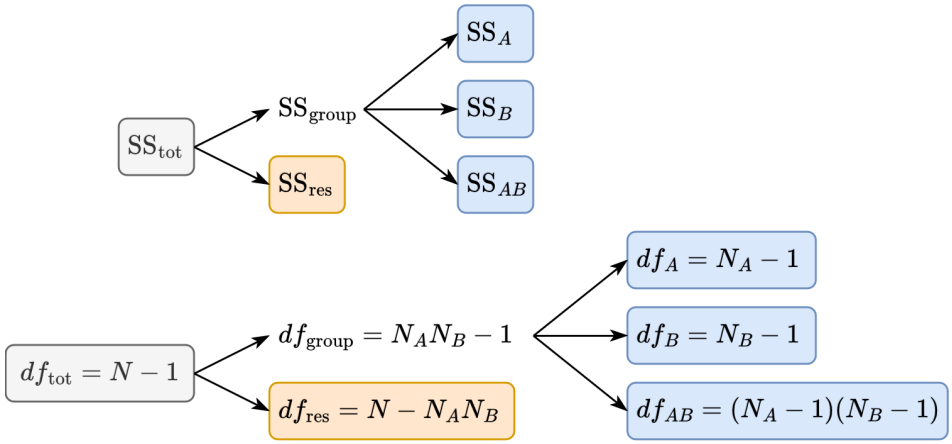



*Figure 2.8: Sum decomposition of sums of squares for two-way ANOVA.*

Since our design table becomes more complex, we also get additional means. All the means for two-way ANOVA are:

- $\bar{Y}$ is the grand mean, same as before.
- $\bar{Y}_{ab}$ is the mean in a single cell: level $a$ for variable $A$ and $b$ for variable $B$.
- $\bar{Y}_a$ is the mean for level $a$ in variable $A$. If one has levels of $A$ plotted as columns (as in Table 2.4), this means averaging over rows.
- $\bar{Y}_b$ is the mean for level $b$ in variable $B$. If one has levels of $B$ plotted as rows (as in Table 2.4), this means averaging over columns.

Using these we can calculate sums of squares using the equations below:

$$\text{SS}_A = N_k N_B \sum_a (\bar{Y} - \bar{Y}_a)^2$$

$$\text{SS}_B = N_k N_A \sum_b (\bar{Y} - \bar{Y}_b)^2$$

$$\text{SS}_{AB} = N_k \sum_a \sum_b (\bar{Y} + \bar{Y}_{ab} - \bar{Y}_a - \bar{Y}_b)^2$$

where $N_k$ is the *number of subjects per cell*, $N_A$ is the number of levels in variable $A$, and $N_B$ the number of levels in variable $B$. It is worth studying the similarities and differences with she expressions

encountered earlier. The expressions for the total variation is:

$$\mathrm{SS}_{\mathrm{tot}} = \sum_a \sum_b \sum_i (\bar{Y} - Y_{abi})^2,$$

where $Y_{abi}$ is the DV value for subject $i$ for level $a$ in IV $A$ and level $b$ in IV $B$. Finally, the residual sum of squares is:

$$\begin{aligned}\mathrm{SS}_{\mathrm{res}} &= \sum_a \sum_b \sum_i (\bar{Y}_{ab} - Y_{abi})^2 \\ &= \mathrm{SS}_{\mathrm{tot}} - \mathrm{SS}_A - \mathrm{SS}_B - \mathrm{SS}_{AB}.\end{aligned}$$

where the second equality is easier to use in practice.

### F-tests and degrees of freedom

One has mean squares, degrees of freedom, and F-tests for each effect. For the effect $A$:

$$df_A = N_A - 1, \quad \mathrm{MS}_A = \frac{\mathrm{SS}_A}{df_A}, \quad F_A = \frac{\mathrm{MS}_A}{\mathrm{MS}_{\mathrm{res}}}$$

where $\mathrm{MS}_{\mathrm{res}} = \mathrm{SS}_{\mathrm{res}}/df_{\mathrm{res}}$ and $df_{\mathrm{res}} = N - N_A N_B$. Similarly, for $B$:

$$df_B = N_B - 1, \quad \mathrm{MS}_B = \frac{\mathrm{SS}_B}{df_B}, \quad F_B = \frac{\mathrm{MS}_B}{\mathrm{MS}_{\mathrm{res}}},$$

and $AB$:

$$df_{AB} = (N_A - 1)(N_B - 1), \quad \mathrm{MS}_{AB} = \frac{\mathrm{SS}_{AB}}{df_{AB}}, \quad F_{AB} = \frac{\mathrm{MS}_{AB}}{\mathrm{MS}_{\mathrm{res}}}.$$

So, for a general effect we can write:

$$\boxed{F = \frac{\mathrm{MS}_{\mathrm{effect}}}{\mathrm{MS}_{\mathrm{res}}},}$$

where $\mathrm{MS}_{\mathrm{effect}} = \mathrm{SS}_{\mathrm{effect}}/df_{\mathrm{effect}}$. Here, *effect* is also often called *treatment* (typical in the medical literature), and *residual* is often called *error*. Note that when using an F-table, one should again make sure that one uses the right degrees of freedom for effects and residuals, respectively!

### Effect size

Intuitively, the effect size for $N$-way ANOVA is:

$$\boxed{\eta^2 = \frac{\mathrm{SS}_{\mathrm{effect}}}{\mathrm{SS}_{\mathrm{tot}}}} \tag{2.4}$$

However, there is a problem in this equation. $\mathrm{SS}_{\mathrm{tot}}$ includes other effects, next to the one we like to study, since $\mathrm{SS}_{\mathrm{tot}} = \mathrm{SS}_{\mathrm{res}} + \sum_{\mathrm{effects}} \mathrm{SS}_{\mathrm{effect}}$. Therefore, this variant of effect size is usually substituted by *partial eta squared* $\eta_p^2$:

$$\boxed{\eta_p^2 = \frac{\mathrm{SS}_{\mathrm{effect}}}{\mathrm{SS}_{\mathrm{effect}} + \mathrm{SS}_{\mathrm{res}}}.} \tag{2.5}$$

It is important to note however that for partial eta squared, the interpretation of explained variance no holder holds exactly, since the sum of effect sizes can be larger than one when using this formula.

### The model

'The model' of two-way ANOVA can be written as[5]:

$$Y_{abi} = \mu + \alpha_a + \beta_b + (\alpha\beta)_{ab} + \epsilon_{abi}, \tag{2.6}$$

where $\alpha_a$ is the effect due to $A$ in group $a$, $\beta_b$ the effect due to $B$ in group $b$, and $(\alpha\beta)_{ab}$ the interaction effect. $\epsilon_{abi}$ is an error like before, now depending on two indices for the cell and one for the subject. Compare this to Eq. 2.3. Notice how these equations give a very compact representation of the technique that is used.

[5]Note that the indices are arbitrary, chosen here to match the rest of our discussion. Another choice of notation that is often found is $Y_{jki} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{jki}$.

### 2.2.2 An example

Fig. 2.9 shows sketched calculations for the medical trial example, continued, now with a second IV. It is worth trying this by hand as well as with R/SPSS/Python.

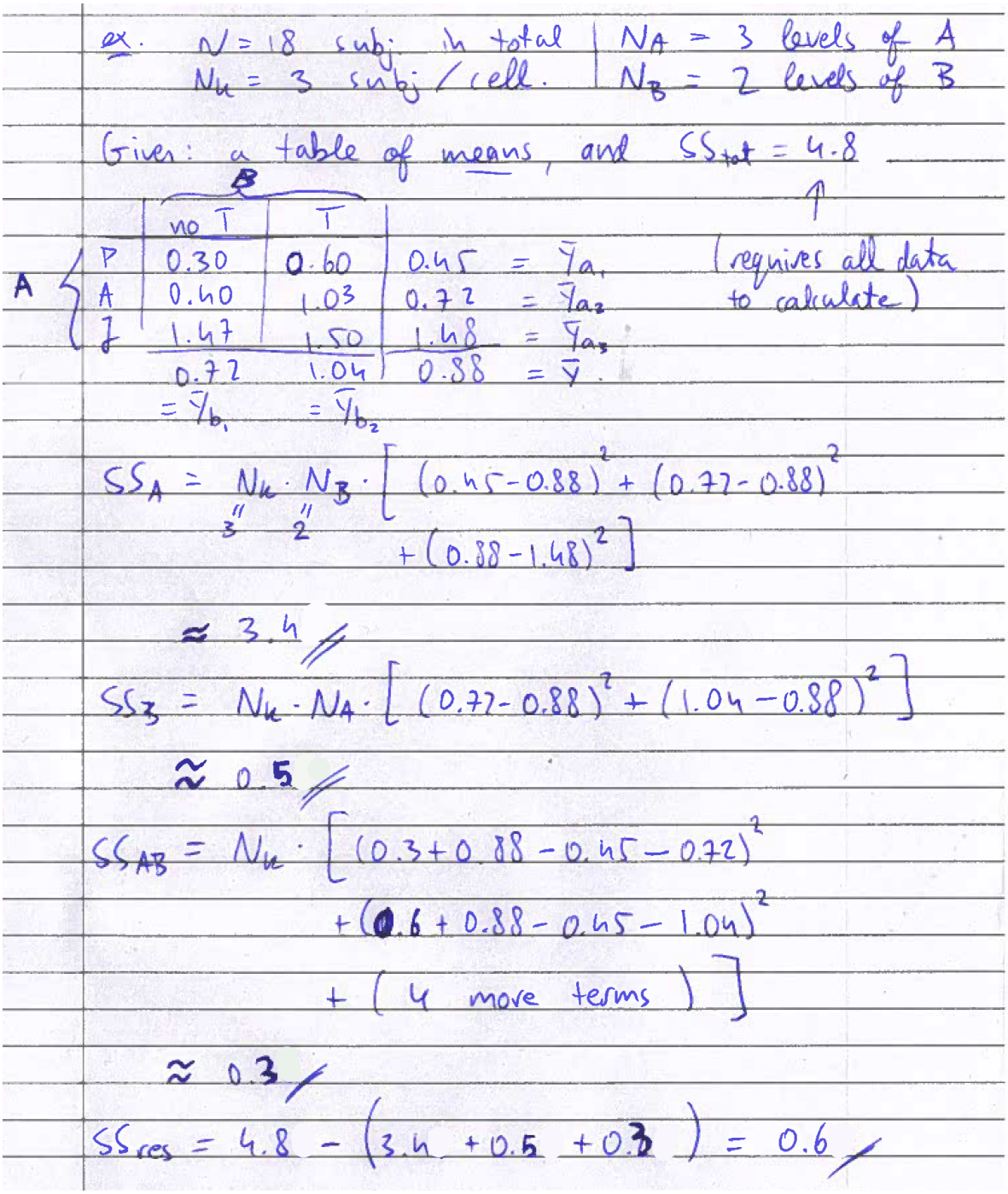



*Figure 2.9: Calculations for the medical trial example, with a second IV.*

### 2.2.3 Further notes

#### Interactions and line graphs

The interpretation of a one-way ANOVA is simple: one simply has a significant difference or not. Often, one can *guess* the outcome by looking at confidence intervals around the calculated means (if

means are included in other points' error bars it is to be expected that differences are not significant). For two-way ANOVA, it already becomes a bit more difficult to guess whether there might be an effect based on numbers alone. Hence, it is often useful to plot *line graphs*. These are very helpful for getting a sense of what effects might be significant or not with a single glance.

As an example, let us now consider a $(2 \times 2)$-design for the variables $A$ and $B$. In our line plot, we will plot the DV value on the vertical axis, the different levels of $A$ on the horizontal axis, and the different levels of $B$ with a different color. *Qualitatively*, there are 8 possible outcomes to this experiment. These are shown in the figures below: the first with interactions, and the second without interactions. It is worth studying these in detail to understand how they work. Notice that one can see at a single glance whether there might be an interaction effect or not, based on whether the lines are (approximately) parallel. It is also important to note that these graphs do not show error bars, which is bad practice.

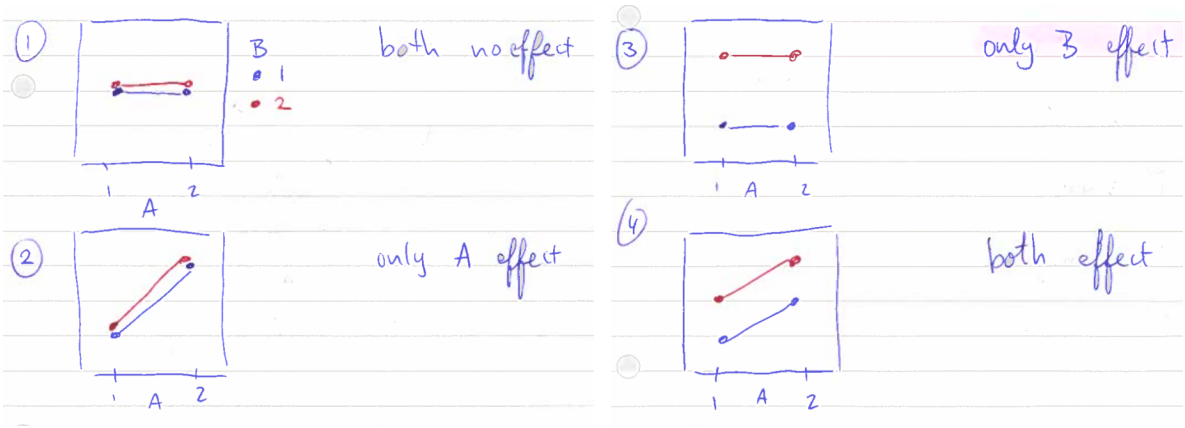



*Figure 2.11: Line graphs for a $(2 \times 2)$-design, without interactions.*

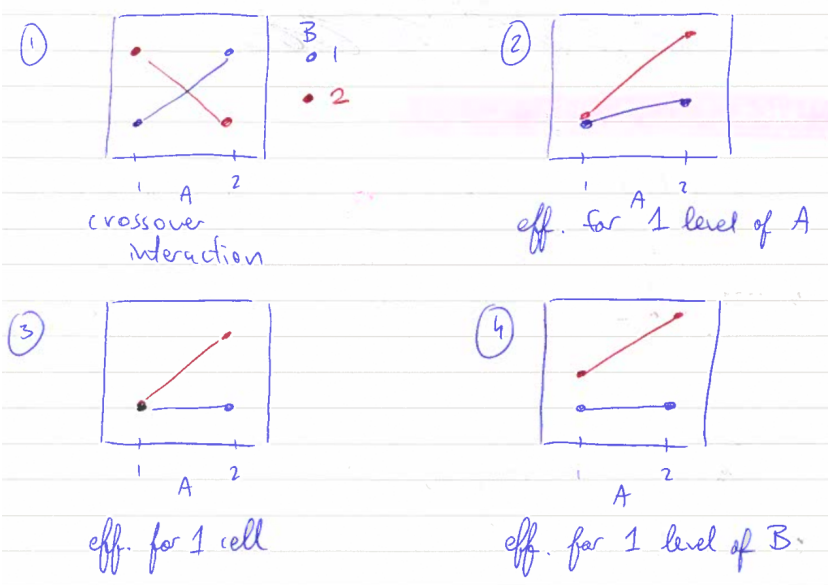



*Figure 2.12: Line graphs for a $(2 \times 2)$-design, with interactions.*

### What remains the same?

We have discussed what changes when going from one-way to two-way or $N$-way ANOVA. So what remains the same? Most importantly, the general 'philosophy' of comparing variation between groups, to variation within groups. Second, the assumptions remain unchanged: on still has to have normality, homogeneity of variance, and independence, as in Table 2.1.

