R wie Rotz

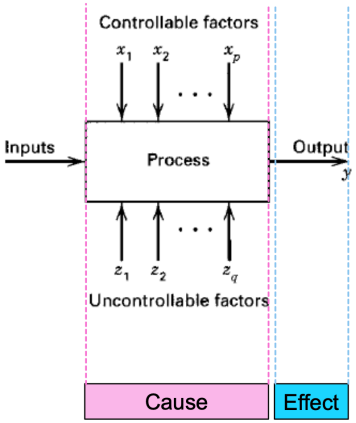
by dcamenisch

1 Introduction

This document is a summary of the 2022 edition of the lecture *Applied Analysis of Variance and Experimental Design* at ETH Zurich. I do not guarantee correctness or completeness, nor is this document endorsed by the lecturers. If you spot any mistakes or find other improvements, feel free to open a pull request on `https://github.com/DannyCamenisch/anova-summary`. This work is published as CC BY-NC-SA.

2 Learning from Data

We are in the abstract situation where we have a "system" or a "process" with many input variables (**predictors**) and an output (**response**). We want to find **cause-effect relationships**, meaning that when we actively change one of the inputs (intervention), this will cause the output to change. This is what we do in **experimental studies**. If we can just observe a system under different settings (observational studies), it is much harder to make a statement about causal effects. With observational data, we can typically just make a statement about an association between two variables. One potential danger is the existence of **confounders** (a common cause for two variables).



2.1 Experimental Studies

Before designing an experimental study, we must have a precise research question that is actually testable, i.e., that we can do the appropriate interventions and that we can measure the right response.

An experimental study consists of:

- **Treatments / Predictors**: the different interventions on the system
- **Experimental units**: the actual objects on which we apply the treatments
- A method that assigns experimental units to treatments, typically **randomization**
- **Response(s)**: the output that we measure

2.1.1 Treatments or Predictors

We distinguish between the following types of predictors:

- Predictors that are of primary interest and that can (ideally) be varied according to our wishes
- Predictors that are systematically recorded such that potential effects can later be eliminated in our calculations (**covariates**)
- Predictors that can be kept constant and whose effects are therefore eliminated
- Predictors that we can neither record nor keep constant

2.1.2 Randomization

Randomization ensures that the only systematic difference between the groups is the treatment. This protects us from confounders and is the reason why a properly randomized experiment allows us to make a statement about a cause-effect relationship between treatment and response. Typically, we do a randomization within homogeneous blocks. This restricted version of randomization is called **blocking**. A block is a subset of experimental units that is more homogeneous than the entire set.

2.1.3 Experimental and Measurement Units

An **experimental unit** is defined as the object on which we apply the treatments by randomization. On the other hand, a **measurement unit** is the object on which the response is being measured. They do not have to be the same.

2.1.4 Experimental Error

Different experimental units will give different responses to the same treatment (**experimental error**). Therefore we need multiple replicates receiving the same treatment. If the difference between the treatments is much larger than the experimental error, we can conclude that there is a treatment effect.

2.1.5 Blinding

Blinding means that those who measure the response do not know which treatment is given. With humans it is common to use **double-blinding** where in addition the patients do not know the assignment either. Blinding protects us from (unintentional) bias due to expectations.
A **control treatment** is typically a standard treatment with which we want to compare. It can also be no treatment at all.

3 Completely Randomized Design

We assume for the moment that the experimental units are homogeneous. We know how to compare two independent groups using the two-sample t-test. If we have more than two groups, this is not applicable anymore.

3.1 One-Way Analysis of Variance

On an abstract level we want to compare $g \geq 2$ treatments, having N experimental units, that we assign randomly to the different treatment groups having n_i observations each. This is what we call **completely randomized design**, it is the most elementary experimental design. If all the treatment groups have the same number of experimental units, we call the design **balanced**. Such random assignments can be done as follows:

```
sample(treat.ord) ## Random Permutation of treat.ord
```

3.1.1 Cell Means Model

Let y_{ij} be the observed response from the j-th experimental unit in treatment group i. In the **cells mean model** we allow each treatment group (cell) to have its own expected value. This means that y_{ij} is the realised value of the random variable:

$$Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2), \quad \text{or} \quad Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

As for the standard two-sample t-test, the variance is assumed to be equal for all groups. We say that Y is the response and the treatment allocation is a categorical predictor. A categorical predictor is also called a factor. We sometimes distinguish between unordered (or nominal) and ordered (or ordinal) factors. We can rewrite the equation as:

$$\mu_i = \mu + \alpha_i$$

Where α_i is called the **treatment effect**. This will later help us to untangle the influence of multiple treatment factors on the response. Through this rewrite we have introduced an additional parameter, to remove it again we need a side constraint. Possible constraints could be:

