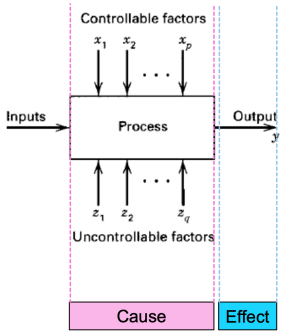
R wie Rotz

by dcamenisch

1 Learning from Data



We are in the abstract situation where we have a system 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).

1.1 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

1.1.1 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

1.1.2 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.

1.1.3 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.

1.1.4 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.

1.1.5 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.

2 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.

2.1 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**.

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

2.1.1 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: $\sum_{i=0}^{g} \alpha_i = 0$
- 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"))
```

2.1.2 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:

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

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

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

The estimate of the **error variance** is 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**:

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

Alternatively we can write this as:

$$MS_E = \frac{1}{N-g} \sum_{i=1}^{g} (n_i - 1) s_i^2, \qquad s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2$$

Where s_i^2 is the empirical variance in treatment group i. The demoninator $N-g$ ensures that σ^2 is an unbiased estimator (the error estimate has $N-g$ degrees of freedom).

```
fit <- aov(y ~ x, data = d)
## Get the estimated coefficients
coef(fit) ## or dummy.coef(fit)
## (Intercept) grouptrt1 grouptrt2
##       5.032     -0.371      0.494
```

2.1.3 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 **single mean model**, a special case of the cell means model. We have the following global H_0 and H_A:

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

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

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

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

Where SS_T is the total sum of squares, SS_{Trt} the treatment sum of squares (between groups) and SS_E the error sum of squares (within groups).

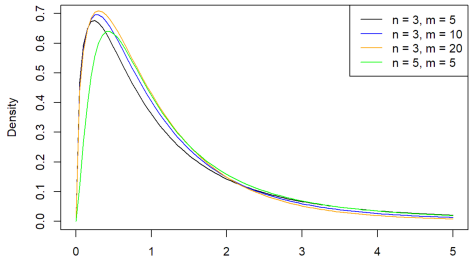
This information can be summarized in a **ANOVA** table.

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	$MS_E = \frac{SS_E}{N-g}$	

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:



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:

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

Where α is often chosen as 0.05. Since this test is based on the F-ratio we call it an 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
```

To perform statistical inference on the individual α_i's we use:

```
summary.lm(fit) ## for the tests
confint(fit)    ## for the confidence intervals
```

2.2 Checking Model Assumptions

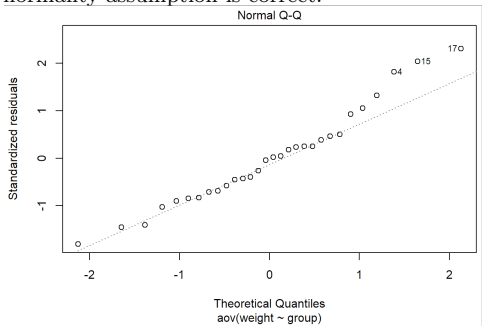
Statistical inference is only valid if all model assumptions are fulfilled:

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

The errors ϵ_{ij} cannot be observed, but the redisuduals $r_{ij} = y_{ij} - \hat{\mu}_i$ can be used as an estimate.

2.2.1 QQ-Plot

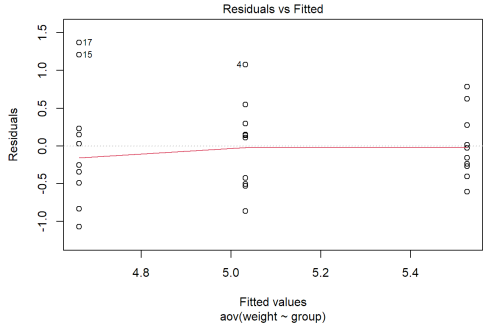
In a QQ-plot we plot the empirical quantiles of the residuals vs. the theoretical quantiles. The plot should show a more or less straight line if the normality assumption is correct.




```
plot(fit, which = 2)
```

2.2.2 Tukey-Anscombe Plot

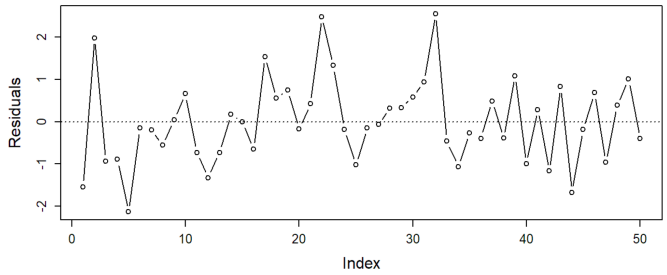
The Tukey-Anscombe plot (TA-plot) plots the residuals r_{ij} vs. the fitted values $\hat{\mu}_i$ (estimated cell means). It allows us to check that the residuals have constant variance and zero mean.




```
plot(fit, which = 1)
```

2.2.3 Index Plot

If the data has some serial structure, i.e. a 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. 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.