### Post hoc analysis/multiple comparisons

When doing multiple comparisons with two-way or $N$-way ANOVA, the complicated thing is that one gets a very large number of pairwise comparisons to make. For instance, without interactions, one has $\mu_A$ vs. $\mu_P$, $\mu_A$ vs. $\mu_J$, $\mu_P$ vs. $\mu_J$, $\mu_T$ vs. $\mu_{noT}$. Including interactions, one has 19 comparisons! This is not a problem by itself, as one could use Bonferroni or Holm corrections as discussed earlier. Another useful test in this case, however, is Tukey's *honest significant difference test* (HSD). This allows testing *all* pairwise correlations. It provides a *simultaneous confidence interval*, which gives limits such that one is 95% confident that *all* means fall within the interval defined by the limits.

### ANOVA table

The equations for two-way factorial ANOVA are summarized in Table 2.5.

*Table 2.5: ANOVA table for two-way design.*

| Source | Sum of Squares | DoF | $F$ |
|---|---|---|---|
| $A$ | $\mathrm{SS}_A = N_k N_B \sum_a (\bar{Y} - \bar{Y}_a)^2$ | $N_A - 1$ | $\mathrm{MS}_A/\mathrm{MS}_{\mathrm{err}}$ |
| $B$ | $\mathrm{SS}_B = N_k N_A \sum_b (\bar{Y} - \bar{Y}_b)^2$ | $N_B - 1$ | $\mathrm{MS}_A/\mathrm{MS}_{\mathrm{err}}$ |
| $AB$ | $\mathrm{SS}_{AB} = N_k \sum_a \sum_b (\bar{Y} + \bar{Y}_{ab} - \bar{Y}_a - \bar{Y}_b)^2$ | $(N_A - 1)(N_B - 1)$ | $\mathrm{MS}_{AB}/\mathrm{MS}_{\mathrm{err}}$ |
| Residuals | $\mathrm{SS}_{\mathrm{err}} = \sum_a \sum_b \sum_i (\bar{Y}_{ab} - Y_{abi})^2$ | $N - N_A N_B$ | |
| Total | $\mathrm{SS}_{\mathrm{tot}} = \sum_a \sum_b \sum_i (\bar{Y} - Y_{abi})^2$ | $N - 1$ | |

## 2.3 RBDs and repeated measures

This section reviews randomized blocks designs (RBDs) and repeated measures ANOVA. It should be noted that although these techniques can seem conceptually somewhat different, on a technical level they are identical Although RM is perhaps used more often, it becomes easier to understand if one deals with RBDs first. Hence, we follow this order here.

### 2.3.1 Randomized blocks designs (RBDs)

Let us return to our running example of a medical trial, going back to only one IV: medication. Now imagine that we have received some information from other researchers suggesting that Joyzepam works much better for *young adults* – something we did not take into account in our earlier analysis. How do we deal with this?

#### Nuisance variables

A first idea might be to simply add *age* as our second IV in a two-way factorial ANOVA design. This would allow us to study the effect of age its interaction with treatment, thus isolating the effect of medication as a separate effect. However... clearly we would like Joyzepam to work for *all* ages. We could say that in such a case we are *not interested* in these effects.

Hence, we could instead label age as a so-called*nuisance variable* (or *nuisance factor* if we discretize it): a *covariate* that we are not interested in. (Recall that a covariate is an IV that we want to control for.) One way of doing this is using an RBD, also called *blocking*. We can summarize the main idea of RBD as follows:

> In RBD we do a new study where we put *'similar subjects'* in a *block*. This reduces the error term by subtracting variance of blocks due to the nuisance variable. We additionally *randomize* assignment of subjects in levels of the other variable(s).

Let us dissect the meaning of this statement step by step. The first sentence refers to '*similar subjects*': this means similar in some IV/CV that we know relates to the DV. This is the nuisance variable (or blocking variable): age in our case. Then, putting these in *blocks* means assigning levels to this nuisance variables (these are the blocks), and putting subjects in them appropriately. In our example, one could define different age groups as blocks. Finally and importantly, we should assign subjects *randomly* to the blocks – which explains the final part of the term randomized blocks design.[6] The design table in Fig. 2.6 illustrates how RBD has one subject in each cell.

*Table 2.6: Design table for a one-way RBD. Note that the letter b here is used for blocks, not for a standard variable like in two-way ANOVA. Also note that an RBD always has one subject per block per group (i.e. one per cell), and that subjects in each block should be similar w.r.t. the nuisance/blocking variable.*

| | $g_1$ | $g_2$ | $g_3$ |
|---|---|---|---|
| $b_1$ | $s_1$ | $s_2$ | $s_3$ |
| $b_2$ | $s_4$ | $s_5$ | $s_6$ |
| $b_3$ | | | |

#### Reducing the error term

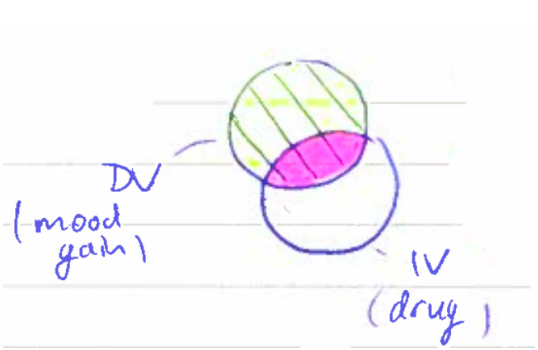
Then the second part of the sentence. How does this 'reduce the error term' and why is this desirable? We can gain an intuition by looking at Fig. 2.18 (note that this holds not jsut for RBDs, but for addition of CVs in general – thus also for ANCOVA, discussed later on). On the left we have an IV that explains some of the variance in the DV, with the yellow area representing $SS_{err}$, i.e. the error term. Now, when adding a CV (right), this will explain some *additional* variance in the DV – hence reducing the yellow area, i.e. the error term.

[6] Note that we here distinguish clearly between the terms *group* and *block*.

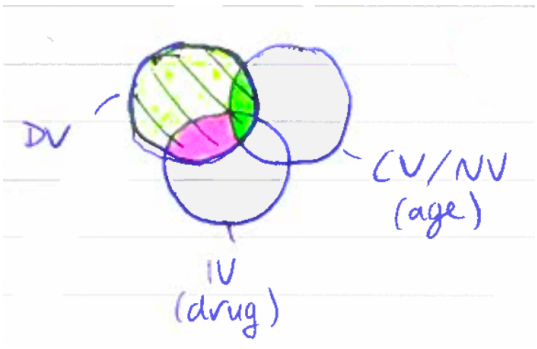
This is useful for two reasons. First, one isolates the effect one is interested in: the pink area on the right. Second, doing so effectively increases the power (or: sensitivity) of the analysis. To understand this, consider the F-test formula:

$$F = \frac{\text{MS}_{\text{effect}}}{\text{MS}_{\text{err}}} \propto \frac{\text{SS}_{\text{effect}}}{\text{SS}_{\text{err}}}$$

where the $\propto$ symbol means 'is proportional to' (i.e. is equal up to a multiplicative constant). We can observe that since $\text{SS}_{\text{err}}$ has *decreased*, $\text{MS}_{\text{err}}$ has also decreased, which means that the F-test statistic *increases* – thus increasing power.[7] The box below briefly discussed error term reduction in factorial ANOVA with the same intuition.




*(a) Without CVs (standard ANOVA).*

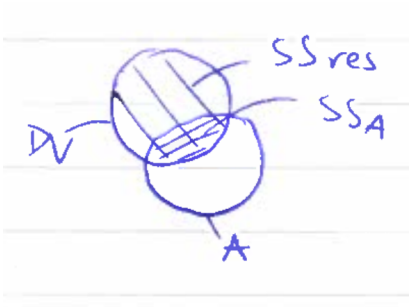



*(b) With CVs (RBD/ANCOVA).*

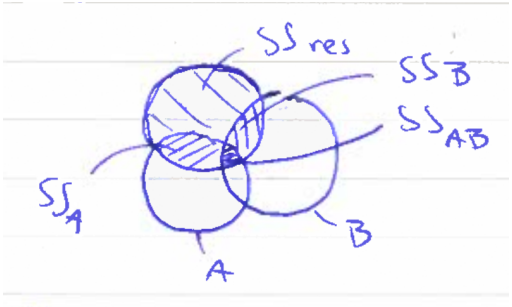
*Figure 2.13: Illustration of decreased error term through addition of a CV. The striped area represents* $\text{SS}_{\text{tot}}$*, the pink area is* $\text{SS}_{\text{effect}}$*, the yellow area is* $\text{SS}_{\text{err}}$*, and the green area is* $\text{SS}_{\text{CV}}$*, also called* $\text{SS}_{\text{block}}$ *when using RBD. By adding a CV, the yellow area decreases. Indeed, the pink area decreases too, but less than the yellow area if the CV is well-chosen. Note that on the right, the yellow and pink area combined represents the residual sum of squares after doing the regression of ANCOVA (Section 3.3).*

**ASIDE: Reducing the error term in factorial ANOVA**

You might be wondering: what happened to the error term when we went from one-way to two-way ANOVA? Indeed, as shown in Fig. 2.14: the area of $\text{SS}_{res}$ decreases.




*(a) One-way ANOVA.*




*(b) Two-way ANOVA.*

*Figure 2.14: Illustration of decreased error term through addition of an IV.*

We now consider in more detail what happens to the sums of squares and degrees of freedom. Consider the decompositions that we had in one-way ANOVA, shown in Fig. 2.15. Notice that we have relabeled $\text{SS}_{\text{within}}$ to $\text{SS}_{\text{err}}$. Now we introduce $N_B$ blocks and $N_G$ groups. We then decompose the $\text{SS}_{\text{within}}$ term into variation due to the blocks, and the interaction between block and group, shown in Fig. 2.16.

[7] Indeed, we are forgetting the degrees of freedom here, which also decrease, compensating for the decrease in $SS$. If one has chosen a good blocking variable, the decrease in $SS$ should be larger than the corresponding decrease in DoFs.

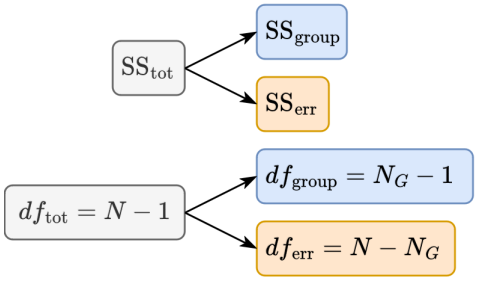



*Figure 2.15: Sum decompositions for one-way ANOVA with renamed terms. Sums of squares and degrees of freedom are shown.*

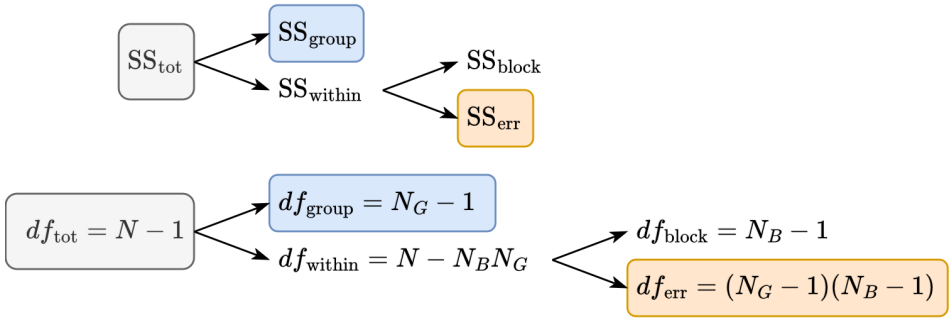



*Figure 2.16: Sum decompositions for one-way RBD ANOVA. Sums of squares and degrees of freedom are shown.*

The F-test does not change from before:

$$F = \frac{\text{MS}_{\text{group}}}{\text{MS}_{\text{err}}},$$

**BUT**, importantly, we *use the block-group interaction as the error term*, i.e. we set

$$\boxed{\text{MS}_{\text{err}} = \text{MS}_{\text{block*group}}.}$$

In RBD, we do not use the original $\text{SS}_{\text{within}}$ term! This is shown in Figs. 2.15 and 2.16 using orange color to denote the term (SS) and corresponding DoF that we use as error term. Notice that this means that we have effectively decreased/*reduced* the error term.[8]

### ANOVA table, effect size

Full expressions for sums of squares and degrees of freedom are shown in Table 2.7. And what about effect sizes with RBD? This can be done using Eqs. 2.4 and 2.5, simply filling in the right term for $\text{SS}_{\text{effect}}$ and $\text{SS}_{\text{err}}$!

### Sphericity

The assumptions for RBD are the same as for one-way ANOVA, with one important change. The assumption of homogeneity of variance now changes to a stronger assumption, called *sphericity*, which says that one should have *equal variances of differences between all combinations of groups* (Table 2.8). This is a mouthful, but can be illustrated as follows. Testing sphericity can be done with *Mauchly's test*.

[8] As a complementary perspective, recall that in factorial ANOVA we decomposed the $\text{SS}_{\text{between}}$ term into three different *effects* that *we are interested in*. In contrast, we now say we are *not* interested in our blocking variable, hence we decompose instead $\text{SS}_{\text{within}}$.

*Table 2.7: ANOVA table for RBDs and repeated measures ANOVA. For RM, 'block' should be replaced by 'subject', $N_B$ by $N_S$, and $b$ by $s$*

| Source. | Sum of Squares | DoF | $F$ |
|---|---|---|---|
| Group | $\mathrm{SS}_{\mathrm{group}} = N_B \sum_g \left(\bar{Y}_g - \bar{Y}\right)^2$ | $N_G - 1$ | $\mathrm{MS}_{\mathrm{group}}/\mathrm{MS}_{\mathrm{err}}$ |
| Block | $\mathrm{SS}_{\mathrm{block}} = N_G \sum_b \left(\bar{Y}_b - \bar{Y}\right)^2$ | $N_B(N_G - 1)$ | |
| Within-group | $\mathrm{SS}_{\mathrm{within}} = \sum_g \sum_b \left(Y_{gb} - \bar{Y}_g\right)^2$ | $N - N_B N_G$ | |
| Residual | $\mathrm{SS}_{\mathrm{err}} = \mathrm{SS}_{\mathrm{within}} - \mathrm{SS}_{\mathrm{block}}$ | $(N_G - 1)(N_B - 1)$ | |
| Total | $\mathrm{SS}_{\mathrm{tot}} = \sum_g \sum_b \left(Y_{gb} - \bar{Y}\right)^2$ | $N - 1$ | |

*Table 2.8: Table illustrating sphericity. On the left a data table is shown for three groups and four blocks. On the right, group differences are calculated, and variances for these are shown at the bottom. Sphericity requires that these variances are of comparable size – which appears not to be the case here, but one should do Mauchly's test for a proper check.*

| **Data table** | | | | **Group differences** | | |
|---|---|---|---|---|---|---|
| | $G_1$ | $G_2$ | $G_3$ | $G_1 - G_2$ | $G_1 - G_3$ | $G_2 - G_3$ |
| $B_1$ | 45 | 50 | 55 | -5 | -10 | -5 |
| $B_2$ | 42 | 42 | 45 | 0 | -3 | -3 |
| $B_3$ | 36 | 41 | 43 | -5 | -7 | -2 |
| $B_4$ | 39 | 35 | 40 | 4 | -1 | -5 |
| | | | **Variance:** | 19 | 16.25 | 2.25 |

### 2.3.2 Repeated measures (RM)

We are now ready to understand repeated measures (RM) ANOVA. RM can be understood in one sentence as follows:

> RM takes RBD to the extreme by taking 'perfect similarity' in blocks: the same subject measured several times.