- weighted sum-to-zero: $\sum_{i=1}^{g} n_i \alpha_i = 0$
- sum-to-zero: $\sum_{i=0}^{g} a_i = 0$
- reference group: $\alpha_1 = 0$

For all of the choices it holds that μ determines some sort of "global level" of the data and α_i contains information about differences between the group means μ_i from that "global level". If we know $g-1$ of the α_i, we automatically know the remaining α_i, we also say that the treatment effect has $g-1$ **degrees of freedom** (df).

```
## Options takes two args, the first for unordered
## and the second for ordered factors.
## contr.poly (weighted sum-to-zero) DEFAULT
## contr.sum (sum-to-zero)
## contr.treatment (reference group)
options(contrasts = c("contr.sum", "contr.poly"))
```

3.1.2 Parameter Estimation

We estimate the parameters using the least squares criterion:

$$\hat{\mu}, \hat{\alpha}_i = \underset{\mu, \alpha_i}{\operatorname{argmin}} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)^2$$

Some notation:

$$y_{i.} = \sum_{j=1}^{n_i} y_{ij} \qquad \bar{y}_{i.} = \frac{1}{n_i} y_{i.}$$

$$y_{..} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij} \qquad \bar{y}_{i.} = \frac{1}{N} y_{..}$$

As we can independently estimate the values of μ_i, one can show that $\hat{\mu}_i = \bar{y}_{i.}$. From $\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}$ we can get all the other parameters needed (they still depend on the side constraint).

The estimate of the error variance is also called **mean squared error** MS_E:

$$\hat{\sigma}^2 = MS_E = \frac{1}{N-g} SS_E$$

Where SS_E is the **error** or **residual sum of square**:

$$SS_E = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2$$

We can fit the model as follows:

```
fit <- aov(dresponse ~ dfactor, data = d)
## Have a look at the estimated coefficients
coef(fit) ## or dummy.coef(fit)
```

3.1.3 Tests

With the two-sample t-test, we could test whether two samples share the same mean. We will now extend this for $g > 2$. Saying that all groups share the same mean is equivalent to saying:

$$Y_{ij} = \mu + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

This is the so-called **single mean model**, a special case of the cell means model. We have the global null hypothesis

$$H_0 : \mu_1 = ... = \mu_g$$

vs. the alternative hypothesis

$$H_A : \mu_k \neq \mu_l \text{ for at least one pair } k \neq l$$

The idea is to check whether the variation between the different treatment groups (the "signal") is larger than the variation within the groups (the "noise"). We can decompose the total variation as follows:

$$\underbrace{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (\bar{y}_{ij} - \bar{y}_{..})^2}_{SS_T} = \underbrace{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2}_{SS_{Trt}} + \underbrace{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2}_{SS_E}$$

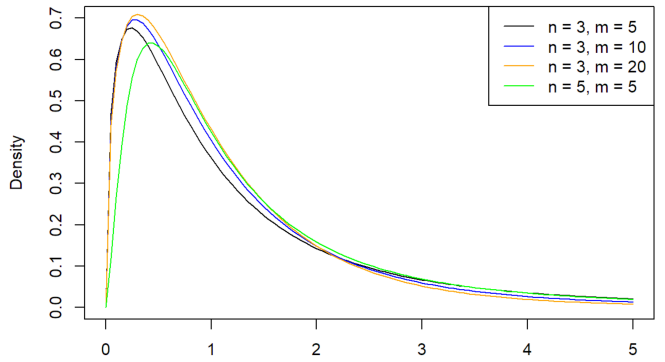
This information can be summarized in a **ANOVA** table.

Source	df	Sum of Squares	Mean Squares	F-ratio
Treatment	$g-1$	SS_{Trt}	$MS_{\text{Trt}} = \frac{SS_{\text{Trt}}}{g-1}$	$\frac{MS_{\text{Trt}}}{MS_E}$
Error	$N-g$	SS_E	$MS_E = \frac{SS_E}{N-g}$	

The MS and SS are normalized with the corresponding degrees of freedom. This is a so-called one-way ANOVA, because there is only one factor involved. If all groups share the same expected value, the treatment sum of squares is typically small. We introduce the so called F-ratio.

$$F\text{-ratio} = \frac{MS_{Trt}}{MS_E} \sim F_{g-1, N-g}$$

If the variation between groups is substantially larger than the variation within groups (higher F-ratio), we have evidence against H_0. The F-distribution looks as follows:



As with any other statistical test, we reject H_0 if the observed value of the F-ratio, our test statistics, lies in an "extreme" region of the corresponding F-distribution:

$$F\text{-ratio} > F_{g-1, N-g, 1-\alpha}$$

As this test is based on the F-ratio we call it an F**-test**. In R, we can use the following function to get the ANOVA table and the p-value of the F-test.

```
summary(fit)
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        2   3.77   1.883    4.85     0.016
## Residuals   27  10.49   0.389
```

3.2 Checking Model Assumptions

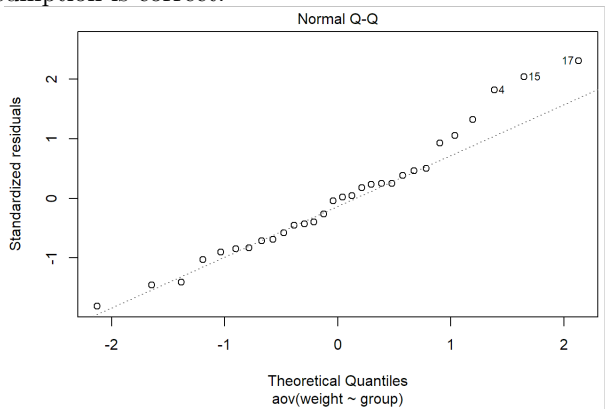
Statistical inference is only valid if all model assumptions are fulfilled. So far this means:

- The errors are independent
- The errors are normally distributed
- The error variance is constant
- The errors have mean zero

We now introduce different plots to check these assumptions. This means that we use graphical tools to perform qualitative checks.

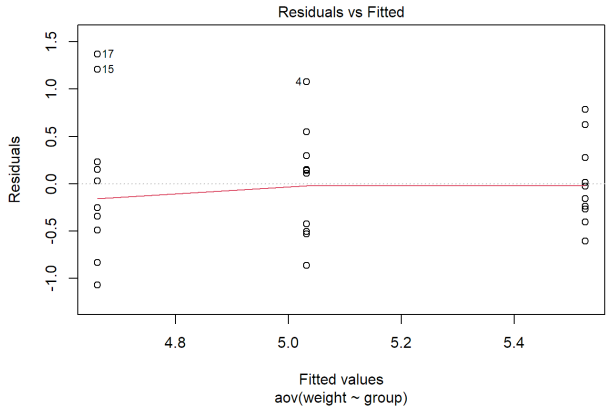
3.2.1 QQ-Plot

In a QQ-plot we plot the empirical quantiles of the residuals or "what we see in the data" vs. the theoretical quantiles or "what we expect from the model". The plot should show a more or less straight line if the normality assumption is correct.



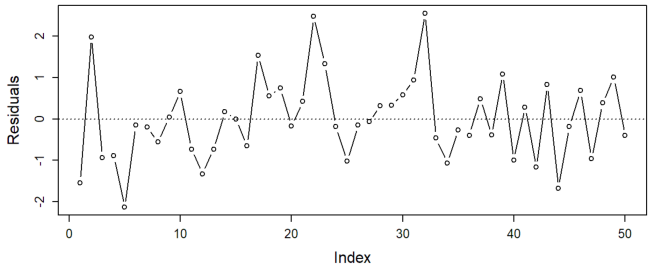
3.2.2 Tukey-Anscombe Plot

The Tukey-Anscombe plot (TA-plot) plots the residuals r_{ij} vs. the fitted values $\hat{\mu}_i$ (estimated cell means). It allows us to check whether the residuals have constant variance.



3.2.3 Index Plot

If the data has some serial structure, i.e., if observations were recorded in a certain time order, we typically want to check whether residuals close in time are more similar than residuals far apart. For this we use the index plot where we plot the residuals against time. For positively dependent residuals, we would see time periods where most residuals have the same sign, while for negatively dependent residuals, the residuals would jump too often from positive to negative compared to independent residuals.



3.3 Power or "What Sample Size Do I Need?"

By construction, a statistical test controls the so-called type I error rate with the significance level α. This means that the probability that we falsely reject the null hypothesis is less than or equal to α. Besides the type I error, there is also a type II error. It occurs if we fail to reject the null hypothesis even though the alternative hypothesis holds. The probability of a type II error is typically denoted by β.

The **power** of a statistical test is defined as $P(\text{reject } H_0 \text{ given that a certain setting under } H_A \text{ holds}) = 1 - \beta$. Intuitively, it seems clear that the "further away" we choose the parameter setting from H_0 the larger will be the power, or the smaller will be the probability of a type II error.