2.2.4 Transformations Affect Interpretation

Whenever we transform the response we implicitly also change the interpretation of the model parameters. Therefore, while it is conceptually attractive to model the problem on an appropriate scale of the response, this typically has the side effect of making interpretation more difficult. For example, if we use the logarithm:

$$\log(Y_{ij}) = \mu + \alpha_i + \epsilon_{ij}$$

All the α_i have to be interpreted on the log-scale. For example, if we us *contr.treatment* and we have $\hat{\alpha}_2 = 1.5$. This means: on the log-scale we estimate that the average value of group 2 is 1.5 larger than the average value of group 1. What about the original scale? We know that $\mathbb{E}[\log(Y_{ij})] = \mu + \alpha_i$, but the expected value on the original scale does not directly follow the transformation. However, we can make a statement about the median. On the log-scale the median is equal to the mean, hence:

$$\text{media}(\log(Y_{ij})) = \mu + \alpha_i$$

In contrast to the mean, any quantile directly transforms with a strictly monotone increasing function. As the median is nothing else than the 50% quantile, we have:

$$\text{media}(Y_{ij}) = e^{\mu+\alpha_i}$$

Similarly, for the ratio:

$$\frac{\text{media}(Y_{2j})}{\text{media}(Y_{1j})} = \frac{e^{\mu+\alpha_2}}{e^{\mu}} = e^{\alpha_2}$$

Hence, we can make a statement that on the original scale the median of group 2 is $e^{\alpha_2} = 4.48$ as large as the median of group 1. This means that additive effects on the log-scale become multiplicative effects on the original scale. Unfortunately, the statement is only about the median and not the mean on the original scale.

If we also consider a confidence intervals for α_2, e.g. [1.2, 1.8], the transformed version $[e^{1.2}, e^{1.8}]$ is a confidence interval for e^{α_2} which is the ratio of medians on the original scale.

2.3 Power / What Sample Size Do I Need?

By construction, a statistical test controls the type I error rate with the significance level α. This means that the probability that we falsely reject H_0 is less than or equal to α. There is also a type II error, occuring if we fail to reject H_0 even though H_A holds. The probability of a type II error is denoted by β.

The **power** of a statistical test is defined as

$$\text{P(reject } H_0 \mid \text{a certain setting under } H_A \text{ holds)} = 1 - \beta.$$

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.

2.3.1 Calculating Power for a Certain Design

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 do I want". This depends on:

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

We mainly the first two to maximize the power. Instead of doing the exact calculations, we choose an alternative way. We can simulate a lot of data sets under H_A that we believe in and check how often we are rejecting the corresponding H_0. The empirical rejection rate is then an estimate of the power. A nice side effect of doing a power analysis is that you do the whole data analysis on simulated data and you immediately see whether it works as planed. From a conceptual point of view, we can do this for any design. However, the number of parameters grows 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 an idea about the required sample size in the sense of whether we need 6 or 60 observations.

```
mu <- c(57, 63, rep(60, 3))
sigma2 <- 7

## This will give us the estimated power
power.anova.test(groups = length(mu), n = 4, between.var =
    var(mu), within.var = sigma2)

## We can replace the argument n with power to get
## and estimate for the needed sample size per group
power.anova.test(groups = length(mu), between.var =
    var(mu), within.var = sigma2, power = 0.8)
```

3 Contrast and Multiple Testing

3.1 Contrast

The F-test is rather unspecific, giving a yes/no answer. It does not tell us what treatment (or combination of treatments) is significant. Such 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.

```
library(multcomp)
## linfct is our contrast
fit.glht <- glht(fit, linfct = mcp(group = c(1, -1/2,
    -1/2)))
summary(fit.glht)
```

Every contrast has an associated sum of squares:

$$SS_C = \frac{(\sum_{i=1}^{g} c_i \bar{y}_{i\cdot})^2}{\sum_{i=1}^{g} \frac{c_i^2}{n_i}}$$