This explains how another name for repeated measures is *within-subject*: we have the same subject in several cells. This term can be contrasted with *between-subject*, where different cells in a design contain different subjects. These can also be combined, which is discussed in the next section.

#### What changes?

*Table 2.9: Design table for a one-way repeated measures design. Note that this design table is the same as Table 2.6, but with blocks now containing only one subject, measured several times on variable $G$.*

| | $g_1$ | $g_2$ | $g_3$ |
|---|---|---|---|
| $s_1$ | $s_1$ | $s_1$ | |
| $s_2$ | $s_2$ | $s_2$ | |
| $s_3$ | $s_3$ | $s_3$ | |

What does that mean in practice? Everything described in the previous paragraph still holds: for our calculations we need only make minor changes in our terminology and notation. The design table in Fig. 2.6 visualizes how $b$ simply becomes $s$. Sums of squares stay the same, replacing *block* by *subject*.

In the equations in Table 2.7 the following changes are made:

$$\begin{aligned} \mathrm{SS}_{\mathrm{block}} &\rightarrow \mathrm{SS}_{\mathrm{subj}} \\ \mathrm{SS}_{\mathrm{block*group}} &\rightarrow \mathrm{SS}_{\mathrm{subj*group}} = \mathrm{SS}_{\mathrm{err}} \\ N_B \ (\#\ \text{blocks}) &\rightarrow N_S \ (\#\ \text{groups}). \end{aligned}$$

Finally, the F-test, then, is still simply

$$F = \frac{\mathrm{MS}_{\mathrm{group}}}{\mathrm{MS}_{\mathrm{err}}},$$

where we remember that we need to *choose the subject-group interaction as the error term*:

$$\mathrm{SS}_{\mathrm{err}} = \mathrm{SS}_{\mathrm{subject*group}}.$$

The ANOVA table 2.7 and formula for effect size of RBD also hold for RM.

### An example

A typical example use case for repeated measures is studying the effect of time. For instance, in our medical trial, imagine we are interested in the effects of Joyzepam just after treatment, and three months later. As such, we could measure mood at three moments in time: just before, just after, and after three months. The figure below shows sketched example calculations.

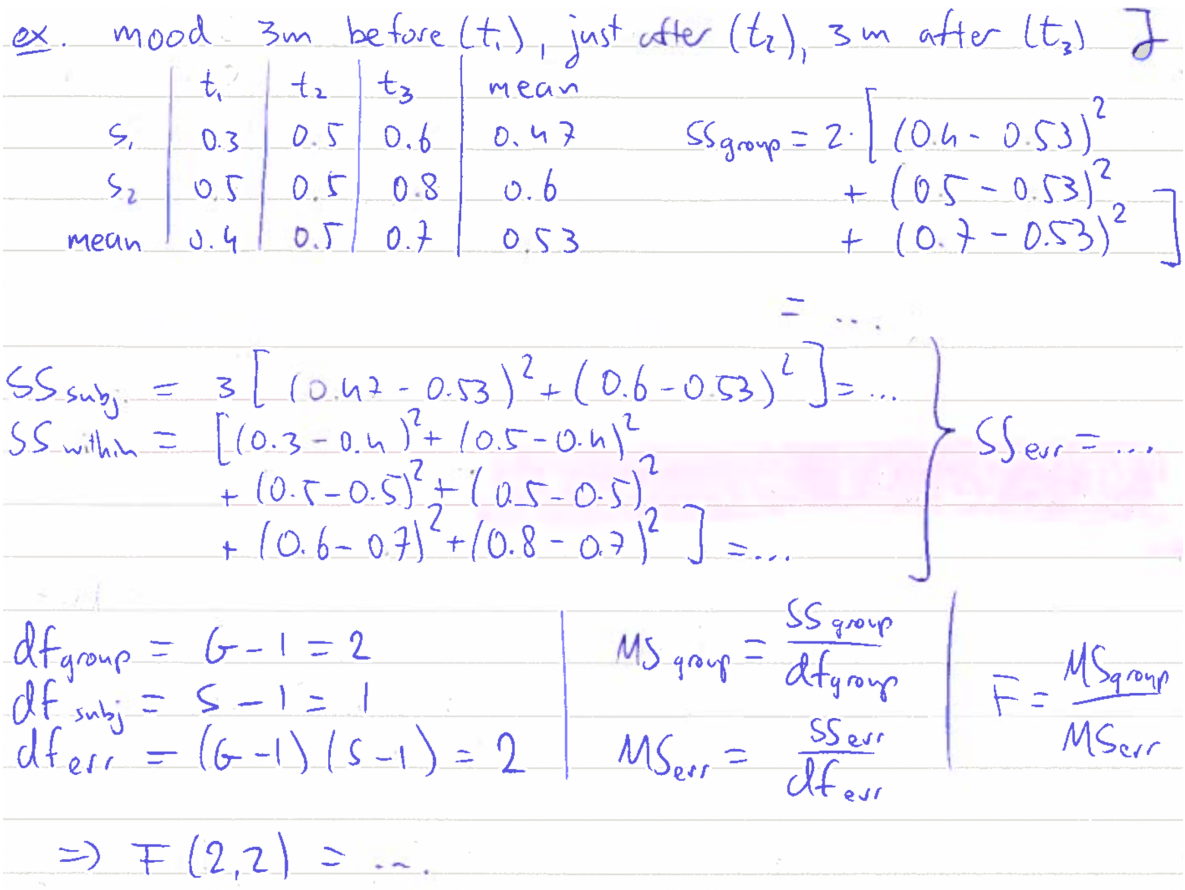



*Figure 2.17: Calculations for a simple example of repeated measures ANOVA.*

## 2.4 More complex designs

The basics for ANOVA have now been covered. We are now ready to increase complexity! The first way we do this is by considering the combination of repeated measures with factorial ANOVA.

### 2.4.1 Combining factorial with RM

How do we do that?

There are two ways of combining RM with factorial ANOVA:

1. **Factorial-within subject design**: each subject in each cell.
2. **Mixed design**: not all subjects in each cell.

Note that we use the word 'cell' with the precise meaning discussed before. We review these two techniques below.[9]

#### Factorial within-subject designs

For this case, one has several IVs that are all within-subject. Table 2.10 shows the design table. Two IVs ($A$ and $B$) are shown, with 2 and 3 levels respectively. Note how each cell includes all subjects.

*Table 2.10: Design table for a factorial within-subject design. Notice the similarity with Table 2.4*

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $b_1$ | $s_1$<br>$s_2$ | $s_1$<br>$s_2$ | |
| $b_2$ | $s_1$<br>$s_2$ | $s_1$<br>$s_2$ | |

What could be a concrete example of such a design? This is the first time where our running example of the medical trial is inappropriate, since the complicated nature of medication and the duration of experiments makes it unlikely that applying all 6 combinations of therapy and medication to the same patients is desirable.

Hence, a different example is appropriate. In a HCI context, one could imagine an example of testing the user-friendliness of some website, where the website is changed on two variables: e.g. three different fonts (variable $A$) and two color schemes (variable $B$). Since testing a website can be done quickly, this is somewhat more realistic. Of course, one should make sure to consider learning effects; if one uses the same subjects for the same task, it is conceivable that subjects perform 'better' on the task with time as they get more practice. This relates to Latin-Square designs (Section 2.4.4).

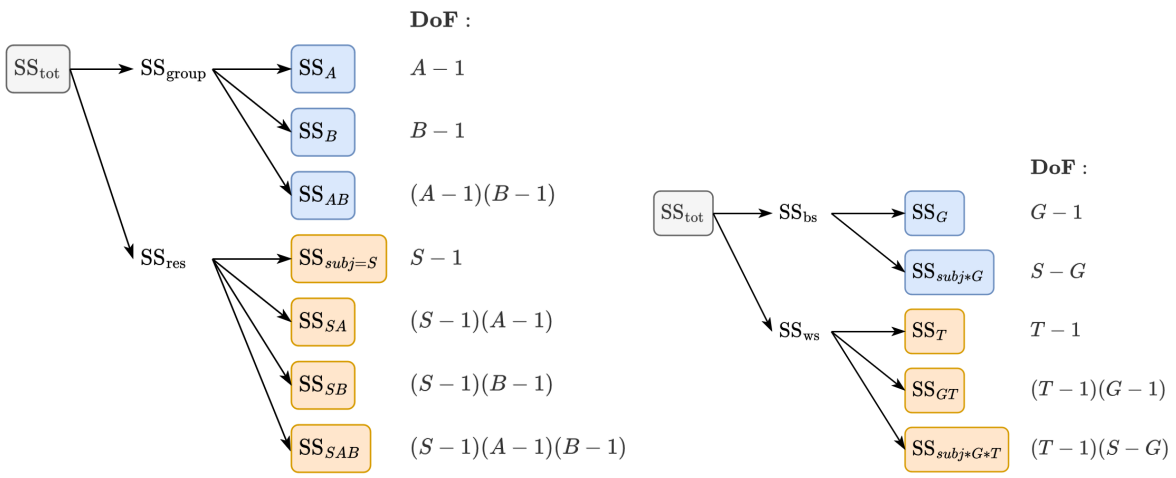
The sum decomposition with corresponding degrees of freedom for this case is shown in Fig. 2.18a. The appropriate equation for the F-test now involves a separate error term for each effect – which we shall ignore here to keep the discussion brief[10]. For effect size one uses Eqs. 2.4 and 2.5 as before.

#### Mixed designs

These sometimes also called 'split-plot' ANOVA. For this case, one has some IVs that are between-subject, and some IVs that are within-subject. A design table is shown in Table 2.11. Notice how the cells for the within-subject variable contain the same subjects, while the between-subject variable corresponds to two different groups. It shows a typical example where the within-subject IV is time (note that this doesn't need to be the case). Here, our running example is appropriate again: we might want to study the effect of medication (the groups, J/A/P) at three different moments in time (as in Section 2.3.2).

[9] Note that due to their complexity, we will not ask you to calculate sums of squares by hand for these designs. We provide the sum decompositions and degrees of freedom for understanding their formal structure.

[10] See e.g. 3.2.3 in [2].




*(a) Factorial within-subject ANOVA.* *(b) Mixed design.*

*Figure 2.18: Sum of squares decompositions with corresponding degrees of freedom for the two ways of combining factorial ANOVA with repeated measures.*

*Table 2.11: Design table for a mixed design.*

| | $t_1$ | $t_2$ | $t_3$ |
|---|---|---|---|
| $g_1$ | $s_1$<br>$s_2$ | $s_1$<br>$s_2$ | |
| $g_2$ | $s_3$<br>$s_4$ | $s_3$<br>$s_4$ | |

The sum of square decomposition, along with corresponding degrees of freedom, is shown again in Fig. 2.18b. Similar to factorial within-subject designs, choosing the right error term for the F-test requires some bookkeeping – we shall spare you the details here. Effect sizes are calculated as before.

### 2.4.2 Nesting

Consider a completely different example. Imagine that we are trying to study the performance of students in four different classrooms $c_1, c_2, c_3, c_4$, based on two different teaching methods, $t_1, t_2$. Your first thought might be: "two IVs, so let's do two-way ANOVA". This would give the design table in Table 2.12. However, at second thought... how can we teach students in the same classroom with different methods? Clearly, we have to assign a teaching method to an *entire* classroom! Thus, we should instead have the design table in Fig. 2.13. This is an example of a *nested* design: levels of $T$ occur *within* levels of $C$. The previous examples of factorial ANOVA that we have seen are called *crossed* designs. Here is the general definition.

*Table 2.12: Erroneous(!) design table for the classroom-teaching method example. A classroom can only have one teaching method.*

| | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
|---|---|---|---|---|
| $t_1$ | $s_1$<br>$s_2$ | $s_5$<br>$s_6$ | | |
| $t_2$ | $s_3$<br>$s_4$ | $s_7$<br>$s_8$ | | |

*Table 2.13: Correct design table for the classroom-teaching method example: a nested design where each classroom has one teaching method. Grey cells are not used.*

| | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
|---|---|---|---|---|
| $t_1$ | $s_1$<br>$s_2$ | $s_3$<br>$s_4$ | | |
| $t_2$ | | | $s_5$<br>$s_6$ | $s_7$<br>$s_8$ |

*Table 2.14: Design table for a nested design, showing means instead of individual subjects. Note that since subjects can only be in one classroom, we cannot visualize the classrooms as rows in this table – and only show the means. Note also that classrooms $c_i$ with $i = 1, 2, ...$ are assigned at random to different teaching methods.*

| $t_1$ | $t_2$ |
|---|---|
| $c_1$ mean<br>$c_2$ mean | $c_3$ mean<br>$c_4$ mean |

We say that the two-way layout is *crossed* when every level of Factor A occurs with every level of Factor B. We say we have a *nested* layout when fewer than all levels of one factor occur within each level of the other factor.

In such a design, what effects can we consider? Most simply, if we want to know the effect of teaching method, the simplest solution is to use the classroom *mean* as the DV, in an otherwise standard one-way between-subject design. This would give the design table in Fig. 2.14 (for the two-level case here doing a t-test would of course be simpler).

However, what if we also want to test the effect of the classroom? For this, consider the sum of squares decomposition in Fig. 2.19. Here, $A$ is the main variable (teaching method), and $B$ is the nested variable (classroom), writing $B(A)$ to emphasize that it occurs within $A$. We can do two tests, which use *different error terms*! First for the main effect:

$$F_A = \frac{\text{MS}_A}{\text{MS}_{B(A)}}.$$

This corresponds to the one-way design using the means of Table 2.14. Then, for the nested variable:

$$F_{B(A)} = \frac{\text{MS}_B(A)}{\text{MS}_{res}}.$$

where $MS_{res}$ is the orange box in Fig. 2.19, equal to $\text{SS}_{\text{res}} = \text{SS}_{\text{tot}} - \text{SS}_A - \text{SS}_{B(A)}$. Note that this is the same error term as in the crossed two-way ANOVA. Equations for two-way nested ANOVA are

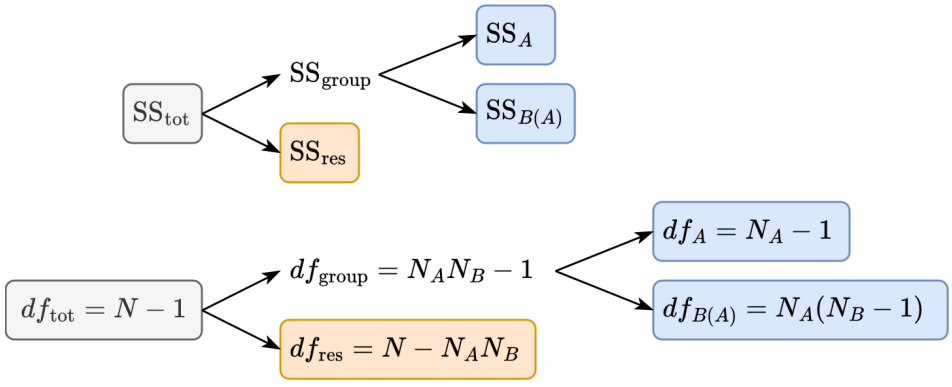



*Figure 2.19: Sum decompositions for nested two-way ANOVA. Compare this to Fig. 2.8.*

*Table 2.15: ANOVA table for a two-way nested design.*

| Source | Sum of Squares | DoF | $F$ |
|---|---|---|---|
| A | $\mathrm{SS}_A = N_k N_B \sum_a (\bar{Y} - \bar{Y}_a)^2$ | $N_A - 1$ | $\mathrm{MS}_A/\mathrm{MS}_{\mathrm{B(A)}}$ |
| B within A | $\mathrm{SS}_{B(A)} = N_k \sum_a \sum_b (\bar{Y}_{ab} - \bar{Y}_a)^2$ | $N_A(N_B - 1)$ | $\mathrm{MS}_A/\mathrm{MS}_{\mathrm{err}}$ |
| Residuals | $\mathrm{SS}_{\mathrm{err}} = \sum_a \sum_b \sum_i (\bar{Y}_{ab} - Y_{abi})^2$ | $N - N_A N_B$ | |
| Total | $\mathrm{SS}_{\mathrm{tot}} = \sum_a \sum_b \sum_i (\bar{Y} - Y_{abi})^2$ | $N - 1$ | |

summarized in Table. 2.15. The model for a nested design is:

$$Y_{abi} = \mu + \alpha_a + \beta_{b(a)} + \epsilon_{abi},$$

where $b(a)$ indicates that levels of $B$ occur within levels of $A$. Compare this to Eq. 2.6. Of course, the concept of nesting can be extended in various ways to much more complex situations. Such designs are best modeled using *multilevel modeling*, which is outside the scope of this course.