3.3.1 Calculating Power for a Certain Design

Why should we be interested in such an abstract concept when planning an experiment? Power can be thought of as the probability of success, i.e. getting a significant result. The question of "what sample size do I need?" depends on the question of "what power you like". The power depends on:

- design of the experiment
- significance level
- parameter setting under the alternative
- sample size

We mainly use sample size and experimental design to maximize the power. We are not gonna look at how exactly the calculations work, rather we choose an alternative way. What we can always do is to simulate a lot of data sets under the alternative that we believe in and check how often we are rejecting the corresponding null hypothesis. The empirical rejection rate is then an estimate of the power. A nice side effect of doing a power analysis is that you actually do the whole data analysis on simulated data and you immediately see whether this works as intended. From a conceptual point of view, we can use such a simulation-based procedure for any design. However, the number of parameters grows rather quickly with increasing model complexity.

In that sense, the results of a power analysis are typically not very precise. However, they should still give us a rough idea about the required sample size in the sense of whether we need 6 or 60 observations per group.

4 Contrast and Multiple Testing

4.1 Contrast

The F-test is rather unspecific and gives us basically a yes/no answer. Often we have a more specific question than the global null hypothesis we want to answer. Such kind of questions can be formulated as so-called **contrasts**. As hypothesis we choose:

$$H_0 : \sum_{i=1}^{g} c_i \mu_i = 0 \text{ and } H_A : \sum_{i=1}^{g} c_i \mu_i \neq 0$$

Typically we have the side constraint that $\sum_{i=1}^{g} c_i = 0$. The contrast is about the differences between treatments and not about the overall response.
We estimate the value of $\sum_{i=1}^{g} c_i \mu_i$ with:

$$\sum_{i=1}^{g} c_i \hat{\mu}_i = \sum_{i=1}^{g} c_i \bar{y}_{i\cdot}$$

In addition, we could derive its accuracy (standard error), construct confidence intervals and do tests.

(3.1.2 Some Technical Details are left out on purpose)

4.2 Multiple Testing

The problem with all statistical tests is the fact that the overall type I error rate increases with increasing number of tests. This means that if we perform many tests, we expect to find some significant results, even if all null hypotheses are true. Somehow we have to take into account the number of tests that we perform to control the overall type I error rate.

We list the potential outcomes of a total of m tests, among which m_0 null hypotheses are true:

	H_0 true	H_0 false	Total
Significant	V	S	R
Not significant	U	T	W
Total	m_0	$m - m_0$	m

For example, V is the number of wrongly rejected null hypotheses (type I errors, also known as FP). Using this notation, the overall or family-wise error rate (FWER) is defined as the probability of rejecting at least one of the true H_0's:

$$\text{FWER} = P(V \geq 1)$$

The family-wise error rate is very strict in the sense that we are just interested in whether there is at least one wrong rejection. We say that a procedure controls the family-wise error rate in the strong sense at level α if FWER $\leq \alpha$ for any configuration of true and non-true null hypotheses.

Another error rate is the FDR which is the expected fraction of false discoveries:

$$\text{FDR} = E\left[\frac{V}{R}\right]$$

Controlling FDR at level 0.2 means that on average in our list of significant findings only 20% are false positives. If a procedure controls FWER at level α, FDR is automatically controlled at level α too. This does not hold the other way around.

We can also control the error rates for confidence intervals. We call a set of confidence intervals simultaneous confidence intervals at level $(1-\alpha)$ if the probability that all intervals cover the corresponding true parameter value is $(1-\alpha)$. This means that we can look at all confidence intervals at the same time and get the correct big picture with probability $(1-\alpha)$.
In the following, we typically start with individual p-values (the ordinary p-values corresponding to the $H_{0,j}$ and modify them such that the appropriate overall error rate (like FWER) is being controlled. The modified p-values should be interpreted as the smallest overall error rate such that we can reject the corresponding null hypothesis.

4.2.1 Bonferroni

The Bonferroni correction is a very generic but conservative approach. The idea is to use a more restrictive (individual) significance level of $\alpha^* = \alpha/m$. This procedure controls the FWER in the strong sense for any dependency structure of the different tests. Especially for large m, the Bonferroni correction is very conservative leading to low power.

4.2.2 Bonferroni-Holm

The Bonferroni-Holm procedure also controls the FWER in the strong sense. It is less conservative and uniformly more powerful, which means always better, than Bonferroni. It works in the following sequential way:

1. Sort p-values from small to large
2. For $j = 1, ...$: Reject null hypothesis if $p_j \leq \frac{\alpha}{m-j+1}$
3. Stop when reaching the first non-significant p-value and do not reject the remaining null hypotheses.

Note that this procedure only works with p-values but cannot be used to construct confidence intervals.