It has one degree of freedom and therefore $MS_C = SS_C$. We have:

$$\frac{MS_C}{MS_E} \sim F_{1,N-g}$$

Two contrasts c, c^* are orthogonal (estimates are independent) if:

$$\sum_{i=1}^{g} \frac{c_i c_i^*}{n_i} = 0$$

If we have g treatments, we can find $g-1$ different orthogonal contrasts (one dimension is already used by the global mean). A set of orthogonal contrasts partitions the treatment SS meaning that if $c^1, ..., c^{g-1}$ are orthogonal it holds that:

$$SS_{c^1} + ... + SS_{c^{g-1}} = SS_{Trt}$$

Multiple contrasts are all orthogonal if and only if for the matrix C that represents them, $C^\top C$ is diagonal.

3.2 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 H_0 are true. Somehow we have to take into account the number of tests that we perform to control the overall type I error rate.

If all H_0 hold, the probability of at least one false rejection is $1 - (1-\alpha)^m$.

We list the potential outcomes of a total of m tests, among which m_0 H_0 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 H_0 (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 H_0's.

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 $H_{0,j}$'s) 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.

What about the F-test? Should we only do pairwise comparison if the F-test is significant? No, the F-test is too conservative (already built-in multiple testing correction) and conditional error rates can be very bad.

3.2.1 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. This can also be performed by multiplying the individual p-values by m and using the original α.

```
library(multcomp)
## K is a matrix with each row being a contrast
fit.glht = glht(fit, linfct = mcp(group = K))
summary(fit.glht, test = adjusted("bonferroni"))
```

3.2.2 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.

```
summary(fit.glht, test = adjusted("holm"))
```

3.2.3 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:

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

3.2.4 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.

```
TukeyHSD(fit)
```

3.2.5 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.

```
fit.glht <- glht(fit, linfct = mcp(group = "Dunnett"))
summary(plant.glht)
```

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.

4 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**.

```
xtabs(~ factor1 + factor2, data = d)
```

This typically leads to questions about the interaction of the different factors (or if the interact at all).

4.1 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**:

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

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 $(a-1)(b-1)$.

4.1.1 Parameter Estimation

We estimate parameters using least squares and the 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}.

```
fit <- aov(y ~ a * b, data = d)
## alternatively: aov(y ~ a + b + a:b, data = d)
```

4.1.2 Tests

The total sum of squares SS_T can be partitioned into different sources.

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

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")	$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij\cdot})^2$
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	$MS_E = \frac{SS_E}{ab(n-1)}$	

In general, the degree of freedom of the error term is given by $N - (\text{DF } A) - (\text{DF } B) - (\text{DF } AB) - 1$.

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:

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

In R, we get the ANOVA table and the corresponding p-values again with the *summary* function. 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. The degree of freedom of the interaction form is the product of factors involved, e.g. $(a-1)(b-1)$.

4.1.3 Individual Analysis

If we have two factors A, B then instead of a full model, we might want to choose one model per individual level of A (e.g. due to some interaction). This can be improved by reusing the MS_E with the degree of freedom of the full model. This leads to a better power because the quantiles of the F-distribution will be smaller. Similar for contrasts we can use σ^2 estimates given by the MS_E of the full model. This is especially useful if the degree of freedom of the error term is small (< 10).

The trade off in power between these two tests is that, given the same degrees of freedom, the test with the larger F-value returns a smaller p-value and given the same F-values the test with the larger degrees of freedom in the denominator will return a smaller p-value. Often, the gain in degrees of freedom in the denominator outweighs the loss in the F-value.

4.1.4 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.

Quite often, we can get rid of interactions if we look at the problem on a different scale, i.e. if we transform the response appropriately. A famous example is the logarithm. Effects that are multiplicative on the original scale become additive on the log-scale, i.e. no interaction is needed on the log-scale.

If we have no replicates and more than two factors we often remove higher-order interaction terms (goes into error term).

4.1.5 Tukey One DF Interaction

The idea is to use only one additional term for the interaction. For a two-factor model this looks like the following:

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

λ is the new term and $\alpha_i\beta_j$ is the product of the main effects.

4.1.6 Checking Model Assumptions

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

4.1.7 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 and we cannot decompose the variability**. This problem can be solved by using a model comparison approach.

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): $SS(A|1, B, AB) \rightarrow SS(B|1, A, AB) \rightarrow SS(AB|1, A, B)$

The tests are the same for the interaction term. For the B factor type 1 and type 2 are the same.