### 2.4.3 Random and fixed effects

In ANOVA designs discussed so far, levels of IVs were chosen to test significant differences in that IV. Indeed, this is how the very idea of ANOVA was presented. In reality, this is just *part* of ANOVA: the *fixed effect model* (with the IV modelled as a *fixed effect*). Sometimes, however, one would like to *generalize* to a *population of levels* of an IV, in which case one has a *random effects model*.

Let's take an example: consider (again) our medical trial with a treatment and placebo group. So far, we discussed two ways of dealing with age: if we are *not* interested in age we add it as a nuisance variable in an RBD, and if we *are* interested, we include it as a second IV in a two-way factorial ANOVA.

In this latter case, in the default approach it is *differences between age groups* that are of primary interest. There is, however, another way of 'being interested in age': we could be interested only in its *general* effect, *without* being interested in the *specific age groups* (and their differences). In this case, we want to model the effect of age, but *generalized to the full population* of age groups.

> Fixed effects are used when specific levels of an IV are of primary interest. Random effects are used when IV levels are randomly selected from a larger population, and you want to generalize results to that population.

Importantly, in order to generalize to the population of levels of the IVs, the levels need to be *randomly selected* from the population (just as subjects are randomly selected from the population of subjects when one wants to generalize results to the population of subjects).

It is enlightening to consider the math. Imagine for a minute that we have only one IV (e.g. for some reason we look at mood gain based *only* on age – we do not compare against placebo). The one-way ANOVA model is:

$$Y_{ik} = \mu + \alpha_k + \epsilon_{ik}.$$

This equation has two types of variables: $\mu$ is a *fixed parameter* whereas, $\epsilon_{ik}$ (the random error) is a *random variable*. Mathematically, that means it is described by a probability distribution (a normal distribution in this case): $\epsilon_{ik} \sim \mathcal{N}(0, \sigma_\epsilon)$, where $\sigma_\epsilon$ is the standard deviation of this distribution, giving the average size of random error. Now consider $\alpha_k$: if age is a fixed effect, then it is a fixed parameter like $\mu$; and if it is a random effect, then it is a random variable: $\alpha_k \sim \mathcal{N}(0, \sigma_\alpha)$, where $\sigma_\alpha$

is the standard deviation of the distributions of levels (i.e. roughly speaking the spread in ages in the population) .

Now, given that ANOVA was introduced as a method for studying significant differences between groups, you might be confused as to what the *purpose* of fitting a random effects model is, given that it's apparently not about significant differences! The answer is that the purpose is often to obtain estimates of the contributions that different experimental factors make to the overall variation of the data, as expressed by their variance. These contributions are called *variance components* – $\sigma_\epsilon$ and $\sigma_\alpha$ are examples of these. These become more relevant in so-called *mixed effects models*[11], in which one can add terms for many different (random and fixed) effects, and systematically study their variance components in turn.

Some additional examples:

- Medical trials: investigating the effectiveness of a treatment across multiple sites or hospitals, where each site is treated as a random effect.
- Educational research: studying the performance of students across different schools, where school-specific effects are treated as random.
- Ecological studies: examining the growth rates of plants in different environments, where environmental conditions are considered random effects.

### 2.4.4 Latin-square design

It should be noted that it often matters in what order we present different levels of IVs. For within-subject variables, subjects might gain practice or get fatigued as the experiment proceeds, such that one performs differently later. In such cases, one can *randomize the order of the IV levels* using a *Latin-square design.* Table 2.16 shows how one can show different levels of a variable $A$ in different orders. Note that this is different from the design tables we have seen before, since it is filled with levels of an IV, rather than subjects. Observe that *each row and each column only contain each level of $A$ once*. This is in fact the defining feature of a Latin-square design, which holds also for much larger designs.[12] (Observe that this is only possible when the number of row blocks = the number of column blocks = the number of levels in the IV.) Now, three F-tests can be done: one for the IV ($A$), for subjects if desired, and a test of the *order*. The effect of order (like the effect of subjects) is subtracted out of the error term.

*Table 2.16: Table showing order of levels for each subject in a one-way latin square design, with $a_i$ the levels of the variable, $s_i$ the subjects and the horizontal numbers at the top showing the order in which levels are measured/presented. Note that this is not a design table as we have used them earlier, since cells here contain levels instead of subjects.*

| | 1 | 2 | 3 |
|---|---|---|---|
| $s_1$ | $a_2$ | $a_1$ | $a_3$ |
| $s_2$ | $a_1$ | $a_3$ | $a_2$ |
| $s_3$ | $a_3$ | $a_2$ | $a_1$ |

### 2.4.5 Unequal $n$/unbalanced designs

A very important note about all the theory covered so far is that it holds only for *balanced designs*, i.e. cases where *the number of subjects in all cells ($n$) is equal*. When this is not the case, i.e. when we have *unequal $n$* (an *unbalanced design*), this immediately makes the analysis much more complicated. A number of problems occur:

- It affects the assumption of homogeneity of variance; one obtains an inflated type I error.
- In general, the different sums of squares are not additive anymore.

[11] Also simply called *mixed models*, not to be confused with mixed *designs* as discussed above!

[12] Latin-square designs can also be understood as more general RBD, where one does not just have a single nuisance variable, but *two* nuisance variables.

- In general, variance can often be attributed to more than one source.
- For factorial ANOVA it makes that the main and interaction effect are not independent.

Thus, it is clear that *having a balanced design is highly desirable*. However, in the real world there are of course cases when this is simply not possible: e.g. if one does an expensive experiment with patients but results are somehow destroyed and cannot be done again, one might have a data table with a missing value.

Then, a number of different strategies exist to remedy this, noting that none of them is completely satisfactory. One strategy is to randomly delete subjects from other cells, until $n$ is equal in all cells. Of course, this neglects data that can be very valuable. A better strategy might be a so-called *unweighted means analysis*. Several other strategies exist, which we do not have time to go through in this course. Hence, we refer to [2] for further reading on this topic.

# Chapter 3

# Regression

### Further reading for this chapter

[1]: Ch. 15 (linear regression).

[2]: 3.5 (simple linear regression). Ch. 5 (multiple regression). Ch. 6 (ANCOVA). Ch. 10 (logistic regression).

[3]: 4.1.4.1 (linear regression).

[7]: 3.1 (linear regression), 4.3.2 (logistic regression).

Regression is one of the most widely used techniques in all of science. It's worth learning! So, how does it connect to ANOVA? With regression we typically have a slightly different research question in mind: ANOVA focuses on finding significant differences, whereas regression focuses on predicting a DV from an IV. Another difference is that with regression, the IV needs to be continuous rather than categorical (this means one does not have *levels* in the IV, and that one does not use the word *factor* in regression).[1]

We consider two main types of regression: linear regression and logistic regression. They can be distinguished by the *type* of the DV: for linear regression it is continuous and for logistic regression it is *binary*, meaning categorical with two levels. Then, we discuss *analysis of covariance* (ANCOVA), which can be seen as a combination of linear regression with ANOVA.

[1]Note that the underlying model and mathematics are in fact identical – something we discuss further in Ch. 4.2.

## 3.1 Linear regression

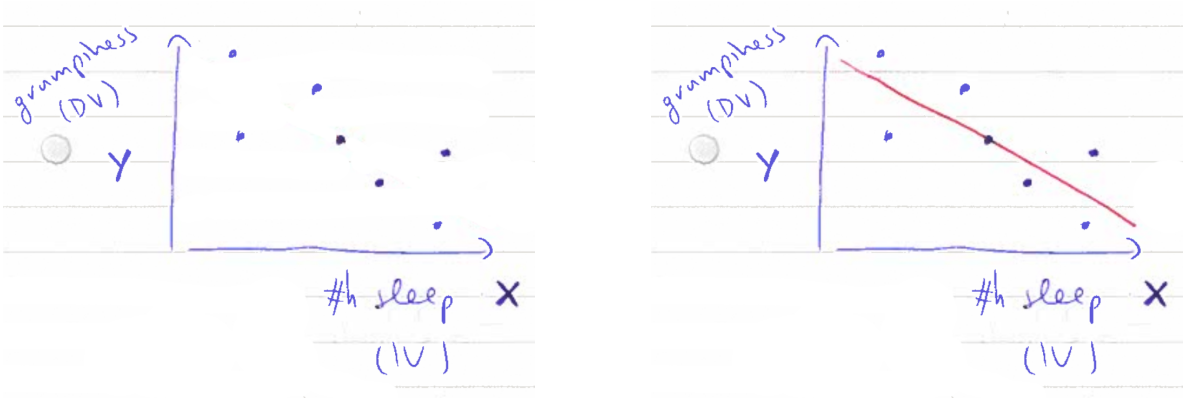
Let's consider some examples of linear regression in different fields.

- Medicine: predicting effectiveness of treatment from the dosage.
- HCI: predicting user satisfaction from the responsiveness of the system (e.g. a website)
- Real estate: predicting the price from the location and the number of rooms.
- [1]: predicting Danielle Navarro's grumpiness level from the hours of sleep she slept.

In all these cases, we want to predict a DV from an IV. Hence in regression we often use the term *predictor* instead of IV. We discuss below *simple* linear regression, in which you have a single predictor. Then we discuss *multiple* regression, in which you can have any number.

### 3.1.1 Simple linear regression

How does regression work, then? Taking Danielle Navarro's example [1], assume we have collected a bunch of data points on how grumpy she was (on a scale from 1 to 10, say) together with how many hours she slept that night. We could plot these points on a graph as in Fig. 3.1a – where we will always plot the DV on the $y$-axis and the IV on the $x$-axis. The point of linear regression then is to find the red line in Fig. 3.1b.




*(a) Datapoints.* *(b) Regression line plotted through the datapoints.*

*Figure 3.1: Regression example: Danielle Navarro's grumpiness together with how many hours she slept.*

> The idea of linear regression is to find the *best fitting straight line.*

Let's dissect this statement.

#### A straight line

First, let's remind ourselves of what a straight line is in mathematical terms. In your math classes in high school, you likely used $y$ for the DV and $x$ for the IV, such that a straight line is described by $y = ax + b$ where $a$ is the *slope* of the line and $b$ is the *intercept*. Together, $a$ and $b$ may be called the parameters or *coefficients* of our line. This is illustrated in Fig. 3.2. Note that the letters used for variables here are just a convention, and that many conventions exist, like $y = mx + c$ with $m$ the slope and $c$ the intercept. Here, we will use notation that is most often found in statistical literature, namely

$$\boxed{\hat{Y} = b_1 X + b_0} \tag{3.1}$$

where $b_1$ is the slope and $b_0$ is the intercept. Importantly, note that we use a hat for the DV in this equation. This is to emphasize that our line describes a *fitted model* or *prediction*. We use $Y$ without

hat to describe the *true values* of the DV, the actual datapoints or measurements. Furthermore, note that for some *specific* datapoint $(X_i, Y_i)$ with $i = 1, 2, ..., N$, we can write $\hat{Y}_i = b_1 X_i + b_0$.

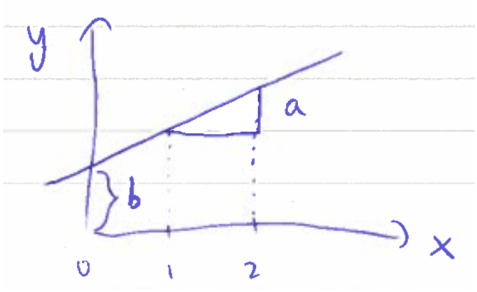



*Figure 3.2: Slope and intercept of a straight line illustrated.*

### Residuals

Say we have finished our regression analysis, and we have a prediction $\hat{Y}_i$ together with a true value $Y_i$. The *difference* between these,

$$\epsilon_i = \hat{Y}_i - Y_i,$$

is called the *residual* or *error*.[2] This is illustrated in Fig. 3.3.

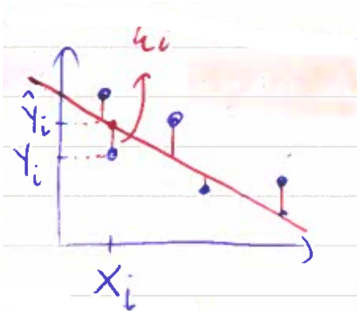



*Figure 3.3: Illustration of errors/residuals in linear regression.*

### The best fit

Great, we have reminded ourselves of what a straight line is. Now, how do we find the best fit?

> The best fitting straight line is described by the $b_0, b_1$ that minimize the sum of squared residuals, $\sum_i \epsilon_i^2$.

This is also called *ordinary least squares* (OLS). It seems sensible: the lower the total sum of errors, the better the line fits.[3] Mathematically, we can view this sum of residuals as a function of $b_0$ and $b_1$ – minimizing this then means finding the global *minimum* of this function. This can in turn be done by calculating the *derivatives* (or: gradient) of the function with respect to the coefficients, and setting the results to zero.

So, to proceed we need to do some mathematics. Following the philosophy discussed in the Preface, we shall not shy away completely from the math. We will show you how to do *derive* equations for $b_1$ for the simplest case in which we assume that the intercept $b_0 = 0$ (i.e. our line goes through the origin). Note that this can almost never be assumed in practice! The point of this is to show you where the

[2] Compare this to residuals in ANOVA: we had $Y_{ik} = \mu_k + \epsilon_{ik}$ where the residuals were $\epsilon_{ik}$. Notice that we can write this as $\epsilon_{ik} = \mu_k - Y_{ik}$, i.e. the 'prediction' in ANOVA is the group mean which we compare to the true values $Y_{ik}$.

[3] If you are wondering why we use square instead of just the sum, this is simply because it makes the mathematics simpler.

equations come from. For the general case we will simply provide the expressions which can be used in calculations. We recommend trying to derive the general expressions.