4.2.3 Scheffe

The Scheffe procedure controls for the search over any possible contrast. This means we can try out as many contrasts as we like and still get honest p-values! This is even true for contrasts that are suggested by the data, which were not planned beforehand, but only after seeing some special structure in the data. The price for this is low power.

The Scheffe procedure works as follows: We start with the sum of squares of the contrast SS_C. Then we build the F-ratio:

$$\frac{SS_C/(g-1)}{MS_E}$$

4.2.4 Tukey Honest Significant Differences

A special case of a multiple testing problem is the comparison between all possible pairs of treatments. The output is a matrix of p-values of the corresponding comparisons. We could now use the Bonferroni correction method. However, there exists a better, more powerful alternative which is called Tukey Honest Significant Differences (HSD).

Think of a procedure that is custom tailored for the situation where we want to do a comparison between all possible pairs of treatments. We get both p-values (which are adjusted such that the family-wise error rate is being controlled) and simultaneous confidence intervals.

4.2.5 Multiple Comparisons with a Control

if we want to compare all treatment groups with a control group, we have a so-called multiple comparisons with a control (MCC) problem. The corresponding custom-tailored procedure is called Dunnett procedure. It controls

the family-wise error rate in the strong sense and produces simultaneous confidence intervals.

We get smaller p-values than with the Tukey HSD procedure because we have to correct for less tests; there are more comparisons between pairs than there are comparisons to the control treatment.

5 Factorial Treatment Structure

Often treatments are combinations of the levels of two or more factors, this is called **factorial treatment structure**. If we observe all possible combinations, we call them **crossed**. This typically leads to questions about the interaction of the different factors (or if the interact at all).

5.1 Two-Way ANOVA Model

We assume a setup with a factor A with a levels, a factor B with b levels and n replicates for every combination (a **balanced** design). We denote by y_{ijk} the kth observation of the response of the treatment formed by the ith level of factor A and the jth level of factor B. Instead of setting up a model for each combination, we incorporate the factorial treatment structure directly into the **two-way ANOVA model with interaction**:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij}$$

Hereby α, β are the main effect of factor A, B and $(\alpha\beta)$ is the interaction effect. A model without interaction term is additive, meaning that the effect of A does not depend on the effect of B.

As usual, we'll have to use side constraints for the parameters (we will use the sum-to-zero constraint). For the main effects:

$$\sum_{i=1}^{a} \alpha_i = 0 \qquad \sum_{j=1}^{b} \beta_j = 0$$

Hence they both have $a-1$ / $b-1$ degrees of freedom. For the interaction effect we need to make sure that it contains nothing which is specific to one factor:

$$\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0 \qquad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0$$

Therefore the interaction term has a degree of freedom of $(a-1)(b-1)$.

5.1.1 Parameter Estimation

As usual, we estimate parameters using the principles of least squares and using sum-to-zero side constraints. We get the following parameter estimates:

$$\begin{aligned}
\hat{\mu} &= \bar{y}_{...} \\
\hat{\alpha}_i &= \bar{y}_{i..} - \bar{y}_{...} \\
\hat{\beta}_j &= \bar{y}_{.j.} - \bar{y}_{...} \\
\widehat{(\alpha\beta)}_{ij} &= \bar{y}_{ij.} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j
\end{aligned}$$

We end up with the mean of the observations in the corresponding cell as the expected value of the response Y_{ijk}.

5.1.2 Tests

As in the case of the one-way ANOVA, the total sum of squares SS_T can be partitioned into different sources.

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

Where the individual terms are given by:

Source	Sum of Squares
A ("between rows")	$SS_A = \sum_{i=1}^{a} bn(\widehat{\alpha}_i)^2$
B ("between columns")	$SS_B = \sum_{j=1}^{b} an(\widehat{\beta}_j)^2$
AB ("correction")	$SS_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b} n(\widehat{\alpha\beta})_{ij}^2$
Error ("within cells")	$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2$
Total	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{...})^2$

We can again construct an ANOVA table:

Source	df	SS	Mean Squares	F-ratio
A	$a-1$	SS_A	$MS_A = \frac{SS_A}{a-1}$	$\frac{MS_A}{MS_E}$
B	$b-1$	SS_B	$MS_B = \frac{SS_B}{b-1}$	$\frac{MS_B}{MS_E}$
AB	$(a-1)(b-1)$	SS_{AB}	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$\frac{MS_{AB}}{MS_E}$
Error	$ab(n-1)$	SS_E	$MS_E = \frac{SS_E}{ab(n-1)}$	

We now want to construct global tests for the main effects and the interaction effect:

Interaction Effect: The null hypothesis that there is no interaction effect can be seen as: "The effect of factor A does not depend on the level of factor B or the other way around". $H_0 : \forall ij.\ (\alpha\beta)_{ij} = 0$. Under H_0 it holds that:

$$\frac{MS_{AB}}{MS_E} \sim F_{(a-1)(b-1),ab(n-1)}$$

Main Effect of A: $H_0 : \forall i.\ \alpha_i = 0$. Under H_0 it holds that:

$$\frac{MS_A}{MS_E} \sim F_{(a-1),ab(n-1)}$$

Main Effect of B: $H_0 : \forall j.\ \beta_j = 0$. Under H_0 it holds that:

$$\frac{MS_B}{MS_E} \sim F_{(b-1),ab(n-1)}$$

We first check whether we need the interaction term or not. If there is no evidence of interaction, we continue with the inspection of the main effects.

5.1.3 Single Observations per Cell

If we only have a single observation in each "cell", we cannot do statistical inference anymore with a model including the interaction. The reason is that we have no idea of the experimental error. However, we can still fit a main effects only model. If the data generating mechanism actually contains an interaction, we are fitting a wrong model. The consequence is that the estimate of the error variance will be biased (upward). Hence, the corresponding tests will be too conservative, meaning p-values will be too large and confidence intervals too wide. This is not a problem as the type I error rate is still controlled; we just lose power.

5.1.4 Checking Model Assumptions

As before, we use the QQ-plot and the Tukey-Anscombe plot to check the model assumptions.

5.1.5 Unbalanced Data

We started with the very strong assumption that our data is balanced, i.e., we have the same number of replicates. This assumption made our life "easy" in the sense that we could uniquely decompose total variability into different sources and we could estimate the parameters of the coefficients of a factor by ignoring the other factors. In practice, data is typically not balanced.

We use the following notation: $SS(B|1, A)$ denotes the **reduction in residual sum of squares** when comparing the model $(1, A, B) = y \sim A + B$ with $(1, A) = y \sim A$. The 1 denotes the overall mean μ. Interpretation of the corresponding test is as follows: "Do we need factor B in the model if we already have factor A, or after having controlled for factor A?".

There are three different ways of model comparison approaches:

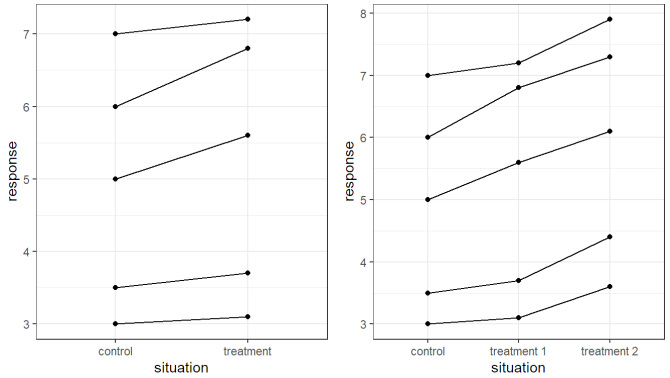
- Type 1 (sequential): $SS(A|1) \rightarrow SS(B|1, A) \rightarrow SS(AB|1, A, B)$
- Type 2 (hierarchical): $SS(A|1, B) \rightarrow SS(B|1, A) \rightarrow SS(AB|1, A, B)$
- Type 3 (fully adjusted): $SS(A|1, B, AB) \rightarrow SS(B|1, A, AB) \rightarrow SS(AB|1, A, B)$

Type 1 is what we will typically get with *summary* in R. Hence we get different results whether we write $y \sim A * B$ or $y \sim B * A$. For type 2 we can either use the function *Anova* in the package *car* or we could compare the appropriate models with the function *anova* ourselves. For type 3 we can use the command *drop1*; we have to be careful that we set the contrast option to *contr.sum* in this special situation for technical reasons, see also the warning in the help file of the function *Anova* of package *car*.

Typically, we take MS_E from the full model (including all terms) as the estimate for the error variance to construct the corresponding F-tests.

6 Complete Block Designs

In many situations we know that our experimental units are not homogeneous. Making explicit use of the special structure of the experimental units typically helps reduce variance. We apply the treatments to the same object / subject. This makes the subject-to-subject variability completely disappear. We also say that we block on subjects or that an individual subject is a block.



6.1 Randomized Complete Block Designs

Assume that we can divide our experimental units into r groups, also known as blocks, containing g experimental units each. The **randomized complete block design** (RCBD) uses a restricted randomization scheme: Within every block, the g treatments are randomized to the g experimental units. The design is called complete because we observe the complete set of treatments within every block. Note that blocking is a special way to design an experiment, or a special "flavor" of randomization. It is not something that you use only when analyzing the data.