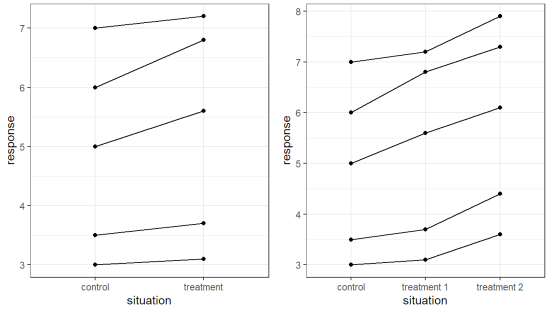
Type 1 is what we will typically get with *summary* in R. Hence we get different results whether we write $y \sim A * B$ or $y \sim B * A$. For type 2 we can either use the function *Anova* in the package *car* or we could compare the appropriate models with the function *anova* ourselves. For type 3 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*.

```r
## Type II sum of squares (Type III is similar)
library(car)
Anova(fit, type = "II", data = d)
```

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.

5 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.



5.1 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:

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

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.

Typically, we are not inspecting the p-value of the block factor, mainly because of the fact that we did not randomize blocks to experimental units and we already knew that blocks are different. 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.

Instead of a single treatment factor we can also have a factorial treatment structure within every block, e.g. a two-factor factorial which we would model as $Y \sim Block + A * B$. Here, we could actually test the interaction between A and B even if every level combination appears only once in every block. As we have multiple blocks, we have multiple observations for every level combination of A and B! However, a randomized complete block design can only be used with one blocking factor.

Source	df	
Block	$r-1$	
A	$a-1$	
B	$b-1$	
AB	$(a-1)\cdot(b-1)$	
Error	$(ab-1)\cdot(r-1)$	← "Leftovers"
Total	$rab-1$	← # observations − 1

We can test for interactions even if we only have one replicate per combination and block.

5.2 Multiple Block Factors

We can also block on more than one factor. A special case is the so-called **Latin Square design** where we have two block factors and one treatment factor having g levels each. Hence, this is a very restrictive assumption.

In a Latin Square design, each treatment appears exactly once in each row and once in each column. A Latin Square design blocks on both rows and columns simultaneously. We also say it is a **row-column design**.

	C_1	C_2	C_3	C_4
R_1	A	B	C	D
R_2	B	A	D	C
R_3	C	D	A	B
R_4	D	C	B	A

To analyze data from such a design, we use the main effects model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}$$

The design is balanced having the effect that our usual estimators and sums of squares are "working". In R, we would use the model formula $y \sim Block1 + Block2 + Treat$. We cannot fit a more complex model, including interaction effects, here because we do not have the corresponding replicates.

5.2.1 Graeco-Latin Squares

If we have another blocking criterion with g levels (denoted by Greek letters, e.g. with levels $\alpha, \beta, \gamma, \delta$), we can use a Graeco-Latin Squares design. The conditions are that the Latin letters (treatments) occur once in each row and column and the Greek letters (third block factor) occur once in each row and column, i.e. we have two superimposed Latin Squares. In addition, each Latin letter occurs exactly once with each Greek letter. We use the main effects model to analyze the data:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijkl}$$

Where α_i is the treatment, β_j the block factor 1, γ_k the block factor 2 and δ_l the block factor 3.

	C_1	C_2	C_3	C_4
R_1	Aα	Bγ	Cδ	Dβ
R_2	Bβ	Aδ	Dγ	Cα
R_3	Cγ	Dα	Aβ	Bδ
R_4	Dδ	Cβ	Bα	Aγ

5.3 Precision

In a RCB design, the squared standard errors are σ^2_{RCB}/r, where r is the number of blocks, and in a completely randomized design σ^2_{CRD}/n. If we want to have the same precision, we need to ensure that:

$$\frac{\sigma^2_{RCB}}{r} = \frac{\sigma^2_{CRD}}{n}$$

Therefore, if we knew both squared standard errors, we would have to use a ratio of:

$$\frac{n}{r} = \frac{\sigma^2_{CRD}}{\sigma^2_{RCB}}$$

σ^2_{RCB} is estimated by the MS_E of our RCB and σ^2_{CRD} can be estimated using a weighted average of MS_E and MS_{Block}. The relative efficiency is then defined as:

$$RE = \frac{\hat{\sigma}^2_{CRD}}{\hat{\sigma}^2_{RCB}}$$

And gives us the ratio n/r, which can be interpreted as how many experimental units would be needed by a CRD to achieve the same efficiency / precision. Easier for a quick check is to look at the ratio MS_{Block}/MS_E, because:

$$\frac{MS_{Block}}{MS_E} > 1 \;\Leftrightarrow\; RE > 1$$

6 Random & 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**.

6.1 Random Effects Model

6.1.1 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

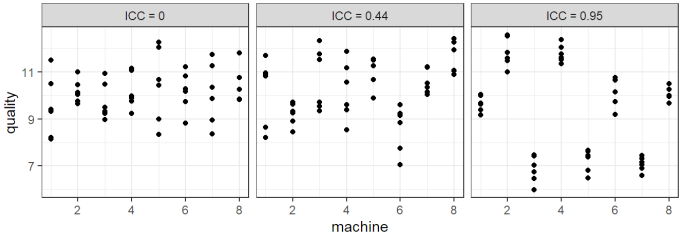
$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \alpha_i \text{ i.i.d.} \sim \mathcal{N}(0, \sigma^2_\alpha)$$

where α_i is the effect of the i samples, it is also called a **random effect**. There are no longer any sideconstraints on α_i. Sometimes, such models are also called **variance components models**. Let us inspect some properties of the model.

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

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

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 ($\sigma^2_\alpha >> \sigma^2$), it means that observations from the same sample are much more similar than observations from different samples.



The same holds for multiple random effects. For them, the correlation is the sum of shared variance components divided by the sum of all variance components. Parameter estimation for the variance components σ^2_α, σ^2 is done with the so-called restricted maximum likelihood technique. The total variance of Y_{ij} is estimated as $\hat{\sigma}^2_\alpha + \hat{\sigma}^2$ and the intraclass correlation as $\hat{\sigma}^2_\alpha/(\hat{\sigma}^2_\alpha + \hat{\sigma}^2)$.

```
library(lme4) ## lmerTest would be an alternative
fit <- lmer(y ~ (1 | x), data = d)
## As usual we can use summary and confint
## Linear mixed model fit by REML ['lmerMod']
## ...
## Random effects:
##  Groups   Name        Variance   Std.Dev.
##  x        (Intercept)      117       10.8
##  Residual                  464       21.5
## Number of obs: 40, groups: sire, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  82.55         5.91    14
```

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. In general, variances are "difficult" to estimate in the sense that we need a lot of observations to have some reasonable accuracy. To verify our model assumptions we can again use QQ-plots, but we have to also plot σ^2_α:

```
## QQ-plots of random effects
qqnorm(ranef(fit)$x[,1], main = "x")
## QQ-plots of residuals
qqnorm(resid(fit), main = "residuals")
```

If we would fit a normal one-way ANOVA model, we could estimate σ^2_α by $\frac{MS_A - MS_E}{N}$.

6.1.2 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.

```
fit <- lmer(y ~ (1 | a) + (1 | b) + (1 | a:b), data = d)
```

6.1.3 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:

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

```
fit <- lmer(y ~ (1 | a/b), data = d)
## Alternatively
fit <- lmer(y ~ (1 | a) + (1 | a:b), data = d)
```

6.2 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.

```
options(contrasts = c("contr.treatment", "contr.poly"))
library(lmerTest)
fit <- lmer(y ~ a + (1 | b) + (1 | b:a), data = d)
```

7 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_j 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.

```
library(lmerTest)
fit <- lmer(mass ~ a * b + (1 | plot), data = d)
```

7.1 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.

7.2 Split-Split Plot Design

If we have more than two factors, a split-split plot design can be performed. For example, consider the following experiment design: The yield of oats from a split-plot field trial using 3 varieties and 4 levels of manurial treatment. The experiment was laid out in 6 blocks of 3 main plots, each split into 4 sub-plots. The varieties were applied to the main plots and the nitrogen treatments to the sub-plots.

A whole plot is given by a plot of land in a block (B), the whole-plot factor is variety (V). A block design (RCB) was used at the whole-plot level. A split plot is given by a subplot of land, the split-plot factor is given by nitrogen treatment (N). The mathematical model is:

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

```
fit <- lmer(Y ~ B + V*N + (1 | B:V), data = d)
```

8 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** (IBD). 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. If we would fit a main effect model to a disconnected design, multiple treatment coefficients can be set to zero (not only one). Intuitively, we should have a good "mix" of treatments in each block.

8.1 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, i.e. all confidence intervalls for $\alpha_i - \alpha_j$ have the same width (for any pair i, j).