**Simplest case: line through the origin**

We are given a set of $N$ datapoints: $(X_1, Y_1), (X_2, Y_2)...(X_N, Y_N)$. To minimize $\sum_i \epsilon_i^2$, we take the derivative with respect to $b_1$ and set the result to zero:

$$\frac{d}{db_1}\left(\sum_i \epsilon_i^2\right) = 0$$

Now we fill in the definition of $\epsilon_i$, and then our model for $\hat{Y}_i$:

$$\sum_i \epsilon_i^2 = \sum_i \left(\hat{Y}_i - Y_i\right)^2 = \sum_i \left(b_1 X_i - Y_i\right)^2$$

Then, we take the derivative of this expression with respect to $b_1$. We use the sum rule for derivatives, and the chain rule:

$$\begin{aligned}\frac{d}{db_1}\left(\sum_i \left(b_1 X_i - Y_i\right)^2\right) &= \sum_i \frac{d}{db_1}\left(\left(b_1 X_i - Y_i\right)^2\right) \\ &= \sum_i 2\left(b_1 X_i - Y_i\right)\left(X_i\right) \\ &= 2\sum_i \left(b_1 X_i^2 - X_i Y_i\right)\end{aligned}$$

This expression should equal zero. We can divide by 2 to get:

$$\begin{aligned} & b_1 \sum_i X_i^2 - \sum_i X_i Y_i = 0 \\ \iff & b_1 \sum_i X_i^2 = \sum_i X_i Y_i \\ \iff & b_1 = \frac{\sum_i X_i Y_i}{\sum_i X_i^2}\end{aligned}$$

So, we see we have an equation for calculating $b_1$ using the datapoints!

**The general case**

As mentioned, we almost never assume the intercept is zero. If we keep the equation general, one can do a similar (slightly longer) derivation for both $b_0$ and $b_1$, which gives the following expressions:

$$\boxed{\begin{aligned} b_0 &= \frac{\sum_i X_i^2 \sum_i Y_i - \sum_i X_i \sum_i X_i Y_i}{N\sum_i X_i^2 - \left(\sum_i X_i\right)^2} \\ b_1 &= \frac{N\sum_i X_i Y_i - \sum_i X_i \sum_i Y_i}{N\sum_i X_i^2 - \left(\sum_i X_i\right)^2}. \end{aligned}} \tag{3.2}$$

These can be used for calculations, and are used under the hood whenever you use R/SPSS/Python to perform regression for you.

**An example**

Fig. 3.4 illustrates a simple example for simple linear regression.

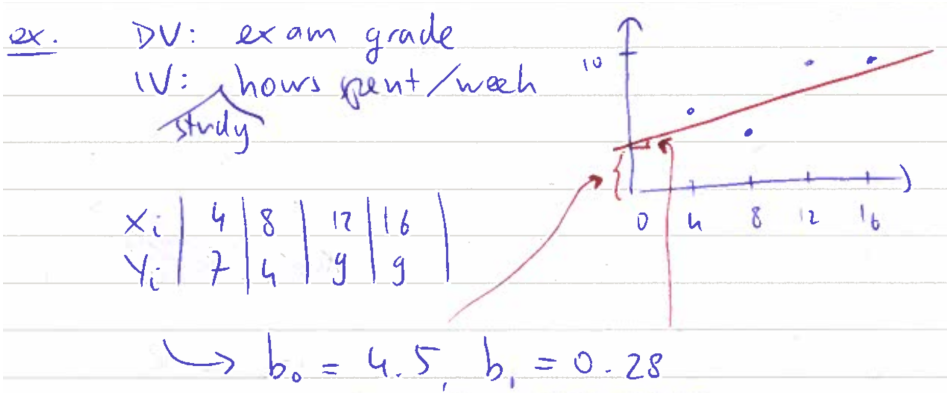



*Figure 3.4: Simple linear regression example. We want to predict students' exam grade from the weekly average number of hours spent studying. We might collect a table of data as given, from which we can calculate the slope and intercept directly using Eqs. 3.2 (calculations are not shown). We could plot this line together with data points as shown.*

### 3.1.2 Multiple regression

In multiple regression we generalize the results from the previous section to any number of IVs (similar to how we generalized from ANOVA to factorial ANOVA). For one IV, $X$ we had the following model (Eq. 3.1):

$$\hat{Y} = b_0 + b_1 X$$

Now we replace $X$ by *two* IVs, $X_1$ and $X_2$:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

where we have added a second slope coefficient $b_2$. For instance, in the example of predicting student's exam grade from their study time, we might want to additionally consider the number of lectures at which they were present. Where we could visualize this using a straight *line* before, this is now generalized to two dimensions to a *plane*, illustrated in Fig. 3.5.

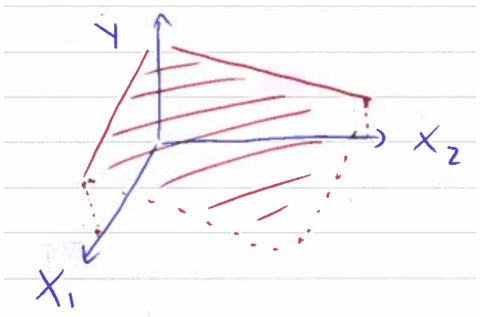



*Figure 3.5: Illustration of the prediction of a single DV from two IVs $X_1$ and $X_2$.*

Of course, we can generalize this further to $K$ IVs, $X_1, X_2, ..., X_K$:

$$\boxed{\hat{Y} = b_0 + \sum_{k=1}^{K} b_k X_k} \tag{3.3}$$

where $b_k$ is the $k$-th regression coefficient, i.e. we get $b_1, b_2, ...b_K$. Although visualizing this becomes hard, this equation still allows us to predict the DV from IVs if we have calculated the coefficients – the general idea remains unchanged. Equations for finding $b_1, b_2, ...b_K$ fall outside the scope of our course[4] – but you should be able to perform multiple regression using SPSS/R/Python.

[4]For readers with a background in linear algebra, you can show that the vector of coefficients is equal to $\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y}$ where $X$ is the data matrix with predictors as rows and datapoints as columns.

#### Standardized coefficients

Let's return to the example of predicting exam grades from study time (IV1) and presence during lectures (IV2). Say we have found coefficients and we have a nice model which we can use to predict the exam grade. Now, we might ask: which helps students more, more study time or being present more often?

Answering this question can be done by calculating *standardized coefficients*, which are a (signed) measure of *feature importance* (here, we use *feature* as a synonym for predictor and IV). Standardized coefficients $\beta_k$ are defined as:

$$\beta_k = b_k \frac{s_k}{s_Y}$$

where $s_Y$ is the *sample standard deviation* of $Y$ (see Ch. 1), and $s_k$ is the sample standard deviation of predictor $X_k$ (we could write $s_{X_k}$ to get somewhat more consistent notation, but this becomes hard to read, so we stick with $s_k$). If we have calculated these, we can compare predictors directly. Say that study time is $X_1$ and lecture presence is $X_2$, we might find that $|\beta_2|$ is larger than $\beta_1$, and conclude that attending lectures improves students' grades more (on average) than studying more.

### 3.1.3 NHST

We have discussed how to get results with regression: how to find the coefficients that describe the best fitting straight line, and how to compare them in multiple regression. But what about the statistical significance of these results? Indeed, this can be done using null hypothesis significance testing (NHST). There are two basic ways to do this: the first is to test significance of the model as a whole, and the second is to test significance of individual coefficients.

#### The whole model

If we want to test significance of our regression model as a whole, the hypotheses are the following:

$$\boxed{\begin{aligned} H_0: &\quad Y_i = b_0 + \epsilon_i \\ H_1: &\quad Y_i = b_0 + \sum_{k=1}^{K} b_k X_{ki} + \epsilon_i \end{aligned}} \tag{3.4}$$

Interestingly, testing this is done using an F-test, like with ANOVA! The sums of squares are very similar to those in one-way ANOVA. The *model* sum of squares, *residual* sum of squares and *total* sum of squares are:

$$\begin{aligned} \mathrm{SS}_{\mathrm{model}} &= \sum_i (\hat{Y}_i - \bar{Y})^2 \\ \mathrm{SS}_{\mathrm{res}} &= \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i \epsilon_i^2, \\ \mathrm{SS}_{\mathrm{tot}} &= \sum_i (Y_i - \bar{Y})^2 \end{aligned}$$

where $\bar{Y}$ is the (grand) mean as before (note that we do not have levels in regression, so we have no group means!). We again have $\mathrm{SS}_{\mathrm{tot}} = \mathrm{SS}_{\mathrm{model}} + \mathrm{SS}_{\mathrm{res}}$. The corresponding degrees of freedom are:

$$\begin{aligned} df_{\mathrm{res}} &= N - K - 1 \\ df_{\mathrm{model}} &= K \\ df_{\mathrm{tot}} &= N - 1 \end{aligned}$$

And to perform an F-test, we finally calculate as follows:

$$\mathrm{MS}_{\mathrm{mod}} = \frac{\mathrm{SS}_{\mathrm{mod}}}{df_{\mathrm{mod}}}, \quad \mathrm{MS}_{\mathrm{res}} = \frac{\mathrm{SS}_{\mathrm{res}}}{df_{res}}, \quad F = \frac{\mathrm{MS}_{\mathrm{mod}}}{\mathrm{MS}_{\mathrm{res}}}$$

That's all there is to it!

**Individual coefficients**

For testing the significance of a coefficient $b_i$, our hypotheses are:

$$\boxed{\begin{aligned} H_0: &\quad b_i = 0 \\ H_1: &\quad b_i \neq 0 \end{aligned}} \tag{3.5}$$

Testing this is done with a 'special t-test', as follows. The test statistic is:

$$t = \frac{b_i}{\text{SE}(b_i)}$$

with $df = N - K - 1$. One can find a confidence interval as

$$\text{CI}(b_i) = b_i \pm t_{crit}\text{SE}(b_i).$$

where $t_{crit}$ is the critical value of the t-distribution for the appropriate significance level (recall that $\alpha = 0.05$ for a CI of 95%). Note that when you have a single predictor, doing a t-test for $b_1$ gives the same result as an F-test for the whole model, since Eq. 3.4 and Eq. 3.5 are then the same hypotheses!

In these equations $\text{SE}(b_i)$ is the *standard error of the intercept*. What does that mean? Well, this is a bit complicated, so we shall defer to using SPSS/R/Python for this type of test. We give this equation regardless to illustrate a connection with *correlations*. Namely, when we have one predictor (i.e. $K = 1$), the coefficient $b_1$ is equivalent to the Pearson correlation $r$! Thus, we see we can also talk about the *significance of a correlation*, using a t-test statistic calculated as follows:

$$t = \frac{r}{\text{SE}(r)}$$

with $df = N - 2$!

### 3.1.4 Effect size

Next up, you guessed it, effect size. For regression, we talk about $R^2$ or *R-squared*, which is defined as follows:

$$\boxed{R^2 = \frac{\text{SS}_{\text{model}}}{\text{SS}_{\text{tot}}} = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}.}$$

One problem with this definition of effect size is that it always increases when increasing the number of predictors. Hence, one often uses the *adjusted R-squared*, defined as:

$$\boxed{\text{adjusted } R^2 = 1 - \frac{\text{MS}_{\text{res}}}{\text{MS}_{\text{tot}}}.}$$

i.e. we swap sums of squares with mean squares, thus taking the degrees of freedom into account and correcting for an increase in $K$. **NOTE** that a disadvantage of this notion of effect size is that *it cannot be interpreted as a proportion of variance*! (Similar to partial eta squared, which we had for factorial ANOVA.)

### 3.1.5 Assumptions

The assumptions for regression are shown in Table 3.1. A1-A3 are assumptions we have seen before. New assumptions are the following:

- *linearity* (A4): the assumption that the relationship between $X$ and $Y$ is indeed linear (i.e. it is well-described by Eq. 3.3.) It can be tested using a *curvature test*.
- *no outliers* (A5), discussed below.
- *cases to IVs* (A6): a heuristic that tells you how many datapoints $N$ you should have compared to the number of predictors, $K$. This depends on the test that you are performing.

- *collinearity* (A7), discussed below.

Many of the can be checked by a *residuals plot*, illustrated in Fig. 3.6. It illustrates how residuals might look when assumptions are met and when A1, A2, A4 and A5 are violated. As before, independence (A3) is difficult to check in practice.

*Table 3.1: Regression assumptions. Note that is only relevant for multiple regression (i.e. $K > 1$).*

| | Assumption | Graphical/heuristic check | Statistical test |
|---|---|---|---|
| A1 | **Normal residuals** | Residuals plot;<br>QQ-plot | Shapiro-Wilk test |
| A2 | **Homogeneity of variance** | Residuals plot | Non-constant variance test |
| A3 | **Independent residuals** | - | - |
| A4 | **Linearity** | Residuals plot | Curvature test |
| A5 | **No outliers** | Residuals plot;<br>Cook's distance;<br>Leverage plot | - |
| A6 | **Cases ($N$) to # IVs ($K$)**[5] | Whole model: $N \geq 50 + 8K$;<br>Individual coeff.: $N \geq 104 + K$ | - |
| A7 | **Collinearity**[6] | Correlations & VIFs | - |

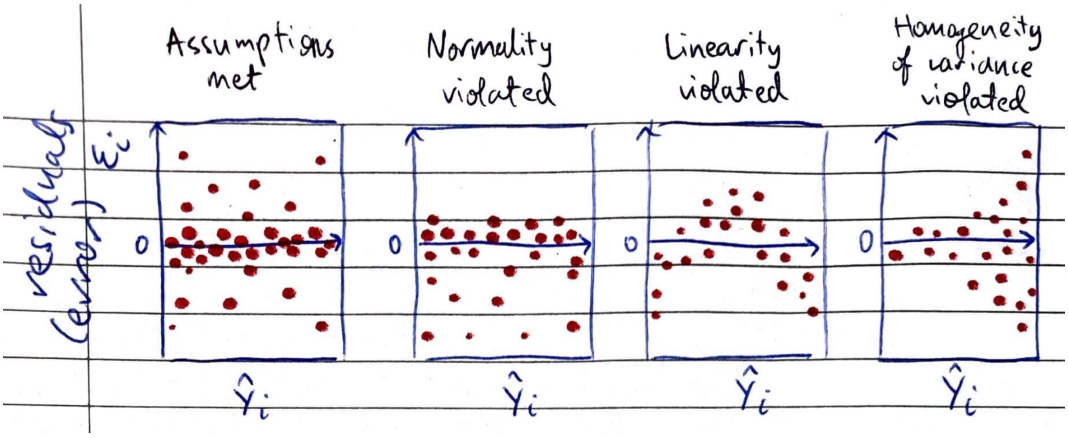



*Figure 3.6: Residuals plots (adapted from [2]). The four plots show residuals $\epsilon_i$ as a function of the fitted DV, i.e. $\hat{Y}_i$, when assumptions are met, when normality is violated, when linearity is violated and where homogeneity of variance is violated.*

### Outliers, leverage and influence (A3)

For A3 we discuss the heuristic checks in some detail. First, observe the middle graph in Fig. 3.7, where data is plotted as $(X_i, Y_i)$. This shows one datapoint $(X_i, Y_i)$ with large *leverage*: a datapoint very different from all the others $(X_j, Y_j)$ where $j \neq i$. The leverage of points can be quantified using

[5] Note: this is not a technical assumption but a practical requirement. Also note that the heuristic checks assume a medium-size relationship between the IVs and the DV, $\alpha = .05$, $\beta = .20$ [2].

[6] Also known as uncorrelated predictors.

*hat values*, $h_i$. Calculating hat values by hand is a bit complicated but can be done easily using a computer. A large hat value means large leverage – but there is no objective point at which a hat value is 'too large': instead, look for points that have a larger hat value than all the others.