The experimental units should be as similar as possible within the same block, but can be very different between different blocks. This design allows us to fully remove the between-block variability from the response because it can be explained by the block factor. In that sense, blocking is a so-called variance reduction technique. The randomization step within each block makes sure that we are protected from unknown confounding variables. Typcial block factors are location, day, machine operator, subjects, etc.

In the most basic form, we assume that we do not have replicates within a block. This means that we only observe every treatment once in each block. The analysis of a randomized complete block design is straightforward. We treat the block factor as "just another" factor in our model. As we have no replicates within blocks, we can only fit a main effects model of the form:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

According to this model, we implicitly assume that blocks only cause additive shifts. Or in other words, the treatment effects are always the same, no matter what block we consider. We would like the block factor to explain a lot of variation, hence if the mean square of the block factor is larger than the error mean square MS_E we conclude that blocking was efficient

7 Random and Mixed Effects Models

Up to now, treatment effects (α_i) where fixed, unknown quantities that we tried to estimate. This means we are making a statement about a specific, fixed set of treatments. Such models are also called fixed effects models.

7.1 Random Effects Model

7.1.1 One-Way ANOVA

We now consider situations where treatments are random samples from a large population of treatments. We are interested in making a statement about some properties of the whole population and not of the observed individuals. We can model such data with the model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \alpha_i \text{ i.i.d.} \sim \mathcal{N}(0, \sigma_\alpha^2)$$

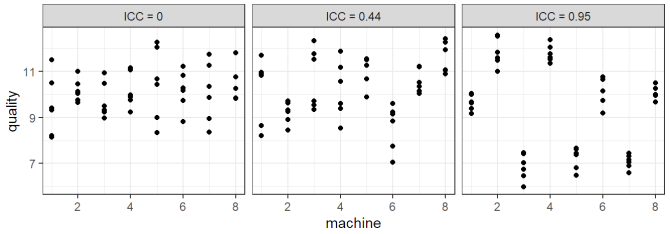
where α_i is the effect of the i samples, it is also called a **random effect**. Sometimes, such models are also called variance components models. Let us inspect some properties of the model.

$$\mathbb{E}[Y_{ij}] = \mu \qquad \text{Var}(Y_{ij}) = \sigma_\alpha^2 + \sigma^2$$

$$\text{Cor}(Y_{ij}, Y_{kl}) = \begin{cases} 0 & i \neq k \\ \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma^2) & i = k, j \neq l \\ 1 & i = k, j = l \end{cases}$$

Observations from different samples are uncorrelated while observations from the same sample are correlated. The correlation within the same sample is also called the intraclass

correlation (ICC). When large, it means that observations from the same sample are much more similar than observations from different samples.



Parameter estimation for the variance components $\sigma^2_\alpha, \sigma^2$ is done with so the called restricted maximum likelihood technique.

Confidence intervals are often larger than with fixed effect models, as we now try to make a statement about a larger population and not only about the measured samples.

7.1.2 More Than One Factor

So far this was a one-way ANOVA model with a random effect. We can extend this to the two-way ANOVA situation and beyond. For the two-way ANOVA situation we have the following model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Hereby α_i and β_j are the random (main) effects. From here we can apply the same techniques as before.

7.1.3 Nesting

We introduce a new data structure, where the level of factor B has a different meaning for every level of factor A. The two factors are **not crossed**, we say B is **nested** in A. We can use the following model:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}$$

Here α_i is the random effect of A and $\beta_{j(i)}$ is the random effect of B within A. We make the usual assumptions for the random effects:

$$\alpha_i \text{ i.i.d.} \sim \mathcal{N}(0, \sigma^2_\alpha), \qquad \beta_{j(i)} \text{ i.i.d.} \sim \mathcal{N}(0, \sigma^2_\beta)$$

7.2 Mixed Effects Models

In practice, we often encounter models which contain both random and fixed effects. We call them **mixed models** or **mixed effects models**. Let assume we have a fixed effect A and a random effect B. We can model our data as follows:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Here α_i is the fixed effect, β_j the random effect and $(\alpha\beta)_{ij}$ the random interaction effect. An interaction effect between a random and a fixed effect is treated as a random effect. We assume that all random effects are normally distributed, this means:

$$\beta_j \text{ i.i.d.} \sim \mathcal{N}(0, \sigma^2_\beta), \qquad (\alpha\beta)_{ji} \text{ i.i.d.} \sim \mathcal{N}(0, \sigma^2_{\alpha\beta})$$

Now the same techniques can be used again to analyse the fixed effects and the random effects.