Now consider the leftmost graph in Fig. 3.7. It shows a single datapoint with a very large residual. This seems to match our intuition for what it means for a point to be an '*outlier*' – suggesting we might *define* an outlier this way. However, since whether a residual is 'large' depends on the *scales* used, it is often more useful to use *standardized residuals*, which scale residuals to have zero mean and standard deviation one (in other words, they take into account the variance of the residuals). If the (standardized) residual is large compared to all others, this suggests an outlier. It is important to investigate such points.

Finally, consider the right graph in Fig. 3.7. This shows a point with large *influence*: an outlier with large leverage. These are *often* a concern, and should definitely be investigated. Moreover, whereas outliers and leverage have no agreed upon limit at which a point is deemed problematic, this does exist for influence, combining both resiudal and leverage. Influence of points are quantified using *Cook's distance* $D_i$, where we say influence is 'large' when $D_i \gtrapprox 1$. (Next to calculating Cook's distances directly, one can also plot leverage against residuals in a so-called *leverage plot*, which in R also shows Cook's distances of 0.5 and 1.0, giving clear boundaries at which points are likely problematic).


*Figure 3.7: Outlier, high leverage and high influence illustration (adapted from [1]). Striped and filled lines indicate the regression line without, and with this point, respectively.*

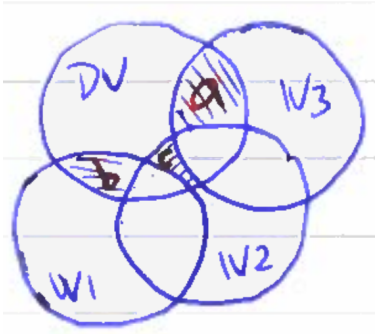
#### Collinearity (A7)

The final assumption we have to explain is A7: collinearity, or (un)correlated predictors. Put simply, if we have collinearity we have *redundancy among predictors*. We can check this by simply calculating correlations between all predictors. A heuristic is that if we get $r \gtrapprox 0.7$ between predictors this suggests a problem. A second way of checking this is using *variance inflation factors*, or VIFs. The square root of a VIF tells you, for each coefficient $b_k$, *how much wider the confidence interval for* $b_k$ *is, compared to what it would be with uncorrelated predictors.* If you have $VIF \gtrapprox 5$ this suggests a problem.

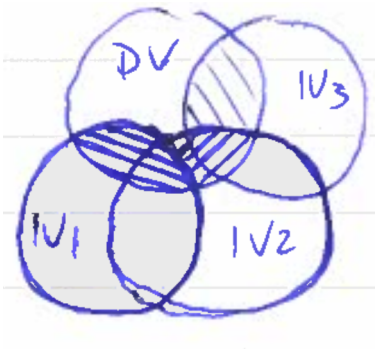
### 3.1.6 Sequential multiple regression

Consider Fig. 3.8a, where we have illustrated the (shared) variance of three IVs and one DV. If one uses multiple regression for this scenario, and calculates (adjusted) $R$-squared, the following happens: IV1 gets credit for $b$, IV2 for $c$ and IV3 for $a$. Two problems can thus be identified in this approach. The first is that IV2 has large overlap with the DV, but appears unimportant if we calculate effect sizes. The second problem is more fundamental: we are using an a-theoretical 'shotgun' approach, treating all IVs on an equal footing. This is often not appropriate, because we are usually more interested in a certain effect than in others (similar to how we were not interested in nuisance variables in Section 2.3.1). More generally, we might have certain logical or theoretical reasons to assign an *order* to IVs.

A solution to this is called *sequential regression* (also called hierarchical regression), illustrated in Fig.




*(a) Standard multiple regression. Shared variance between IV1 and IV2 does not explain variance in the DV.*




*(b) Sequential regression. IV1 takes precedence over IV2 and IV3, and IV2 takes precedence over IV3, such that all variance of IVs that helps explain variance in the DV is taken into account.*

*Figure 3.8: Different ways of treating shared variance between IVs in multiple regression.*

3.8b.[7] It does precisely what we would like: it provides the means to assign an order to IVs based on a given theoretical framework or specific research question. For instance, in the example of predicting students' exam grade, imagine we have three IVs: (1) the average grade in other courses, (2) the presence during lectures, and (3) the number of passed assignments. Using sequential regression with this order gives a systematic way of studying the variance explained by adding IVs sequentially.

[7]Note that this is is not the same as *stepwise* or *statistical regression*, a more controversial technique, for which we refer to [2] or other more advanced textbooks on statistics.

## 3.2 Logistic regression

As was mentioned, with regression we try to predict a *continuous* variable. This can be contrasted with *classification*, in which we try to predict a *discrete* variable.

The simplest method for regression is linear regression. The simplest method for *classification* is (confusingly) called *logistic regression* (the reason is that the underlying mathematics of logistic regression is the same as that of linear regression).

Logistic regression is widely used in a variety of fields. For instance, we might want to classify an e-mail as spam or no spam based on its content. Or, in the medical domain, we might want to classify an image of a tumor as either benign or malignant, based on image features. Logistic regression also forms the basis for the mathematics of *neural networks*, hence also for modern *deep learning* methods. We shall give a very brief introduction.[8]

### Classification

Recall that for linear regression we had $\hat{Y} = b_0 + \sum_{k=1}^{K} b_k X_k$, and that both $Y$ and $\hat{Y}$ are continuous, i.e. they can be *any real number*. With logistic regression, $Y$ and $\hat{Y}$ are restricted to be *either 0 or 1*, where 0 is used as the 'negative' class (e.g. spam) and 1 for the 'positive' class (e.g. no spam).

> Logistic regression takes the result of linear regression and maps it to the interval $[0, 1]$.

Mathematically, this means we define a variable $z$ as:

$$z = b_0 + \sum_{k=1}^{K} b_k X_k$$

(note that this is the 'result' of linear regression, where we have replaced $\hat{Y}$ by $z$). Then, we use this as an input to the *sigmoid* function, also called *logistic* function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

This function does exactly what we want: it maps any real number to $[0, 1]$. This can be observed in Fig. 3.9: for negative values, it asymptotes to zero, while for positive values it asymptotes to one (large positive values of $z$ are mapped close to 1, and large negative values are mapped close to 0). Finally, we assign $\hat{Y}$ to one if $\sigma(z) \geq 0.5$ and to zero if $\sigma(z) \leq 0.5$. I.e. we can write:

$$\hat{Y} = \begin{cases} 1 & \text{if } \sigma(z) \geq 0.5 \\ 0 & \text{if } \sigma(z) < 0.5. \end{cases}$$

This makes sure that our values are mapped to *either* 0 and 1, instead of any value in the interval [0,1]. Thus, given a datapoint $X_{ki}$, we can obtain a prediction for its class $\hat{Y}$ using the given equations, and we have performed a *classification* of $X_{ik}$.

### A probability model

The above section illustrated how logistic regression provides a means to classify new examples from an input $X$. However, in addition, logistic regression is inherently also a model of *probabilities*. Imagine we are given a datapoint $X_i$, for which we are modelling $\hat{Y}$. The sigmoid function gives a model for the *probability that* $\hat{Y}_i = 1$ *given* $X_i$, i.e.:

$$P(\hat{Y}_i = 1 | X_i) = \sigma(z).$$

Since there are two classes and probabilities sum to one, we automatically have that:

$$P(\hat{Y}_i = 0 | X_i) = 1 - P(\hat{Y}_i = 1 | X_i) = 1 - \sigma(z).$$

[8] Refer to other UU courses on Data Mining and Pattern Recognition for more. Good books include [8] and [7].

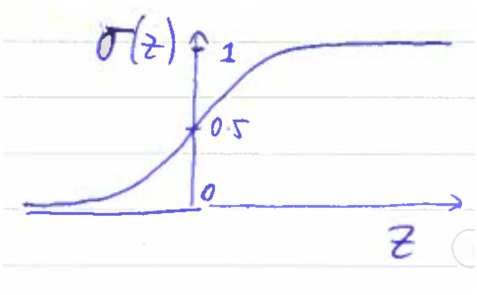



*Figure 3.9: Sketch of the sigmoid function.*

### Finding coefficients

Now, you might be wondering: how do we find the coefficients $b_k$ for logistic regression? Can we use OLS like before? No: things are a bit more complicated. We need to use a so-called *cost* or *loss* function, then do numerical minimization of this function (typically using a technique called *gradient descent*) to find these.[9]

### Multi-class classification

You might also be wondering: what if we our DV is not binary, but has more than two classes? For instance, we might want to classify the weather as sunny, cloudy or rainy, based on various meteorological features that have been measured. How can we do such *multi-class classification*? There are two basic methods for doing this. The first is to do 'one-vs.-all' logistic regression several times. For instance, we might first group rainy and cloudy together as 0 with sunny as 1, followed by rainy and sunny as 0 with cloudy as 1 – and so on. This is illustrated in Fig. 3.10. The second method is to use so-called *softmax regression*.

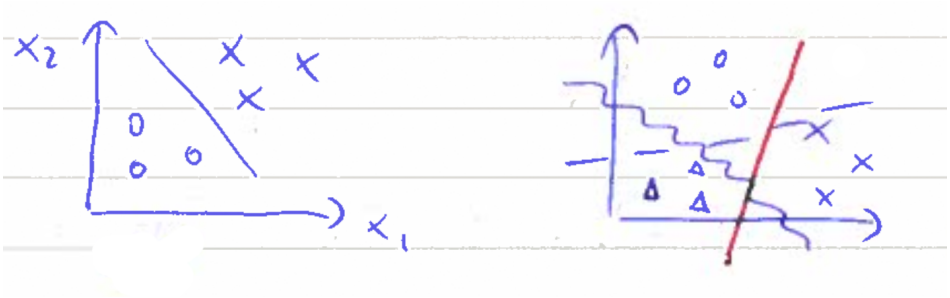



*Figure 3.10: Binary classification vs. multi-class classification using one-vs.-all methods.*

### Neural networks

As a final note, we mention that *neural networks* can be seen as a network of logistic regression units! Consider a very simple neural network, illustrated in Fig. 3.11, with three input features/predictors $X_1, X_2, X_3$, an output variable $Y$, and a single so-called *hidden layer*, with *activation nodes* (or: *neurons*) $a_1$ and $a_2$. The connections between nodes are all assigned a *weight* (which is exactly a coefficient $b_k$ as before). For this network, one calculates the activation nodes as follows:

$$a_1 = \sigma(b_{11}X_1 + b_{21}X_2 + b_{31}X_3)$$
$$a_2 = \sigma(b_{12}X_1 + b_{22}X_2 + b_{32}X_3),$$

where we can see how, at each node, we perform a type of logistic regression. Of course, in *deep* learning we include many more layers, as well as neurons per layer, and extend the model in various other ways. But perhaps this shows that neural networks are in a sense less complicated than they might seem!

[9] See aforementioned courses on pattern recognition and machine learning for more on this!

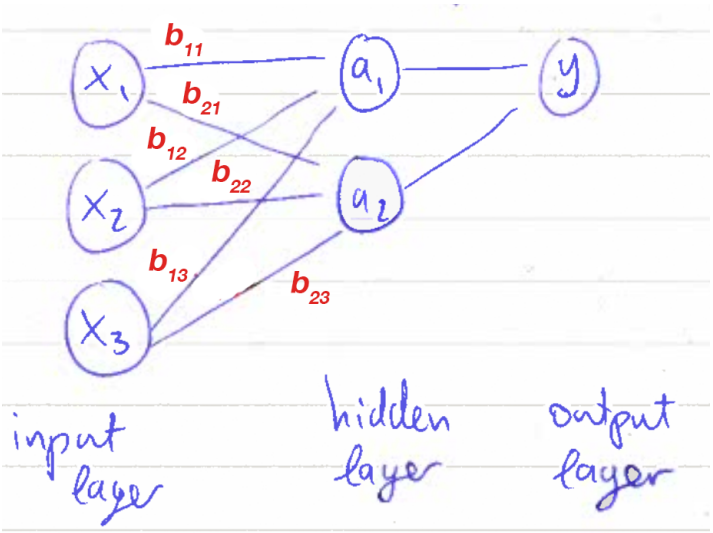



*Figure 3.11: Simple neural network illustration.*

## 3.3 ANCOVA

Having tackled regression, we go on to *analysis of covariance*, or *ANCOVA*. Understanding ANCOVA after the previous chapters is quite easy: namely, on a theoretical level, ANCOVA is nothing more than a *combination of regression and ANOVA*! On a practical level, however, the technique that comes closest is RBDs (Section 2.3.1): it is a method to adjust/control for CVs/NVs. The key difference with RBDs, then, is that the CV is now *continuous*.

### 3.3.1 Basics

Let us now consider again the example we had when discussing RBDs. We considered *mood gain* as the DV, with *medication* as IV (levels Joyzepam/Anxifree/Placebo). We also had *age group* as a CV (or nuisance variable). Now, we imagine we no longer have age groups, but simply people from all ages, distributed randomly. Now, ANCOVA might be appropriate! So how does it work?

The idea of ANCOVA is to:

1. do *regression* with the CV on the DV
2. do ANOVA with *what's left* (i.e. the *residuals*).

Using equations, we can sketch this as follows.

1. $\hat{\text{DV}}_i = b_1 \text{CV}_i + b_0$; $\quad \text{DV}_i - \hat{\text{DV}}_i = \epsilon_i$
2. ANOVA with $\epsilon_i$

Notice how the first step has the form $\hat{Y}_i = b_1 X_i + b_0$, but with $Y$ now as the DV and $X$ as the CV. In a slogan, we can say that:

'ANCOVA = regression + ANOVA'

Some examples cases of using ANCOVA are shown in Table 3.2.

### 3.3.2 ANCOVA vs. RBDs

ANCOVA and RBDs are similar in spirit: just like RBDs, ANCOVA controls for the effect of CVs thus reducing the error term (refer to Section 2.3.1, *Reducing the error term*, and particularly Fig. 2.18,

*Table 3.2: Examples for ANCOVA variable in different domains.*

| Domain | DV | IV | CV |
|---|---|---|---|
| *Politics* | Voting behavior | Region | Socioeconomic status |
| *Education* | Grade | Teaching | Pre-course score |
| *HCI* | User satisfaction | Website layout | User prior knowledge |

which holds exactly for ANCOVA too). There are also three important differences, however.

1. As mentioned, *the CV in ANCOVA is necessarily continuous, where for RBD it is discrete* (although continuous variables can of course be discretized into levels, like we discretized age to get a number of age *groups*).
2. Say that we performed the aforementioned of medication without realizing the effect of age (group) on the results. Now, if we have this realization after doing an experiment, *ANCOVA can typically still be done post hoc*. That is, no new experiment is required, and we can simply collect ages after the expriment and perform ANCOVA as outlined. In contrast, the use of RBD requires *doing a new study*, in which subjects are explicitly blocked in the right way. Clearly, this can be a disadvantage of RBD.
3. The advantage of an RBD is that *it is typically more powerful than ANCOVA*.

### 3.3.3 Further notes

#### Choosing CVs

We make a couple of important remarks on *choosing covariates*.

- *It is recommended to have a small number of covariates*. Including more covariates can reduce power and increase the risk of *overfitting*, leading to models that might describe the sample data well but perform poorly on new data.
- *All CVs should be correlated with the DV*. If it's not, that means it does not help in reducing the error term, making it unnecessary or even detrimental to include it in the model.
- *None of the CVs should be correlated with each other* (this is the assumption of collinearity, discussed in Section 3.1.5).
- *CVs should be independent of the IV*. If it's not, changes in the DV could be attributed to the CV rather than the IV, leading to false conclusions about the effect of the IV (so-called *confounding effects)*. CVs should be used to control for *pre-existing* differences among subjects, before the experimental manipulation has happened. This means CVs should be measured *before* the IV whenever possible, to prevent it from being influenced by the IV.