8 Split-Plot Designs

In this section we are going to focus on experimental designs that contain experimental units of different sizes, with different randomizations. These are called **split-plot designs**.
A split-plot design has a **whole-plot factor**, treatment scheme was applied to plots, and a **split-plot factor** where the treatment gets applies to subplots. In the following example the whole-plot factor is *ctrl, new* and the split-plot factor is *A, B, C, D*.

1	2	3	4	5	6	7	8
ctrl	ctrl	new	ctrl	new	ctrl	new	new
D	*A*	*B*	*C*	*B*	*A*	*D*	*A*
A	*D*	*C*	*D*	*A*	*D*	*C*	*B*
C	*B*	*A*	*A*	*C*	*C*	*A*	*D*
B	*C*	*D*	*B*	*D*	*B*	*B*	*C*

As we now have two different sizes of experimental units, we also need two error terms to model the corresponding experimental errors. One error term acting on the plot level and another one on the subplot level. We end up with the following model:

$$Y_{ijk} = \mu + \alpha_i + \eta_{k(i)} + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where α_i is the fixed effect of the whole-plot factor and β_{ij} is the fixed effect of the split-plot factor. Further $(\alpha\beta)_{ij}$ is the interaction term and $\eta_{k(i)}$, ϵ_{ijk} are the errors on the plot and subplot level. Note the due to the whole-plot error, observations from the same plot are modelled as correlated data.

8.1 Properties of Split-Plot Designs

Typically, split-plot designs are suitable for situations where one of the factors can only be varied on a large scale. For example, fertilizer or irrigation on large plots of land. The price that we pay for this laziness on the whole-plot level is less precision, or less power, for the corresponding main effect because we have much fewer observations on this level. Note that the main effect of the split-plot factor and the interaction between the split-plot and the whole-plot factor are not affected by this loss of efficiency.

Typical signs for split-plot designs are:

- Some treatment factor is constant across multiple time-points, while another changes at each time-point.
- Some treatment factor is constant across multiple locations, while another changes at each location.
- When planning an experiment: Thoughts like, "It is easier if we do not change these settings too often".

If we are not taking into account the special split-plot structure, the results on the whole-plot level will typically be overly optimistic.

9 Incomplete Block Designs

The block designs in a previous section were complete, meaning that every block contained all treatments. This is not always possible, this leads to **incomplete block designs**. We have to decide what subset of treatments we use in an individual block. Bad decision, can lead to flawed designs, in the sense that certain quantities are not estimable anymore.

We cannot fit our standard main effects model to such a design, as it will lead to some linear functions not being

estimable. This can be due to so-called **disconnected design**, meaning part of the treatment / block set do not overlap, they are disjoint. Intuitively, we should have a good "mix" of treatments in each block.

9.1 Balanced Incomplete Block Designs

To achieve this good "mix", we can try to fulfill some optimality criterion. One criterion could be, that we can estimate all treatment differences with the same precision.

A **balanced incomplete block design** is an incomplete block design where all pairs of treatments occur together in the same block equally often, we denote this number by λ. How can we construct a BIBD? Let's define g as the number of treatments and k as the size of a block. For every setting $k < g$ we can find a BIBD by taking all possible subsets, where we have $\binom{g}{k}$. This is an unreduced balanced incomplete block design. In practice this might not be possible. Wether a BIBD exists is a combinatorial problem. A necessary, but not sufficient condition is that

$$\frac{r * (k-1)}{g-1} = \lambda$$

where r is the number of blocks and λ is the number of times two treatments occur together in the same block (hence, an integer).

9.2 Analysis of Incomplete Block Designs

The analysis of an incomplete block design is as usual. We use a fixed block factor and a treatment factor leading to:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

Because we do not observe all the block and treatment combinations equally often (some are simply missing), we are faced with an unbalanced design. We typically use sum of squares for treatment effects that are adjusted for block effects.

9.2.1 Intra- and Inter-block Analysis

Up until now, we estimated treatment effects by adjusting for block effects. This means that whatever is special to a block is fully allocated to the block effect and does not affect the treatment effect. Basically, the estimate of the treatment effect is based on the "leftovers." This is also known as an **intra-block analysis**.

On the other hand, if we treat the block factor as a random effect, the mean of the values of a block implicitly also contain information about the treatment effects. An analysis which is based on this information is known as an **inter-block analysis**. This leads to another estimate of the treatment effects. Both approaches can be combined.

10.1 Two-Sampled T-Test

10.2 Charts

```
stripchart(weight ~ group, vertical = TRUE, pch = 1,
    data = PlantGrowth)

boxplot(weight ~ group, data = PlantGrowth)
```

10 Various