#### Assumptions

The assumptions for ANCOVA are the following. First, one has the *regression assumptions* for the CV-DV relationship (Table 3.1), and the *ANOVA assumptions* for the IV-DV relationship (Table 2.1). We also have *two additional* assumptions for ANCOVA specifically:

- *Reliability of CVs*.
- *Homogeneity of regression*.

For reliability of CVs, we quote from section 6.2.3.8 in [2]:

> In ANCOVA, it is assumed that CVs are measured without error, and that they are perfectly reliable. In case of such variables as sex and age, the assumption can usually be justified. With self-report of demographic variables and with variables measured psychometrically, such assumptions are not so easily made. And variables such as attitude may be reliable at the point of measurement, but fluctuate over short periods.

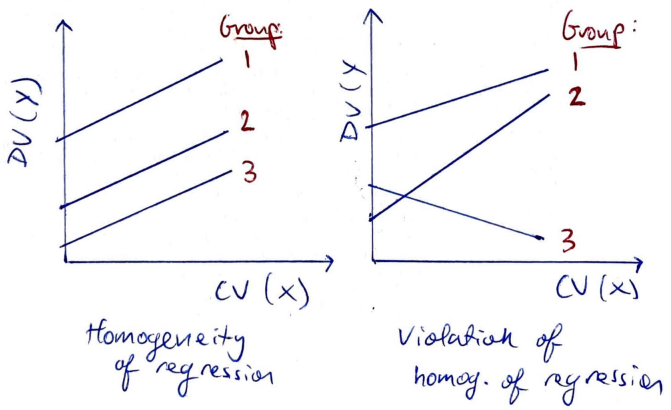



*Figure 3.12: Illustration of the assumption of homogeneity of regression (left) vs. violation of this assumption (right) in ANCOVA. The different groups/levels of the IV are shown with different lines: for all lines, one should have (approximately) equal slopes.*

Homogeneity of regression is illustrated in Fig 3.12. For each (combination of) level(s) of the IV(s), one wants the CV-DV regression line to have (approximately) equal slope.

### 3.3.4 An example

We are given Table 3.2, and we ask: *How does diet type (Mediterranean, vegetarian, control) affect weight loss when controlling for initial weight?* Our DV is thus the weight-loss; the IV is the diet (2x diet, 1x control), and the CV is the weight before the diet program. The result of doing ANCOVA is shown in Fig. 3.4. It is interesting to compare this to what you would find if simply used ANOVA, without correcting for pre-treatment scores. This is shown in Fig. 3.5. Observe that the result now is no longer significant.

*Table 3.3: Small-sample data for illustration of ANCOVA.*

| Vegetarian | | Mediterranean | | Control | |
|---|---|---|---|---|---|
| Pre | Loss | Pre | Loss | Pre | Loss |
| 82 | 5.1 | 85 | 4.4 | 91 | 4 |
| 79 | 4.9 | 87 | 5.9 | 81 | 1.3 |
| 84 | 3.2 | 90 | 6.6 | 82 | 2.6 |

*Table 3.4: Resulting F-table after ANCOVA.*

| | df | Adjusted SS | MS | $F$ | $p$ |
|---|---|---|---|---|---|
| Diet | 2 | 13.642 | 6.821 | 5.981 | 0.0472 |
| PreWeight | 1 | 2.651 | 2.651 | 2.325 | 0.1878 |
| Residuals | 5 | 5.702 | 1.14 | | |

*Table 3.5: F-table after ANOVA instead of ANCOVA, not correcting for the CV.*

| | df | SS | MS | $F$ | $p$ |
|---|---|---|---|---|---|
| Diet | 2 | 13.642 | 6.821 | 4.899 | 0.0548 |
| Residuals | 6 | 8.353 | 1.392 | | |

# Chapter 4

# Multivariate statistics & the GLM

## Further reading for this chapter

[2]: 1.1 (preliminaries), Ch. 7 (MANOVA), Ch. 17 (general linear model)

[1]: 16.6 (linear models)

Techniques discussed in earlier chapter have always involved a single DV. Going beyond this to any number of DVs and IVs, we enter the domain of *multivariate statistics* (Table 4.1). This chapter gives an introduction to *multivariate analysis of variance*, or *MANOVA*. In order to introduce this technique, it is useful to first discuss two preliminaries: the distinction between experimental and non-experimental research, and linearity.

*Table 4.1: Main univariate and multivariate methods covered in these notes. Methods in italic are not discussed.*

| | **Univariate statistics (1 DV)** | | **Multivariate statistics (>1 DV)** | |
|---|---|---|---|---|
| | **1 IV** | **>1 IV** | **1 IV** | **>1 IV** |
| **Discrete IV** | ANOVA | Factorial ANOVA | MANOVA | Factorial MANOVA |
| **Continuous IV** | Regression | Multiple regression | *Canonical correlation* | |
| **Mixed** | *Multilevel modelling* | | | |

### Experimental & non-experimental research

Within quantitative methods, there is an important distinction to be made between *experimental* and *non-experimental* research. Experimental research includes many of the example studies we have discussed so far: medical trials, user studies where websites are rated by participants, etc. In these cases, experiments are typically designed to answer *specific* research questions. Non-experimental research, on the other hand, includes e.g. opinion surveys, common in political science, and is more closely related to *data science*. Here, research questions might be *less specific*, but instead revolve around finding unknown patterns in large sets of data.

Other than this difference in research question, a key is that for experimental studies, the researcher has *control over the levels of the IV(s)*. This means that if subjects are randomly assigned to these levels, and if then a systematic difference in the DV is observed that is associated with the IV, then one can say (with some degree of confidence) that the change in the DV is *caused* by the IV! In contrast, with non-experimental research, one does not have this control, and establishing causality is very difficult.

This is captured by the famous adagium "correlation $\neq$ causation" (illustrated beautifully by Tyler Vigen's *Spurious Correlations*, https://www.tylervigen.com/spurious/random).

## Why use multivariate techniques?

We have seen that when one has multiple IVs, one gets an i*nflated Type I error rate* whenever several tests are done – which could be corrected for by using multiple comparison corrections (e.g. Bonferroni, Holm or Tukey). With multiple DVs, the same problem arises if each DV is tested separately. Now, however, multiple comparisons is typically insufficient, because *some of the DVs are often correlated with one another*. Hence, separate tests will also analyze some of the same variance. Multivariate techniques can take this into account by providing a *joint* analysis, while keeping the overall Type I error rate at the desired level. As such, multivariate techniques are more *comprehensive* than univariate techniques. Additionally, they sometimes provide increased power.

Most multivariate techniques were developed for non-experimental research, but with the widespread availability of computer programs, they are becoming increasingly common also in experimental research. Here, they allow the design of more *efficient* and *realistic* experiments. For example, in psychology, one might have two surveys that gives scores on two different but related constructs, e.g. anxiety and depression. With multivariate techniques, these scores can be analyzed together in a way that takes into account relationships between these constructs.

## Warnings

Although these are all real advantages of multivariate techniques, they also come with some important warnings. First, there is the idea of *garbage in, roses out?*. Quoting from [2]:

> The trick in multivariate statistics is not in computation. (...) The trick is to select reliable and valid measurements, choose the appropriate program, use it correctly, and know how to interpret the output. Output from commercial computer programs, with their beautifully formatted tables, graphs, and matrices, can make garbage look like roses.

Second, using multivariate techniques can be *slippery*, in the following sense. Due to the increased complexity of the analysis, the probability of making errors (e.g. in the code) increases. Interpreting results is also often ambiguous.

## Linearity

Linearity was mentioned briefly in the context of linear regression, where it was presented as one of the necessary assumptions. Here we provide a short, more general discussion of linearity. A *linear combination* of two variables $X_1$ and $X_2$ is formally defined as:

$$Y = w_1 X_1 + w_2 X_2$$

which is reminiscent of the equation we had in linear regression. Using more general terminology, $Y$ is the *composite* variable and $w_i$ are weights. The important point for linearity is that one has a *weighted sum* of the input variables; no powers (e.g. $X_1^2$), cross-products (e.g. $X_1 X_2$), etc. The more general field of study of such relationships is called *linear algebra*, which intuitively deals with the mathematics involving the notion of straightness and flatness.

## 4.1 MANOVA

In the same way as ANOVA is one of the foundational univariate statistical techniques, MANOVA can be seen as a foundational multivariate statistical technique. Some examples of where MANOVA might be appropriate are:

- *Social science*: studying the effect of age, gender & socioeconomic status (IVs) on self-esteem, anxiety & depression (DVs).
- *Medicine*: studying the effect of treatment (IV) on blood pressure, heart rate & cholesterol level (DVs).

So how does it work?

> The idea of MANOVA is to look at significant differences in group means *on a linear combination of DVs* (i.e. a composite DV).

Symbolically, for two DVs, this means:

$$\mathrm{DV}_{\mathrm{composite}} = w_1\mathrm{DV}_1 + w_2\mathrm{DV}_2,$$

where first (1) the weights are picked to *maximize the differences between group means*, and second (2) normal ANOVA is done with the composite DV.

As an example, one might imagine having a composite variable of 'health improvement', composed of the DVs weight loss and cardiovascular fitness, which is studied based on different exercise programs.

### Advantages and limitations

The advantages of and warnings for multivariate techniques described above hold for MANOVA too. An additional advantage is shown in Fig. 4.1, which sketches a possible joint distribution of data on two DVs, $Y_1$ and $Y_2$, for two levels of an IV ($X$), in blue and pink. As is clear from the graph, the pink and blue areas overlap only a little, suggesting a significant difference between the IV levels. However, when this data is analyzed using $Y_1$ and $Y_2$ individually, as illustrated on the vertical and horizontal axes, ANOVA might not find an effect on these DVs, because the marginalized distributions of the IV levels *do* stronly overlap. By using a linear combination of DVs, MANOVA can find a straight line in this space that maximally separates the blue and pink regions, such that a significant effect can be revealed.

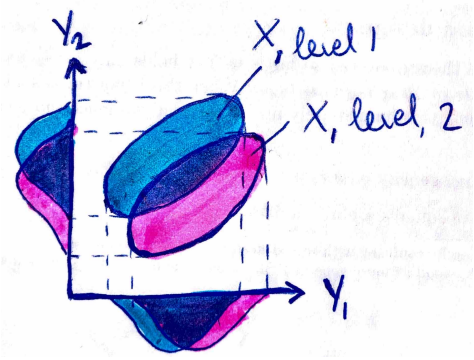



*Figure 4.1: Advantage of MANOVA over ANOVA. A data distribution is shown for two levels of an IV ($X$) using pink and blue, plotting values on two DVs $Y_1$ and $Y_2$. The joint distributions do not overlap a lot but the marginal distributions do, meaning an effect might be found with MANOVA but not with ANOVA.*

Next to this additional advantage, some additional things to pay attention to are listed below.

- It is often true that ANOVA provides *more* power than MANOVA!
- For each cell in the design, one wants more cases (i.e. datapoints) than the number of DVs.

- You want to have large negative correlation between DVs.

Below, we give an outline of the theory underlying (factorial) MANOVA.[1] To do so, we shall start from variables of factorial ANOVA, and simply make some replacements. Mathematically, we replace scalars (one-dimensional objects) with vectors and matrices (two-dimensional objects). We shall outline the case of two DVs for simplicity, but the replacements can be done more generally.

### Values of $Y$

Going from $Y$ to $Y^{(1)}$ and $Y^{(2)}$, we replace:

$$Y_{abi} \to \mathbf{Y}_{abi} = \begin{pmatrix} Y^{(1)}_{abi} \\ Y^{(2)}_{abi} \end{pmatrix}, \text{ e.g. } \quad Y_{111} = 115 \to \mathbf{Y}_{111} = \begin{pmatrix} 115 \\ 108 \end{pmatrix}$$

where $Y^1_{abi}$ is the value for DV1, and $Y^{(2)}_{abi}$ for DV2. In the same way, the *grand mean* now becomes a vector:

$$\bar{Y} \to \bar{\mathbf{Y}} = \begin{pmatrix} \bar{Y}^{(1)} \\ \bar{Y}^{(2)} \end{pmatrix}, \text{ e.g. } \quad \bar{Y} = 89.11 \to \bar{\mathbf{Y}} = \begin{pmatrix} 89.11 \\ 87.22 \end{pmatrix}$$

And likewise for the *group means*:

$$\bar{Y}_{a1} \to \bar{\mathbf{Y}}_{a1} = \begin{pmatrix} \bar{Y}^{(1)}_{a1} \\ \bar{Y}^{(2)}_{a1} \end{pmatrix}, \text{ e.g. } \quad \bar{Y}_{a1} = 95.83 \to \bar{\mathbf{Y}}_{a1} = \begin{pmatrix} 95.83 \\ 96.0 \end{pmatrix}$$

### Design table

Recall the design table we had for two-way factorial ANOVA, where we shall now leave out the subjects for the moment:

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $b_1$ | $s_1$<br>$s_2$ | $s_5$<br>$s_6$ | |
| $b_2$ | $s_3$<br>$s_4$ | $s_7$<br>$s_8$ | |

$=$

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $b_1$ | $Y_{11}$ | $Y_{21}$ | $Y_{31}$ |
| $b_2$ | $Y_{12}$ | $Y_{22}$ | $Y_{32}$ |

Then, the design table for two-way factorial MANOVA becomes:

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $b_1$ | $Y_{11}$ | $Y_{21}$ | $Y_{31}$ |
| $b_2$ | $Y_{12}$ | $Y_{22}$ | $Y_{32}$ |

$\to$

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $b_1$ | $\mathbf{Y}_{11}$ | $\mathbf{Y}_{21}$ | $\mathbf{Y}_{31}$ |
| $b_2$ | $\mathbf{Y}_{12}$ | $\mathbf{Y}_{22}$ | $\mathbf{Y}_{32}$ |

$=$

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $b_1$ | $Y^{(1)}_{11}$<br>$Y^{(2)}_{11}$ | $Y^{(1)}_{21}$<br>$Y^{(2)}_{21}$ | $Y^{(1)}_{31}$<br>$Y^{(2)}_{31}$ |
| $b_2$ | $Y^{(1)}_{12}$<br>$Y^{(2)}_{12}$ | $Y^{(1)}_{22}$<br>$Y^{(2)}_{22}$ | $Y^{(1)}_{32}$<br>$Y^{(2)}_{32}$ |

Here, each *cell* now contain two DVs, where each DV should be understood to have values for all subjects in that cell. I.e. the *number of cells* in a MANOVA design does not increase as the number of DVs increases; this is still uniquely determined by the levels in the IVs!

### Sums of squares

So what does this mean for the sums of squares? First, recall how this was defined for a single DV:

$$\mathrm{SS}_{\mathrm{tot}}(Y) = \sum_{i=1}^{N} (Y_i - \bar{Y})^2.$$

This was closely related to the sample variance, defined as:

$$\mathrm{var}(Y) = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2.$$

[1] Note that the math involved is the same as that of *discriminant analysis*, which might be familiar to those who have studied machine learning. See e.g. [7].

I.e. in other words:

$$\boxed{\mathrm{SS}_{\mathrm{tot}}(Y) = \frac{1}{N-1}\,\mathrm{var}(Y)}$$

Now, with two DVs, we go from variance to the *covariance matrix*, which for two DVs is a two-by-two matrix:[2] :

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{var}(Y^{(1)}) & \mathrm{cov}(Y^{(1)}, Y^{(2)}) \\ \mathrm{cov}(Y^{(2)}, Y^{(1)}) & \mathrm{var}(Y^{(2)}) \end{pmatrix}$$

With several DVs, the role of sum of squares is now taken by *cross-product matrix* $\mathbf{S}$, which is related to the covariance matrix as follows:

$$\boxed{\mathbf{S} = \frac{1}{N-1}\mathbf{\Sigma}}$$

Analogous to the situation with one DV, one gets dedicated matrices for different effects as well as residuals. For instance, in the case of two-way factorial MANOVA, one gets:

$$\mathbf{S}_{tot} = \mathbf{S}_A + \mathbf{S}_B + \mathbf{S}_{AB} + \mathbf{S}_{err}$$

For instance, for an effect $A$ one goes from a single number to a matrix of four values, containing the different variances and covariances associated with the DVs (times $N-1$):

$$\mathrm{SS}_A = 571 \rightarrow \mathbf{S}_A = \begin{pmatrix} 571 & 762 \\ 762 & 1127 \end{pmatrix}$$

Observe how the matrix is symmetric in the diagonal since one has that $\mathrm{cov}(A, B) = \mathrm{cov}(B, A)$.

**F-test and effect size**

In MANOVA one has an *approximate* $F$ instead of an F-test statistic. This is hard to calculate by hand, but is calculated easily using the appropriate computer program. Unlike univariate tests, *the value of $F$ for MANOVA is entangled with the chosen measure of effect size.* Recall that for factorial ANOVA, the effect size was a *ratio of sums of squares*, e.g. $\eta^2 = \frac{\mathrm{SS}_b}{\mathrm{SS}_{\mathrm{tot}}}$. In MANOVA, with cross-product matrices taking the role of sums of squares the effect sizes are now defined using *ratios of determinants of the cross-product matrices*, $|\mathbf{S}| = \det \mathbf{S}$. Most important is *Wilk's lambda*, defined as:

$$\Lambda = \frac{|\mathbf{S}_{\mathrm{error}}|}{|\mathbf{S}_{\mathrm{effect}} + \mathbf{S}_{\mathrm{error}}|}$$

which corresponds to a specific F-value. Note that Wilk's Lambda itself is typically not seen as an effect size, but it allows direct calculation of one through:

$$\boxed{\eta^2 = 1 - \Lambda.}$$

Now, next to Wilk's Lambda there are several different effect sizes, *each with a different approximate $F$-statistic.*

1. Pillai's Trace
2. Hotelling's Trace ($T^2$)
3. Roy's Largest Root (gcr)

When performing MANOVA using a computer, one often obtains a large table with values of each of these (for all effects), along with their corresponding approximate $F$. How should one pick between these? Wilk's Lambda usually has more power than Pillai's Trace, but Pillai's Trace, is more robust against violations of assumptions. Following [2], we suggest using Wilk's Lambda unless one has a good reason to use Pillai's Trace.

### 4.1.1 An example

Refer to slides.

[2] For any number of DVs $Y^{(i)}$ with $i = 1, 2, ..., M$, it is defined by $\Sigma_{ij} = \mathrm{cov}(Y^{(i)}, Y^{(j)})$

### 4.1.2 MANOVA and earlier techniques

Combining MANOVA with earlier techniques is straightforward. We just saw how *factorial MANOVA* works. To get *one-way* MANOVA, one simply leaves out one IV. To get RM/RBD designs with several DVs, one simply does the analysis of Ch. 2 with several DVs – similarly for ANCOVA, in which case one (intuitively) ends up with *multivariate analysis of covariance*, or *MANCOVA*. One interesting use case is when one has a *RM ANOVA* design where *sphericity is violated*; then, one can instead use *MANOVA with the within-subject variable taken as different DVs*.

### 4.1.3 Assumptions

We list assumptions for MAN(C)OVA below.

1. Multivariate normality: sampling distributions of means of various DVs, and all linear combinations of them are normally distributed.
2. No outliers: MANOVA is particularly sensitive to outliers.
3. Homogeneity of covariance matrix: the generalization of the homogeneity of variance assumption. This can be tested using *Box's M test*.
4. Linearity: between all pairs of DVs (and with MANCOVA, between pairs of CVs, and DV-CV pairs).
5. No collinearity between DVs: DVs should not be too correlated/redundant.
6. (In MANCOVA): homogeneity of regression (regression between CVs and DVs should be the same for all groups).
7. (In MANCOVA): reliable CVs.

## 4.2 The General Linear Model (GLM)

Let us recap the methods that we have discussed so far. We have focused on two types of research questions:

1. **Differences between group means**: *ANOVA-type* techniques (including t-tests).
   - Violated assumptions: *non-parametric tests*.
   - >1 IV: *factorial ANOVA* (*between-subject* & *within-subject*, a.k.a. *repeated measures*).
   - CVs: *randomized blocks design* (discrete CV), *ANCOVA* (continuous CV).
   - >1 DV: *MANOVA*
2. **Prediction**: *regression-type* techniques (including logistic regression).
   - Continuous DV: *linear regression*.
   - Binary DV: *logistic regression*.

As was hinted at in Ch. 3, all these methods *share an underlying mathematical core*. They can in fact be seen as specific cases of the so-called *general liner model* (GLM). To understand this, we review in detail how regression and ANOVA can be seen as the same *linear model*. This will provide an intuition for the GLM, which is discussed afterwards.

### 4.2.1 From regression to ANOVA: linear models

We shall start from regression with one predictor, and show how this can be rewritten to obtain one-way ANOVA. Then, we shall do the same for two predictors/two-way ANOVA.

#### One IV

To see where we have to end up, recall the hypotheses for one-way ANOVA:

$$\begin{aligned} H_0: &\quad Y = \mu + \epsilon \\ H_1: &\quad Y = \mu + \alpha + \epsilon \end{aligned} \tag{4.1}$$

where we have left out the subscripts for simplicity. Recall that $\alpha$ was the *effect* associated with the IV. We have made the grand mean blue and the effect red to aid the ensuing discussion. Observe how we can write these hypothesis in a type of design table (including DV values instead of subjects) as follows: Now we return to regression. Recall that the equation for regression with one predictor $X_1$

*Table 4.2: One-way ANOVA with two levels.*

| $a_1$ | $a_2$ |
|---|---|
| $\mu$ | $\mu + \alpha$ |

was:

$$Y = b_0 + b_1 X_1 + \epsilon \tag{4.2}$$

where, compared to before, we have replaced $\hat{Y}$ by $Y$ on the left hand side, compensated by including the residual $\epsilon = Y - \hat{Y}$ on the right hand side.[3] Now to go from regression to ANOVA, we ought to make $X_1$ *discrete*. To do so, we set it to be either 0 or 1 (like we did in *logistic* regression). Defined in this way, $X$ is also called a *binary contrast*. Then, we obtain for $X_1 = 0$:

$$Y = b_0 + \epsilon,$$

and for $X_1 = 1$:

$$Y = b_0 + b_1 + \epsilon.$$

[3] We have included a minus sign compared to Section 2, which is allowed since $\epsilon$ is a random variable with mean zero.

Recall now that in regression, we could do hypothesis testing of individual coefficients $b_i$ (Section 3.1.3), where the hypotheses were:

$$H_0: \quad b_1 = 0$$
$$H_1: \quad b_1 \neq 0$$

Looking now at Eq. 4.1, we see that we can *map* exactly the symbols of regression to those of ANOVA. We associate $X_1 = 0$ with level $a_1$, and $X_1 = 1$ with level $a_2$, and identify $\mu = b_0$ and $\alpha = b_1$. Then, regression with binary contrasts gives exactly(!) the same hypotheses as one-way ANOVA:

$$H_0: \quad Y = b_0 + \epsilon$$
$$H_1: \quad Y = b_0 + b_1 + \epsilon.$$

I.e. we should understand the *intercept/slope* in regression as the *grand mean/main effect* in ANOVA, respectively! Putting this in a table as above gives:

*Table 4.3: Regression with a single binary contrast.*

| $X_1 = 0$ | $X_1 = 1$ |
|---|---|
| $b_0$ | $b_0 + b_1$ |

Putting this side by side with the table we had before we see clearly how regression with binary contrasts is equivalent to one-way ANOVA with two levels.

*One-way ANOVA with two levels.*

| $a_1$ | $a_2$ |
|---|---|
| $\mu$ | $\mu + \alpha$ |

=

*Regression with a single binary contrast.*

| $X_1 = 0$ | $X_1 = 1$ |
|---|---|
| $b_0$ | $b_0 + b_1$ |

**Two IVs**

We now do the same thing for regression with *two predictors*, adding $X_2$ to the regression equation:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \epsilon.$$

Setting now both $X_1$ and $X_2$ to binary contrasts gives the table below.

*Table 4.4: Regression with two binary contrasts.*

| | $X_1 = 0$ | $X_1 = 1$ |
|---|---|---|
| $X_2 = 0$ | $b_0$ | $b_0 + b_1$ |
| $X_2 = 1$ | $b_0 + b_2$ | $b_0 + b_1 + b_2$ |

Now, on the ANOVA side, consider a (crossed) 2x2 ANOVA with variables $A$ and $B$, and associated effects $\alpha$ and $\beta$ (we leave out the interaction $(\alpha\beta)$):

$$\begin{aligned} H_0: \quad & Y = \mu + \epsilon \\ H_1: \quad & Y = \mu + \alpha + \epsilon, && \text{(main effect A), or} \\ & Y = \mu + \beta + \epsilon, && \text{(main effect B), or} \\ & Y = \mu + \alpha + \beta + \epsilon, && \text{(main effect of both).} \end{aligned}$$

Which in our table gives:

*Table 4.5: (2x2) ANOVA.*

| | $a_1$ | $a_2$ |
|---|---|---|
| $b_1$ | $\mu$ | $\mu + \alpha$ |
| $b_2$ | $\mu + \beta$ | $\mu + \alpha + \beta$ |

Thus, we see that if we additionally identify $\beta = b_2$, we get precisely (2x2) crossed ANOVA (without interactions)!

*(2x2) ANOVA.*

| | $a_1$ | $a_2$ |
|---|---|---|
| $b_1$ | $\mu$ | $\mu + \alpha$ |
| $b_2$ | $\mu + \beta$ | $\mu + \alpha + \beta$ |

=

*Regression with two binary contrasts.*

| | $X_1 = 0$ | $X_1 = 1$ |
|---|---|---|
| $X_2 = 0$ | $b_0$ | $b_0 + b_1$ |
| $X_2 = 1$ | $b_0 + b_2$ | $b_0 + b_1 + b_2$ |

**In sum**

To sum up, we have observed that:

> Doing t-tests for regression coefficients ($b_1$, $b_2$) with binary contrasts is equivalent to doing F-test for main effects ($\alpha$, $\beta$) in a crossed one-way ANOVA.

Thus, these examples show that regression and ANOVA are less different than they might appear at first sight. This relation between regression and ANOVA keeps being valid when more IVs are included, which can be done by simplya adding more contrasts. Adding more than two levels for IVs is also possible, but requires a somewhat more elaborate scheme for contrasts.

The key underlying feature of these models is thus their *linearity*, as stated by Eq. 4.1 for one-way ANOVA, and Eq. 4.2 for regression with a single predictor: they are both a *linear model*. Our next step will be to include *all* methods we have discussed (and many more which we have not discussed) in a *single model*.

### 4.2.2 The GLM

This is the *general linear model* (GLM). This is the set of models defined by:

$$\boxed{\mathbf{Y} = \mathbf{XB} + \mathbf{U}} \tag{4.3}$$

where $\mathbf{Y}$ represents the DV(s), and $\mathbf{X}$ the IV(s) and $\mathbf{U}$ is represents the error. We can distinguish three main cases of writing this equation, based on the nature of $\mathbf{X}$ and $\mathbf{Y}$, which we discuss below. Then, depending on whether the *type* of the variables continuous, discrete (with binary/dichotomous as a special case), or mixed, we get all the techniques discussed in previous chapters.

#### One IV, one DV

Here, $\mathbf{X}$ and $\mathbf{Y}$ are one-dimensional, i.e.:

$$Y = b_1 X_1 + \epsilon.$$

This is also called *bivariate form*. Techniques we have discussed:

- Pearson's correlation (= linear regression with one predictor) ($X$ and $Y$ continuous)
- $\chi^2$ independence ($X$ and $Y$ binary)

#### Several IVs, one DV

Here, $\mathbf{Y}$ is one-dimensional, and $\mathbf{X}$ is K-dimensional. This is also called *simple multivariate form*. It can be written as:

$$Y = \sum_k b_k X_k + \epsilon$$

Techniques we have discussed:

- *Factorial ANOVA* ($X$s discrete, $Y$ continuous)
- *Multiple regression* ($X$s continuous, $Y$ continuous)

- *ANCOVA* ($X$s discrete/continuous, $Y$ continous)
- *Two-group logistic regression* ($X$s continuous, $Y$ binary)

Some techniques we have not discussed:

- *Two-group discriminant analysis* ($X$s continuous, $Y$ binary)
- *Multilevel modelling* ($X$s at each discrete/continuous. $Y$ at each level is continuous.)

**Several IVs, several DVs**

Here, both $\mathbf{Y}$ and $\mathbf{X}$ can have any dimension, i.e. one has an arbitrary number of IVs and DVs. The equation is Eq. 4.3, which we can also write as:

$$\begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1q} \\ 1 & X_{21} & X_{22} & \cdots & X_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nq} \end{bmatrix} \begin{bmatrix} b_{01} & b_{02} & \cdots & b_{0p} \\ b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{q1} & b_{n2} & \cdots & b_{qp} \end{bmatrix} + \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \cdots & \epsilon_{n1} \\ \epsilon_{21} & \epsilon_{22} & \cdots & \epsilon_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n1} & \epsilon_{n1} & \cdots & \epsilon_{np} \end{bmatrix},$$

which is also called *full multivariate form*. Techniques we have discussed:

- *MANOVA* ($X$s discrete, $Y$s continuous);
- *MANCOVA* ($X$s discrete/continuous, $Y$s continuous).

Some techniques we have not discussed:

- *Canonical correlation* ($X$s continuous, $Y$s continuous);
- *Discriminant analysis* ($X$s continuous, $Y$s discrete);
- *Factor analysis/principal component analysis* ($Y$s continuous, $X$s are *latent variables*);
- *Structural equations modelling* ($X$s continuous and/or latent, $Y$s continuous and/or latent).

As such, the GLM provides an elegant underlying framework for many of the topics discussed so far. For more on these techniques, we refer to [2].

# Bibliography


[1] Danielle Navarro. Learning Statistics with R, 2016.

[2] Barbara G. Tabachnick and Linda S. Fidell. *Using Multivariate Statistics*. Pearson Education, 2013. Google-Books-ID: ucj1ygAACAAJ.

[3] NIST/SEMATECH e-Handbook of Statistical Methods.

[4] George Casella and Roger L. Berger. *Statistical Inference*. Cengage Learning India, Andover Melbourne Mexico City Stamford, CT Toronto Hong Kong New Delhi Seoul Singapore Tokyo, December 2007.

[5] Paul Cairns. *Doing Better Statistics in Human-Computer Interaction*. Cambridge University Press, Cambridge, 2019.

[6] Alan Dix. *Statistics for HCI: Making Sense of Quantitative Data*. MORGAN & CLAYPOOL, San Rafael, California, April 2020.

[7] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006th edition edition, August 2006.

[8] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. New York Heidelberg Dordrecht London, 2013.