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 replicates per treatment and λ is the number of times two treatments occur together in the same block (hence, an integer). By definition we have $N = b * k = g * r$.

```
library(ibd)
des.bibd <- bibd(v = 6, b = 10, r = 5, k = 3, lambda = 2)
## arguments of function bibd are:
## - v: number of treatments
## - b: number of blocks
## - r: number of replicates (across all blocks)
## - k: number of experimental units per block
## - lambda: lambda
des.bibd$design ## here, blocks are given by rows
des.bibd$NNP    ## gives the concurrency matrix
```

In a partially balanced incomplete block design, some treatment pairs occure together more often than other pairs.

Row-Column IBD - In these designs, we have two block factors (rows and columns) and one or both of them are incomplete blocks.

Youden Squares - A Youden Square is rectangular such that the columns form a BIBD and for the rows each treatment appears equally often in each row. The columns therefore form a BIBD and the rows an RCD.

8.2 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.

```
tab <- xtabs(~ a + b, data = d)
## We have to transform the design information to
## the desired form
library(crossdes)
m <- t(apply(tab, 2, function(x) (1:4)[x != 0]))
## isGYD now tells us if the IBD is balanced
isGYD(m)
```

We fit the model: `y ~ block + treatment` (instead of fitting `y ~ treatment + block`). Since want to check whether the treatment has any influence on the response after having controlled for the variation between blocks (only important with `summary`). As it is an unbalanced data set, we use *drop1* to analyse the data, such that we get the sum of squares.

```
fit <- aov(y ~ a + b, data = d)
drop1(fit, test = "F")
```

8.2.1 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.

```
library(lmerTest)
fit.ibd <- lmer(y ~ treat + (1 | block), data = dat)
summary(glht(fit.ibd, linfct = mcp(treat = c(1, 0, 0, 0, 0,
    -1))))
```

9 Various

Standard Deviation $= \sqrt{\text{Variance}}$, Standard Deviation $= \sigma$, Variance $= \sigma^2$

9.1 Two-Sampled t-Test

9.1.1 Unpaired Data

We have X_i i.i.d. $\sim \mathcal{N}(\mu_X, \sigma^2)$ and Y_j i.i.d. $\sim \mathcal{N}(\mu_Y, \sigma^2)$ with X_i, Y_j independent. For the t-test, $H_0 : \mu_X = \mu_Y$ and $H_A : \mu_X \neq \mu_Y$. Then:

$$T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{s^2}{n} + \frac{s^2}{m}}} \sim t_{n+m-2} \text{ under } H_0$$

9.1.2 Paired Data

We have independent $D_i = X_i - Y_i$ and:

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \sim \mathcal{N}(\mu_D, \sigma_D/\sqrt{n})$$

H_0 and H_A as before and:

$$T = \sqrt{n}\frac{\bar{D}}{S_D} \sim t_{n-1} \text{ under } H_0$$

```
t.Per1 <- d$AGGREG[d$PERIODE == 1]
t.Per2 <- d$AGGREG[d$PERIODE == 2]
t.test(t.Per1, t.Per2, paired = TRUE)
##
##  Paired t-test
##
## data: t.Per1 and t.Per2
## t = -4.27, df = 10, p-value = 0.0016
## alternative hypothesis: true difference in means is not
     equal to 0
## 95 percent confidence interval:
##  -15.6311 -4.9143
## sample estimates:
## mean of the differences
##                -10.273
```

The two-way ANOVA with single replicates and the t-test give exactly the same results.

9.2 Charts

```
stripchart(y ~ x, vertical = T, pch = 1, data = d)
boxplot(y ~ x, data = d)
with(d, interaction.plot(x.factor = a, trace.factor = b,
     response = y))
```

If the lines of an interaction plot are NOT parallel, there is possibly an interaction effect we have to consider.

9.3 Data Generation / Calculations

```
## Generate 10 A's followed by 10 B's
rep(c("A", "B"), each = 10)

## Alternate A, B 10 times
rep(c("A", "B"), times = 10)

## Toss a coin 20 times (1/2 prob. for A, B)
sample(c("A", "B"), 20, replace = T)

## Choose 10 A's at random, the rest B's
sample(rep(c("A", "B"), times = 10), 20, replace = F)

## Overall mean of column A
mean(d$A) ##or aggreagte(A ~ 1, data = d, mean)

## Group mean per B
aggregate(A ~ B, data = d, mean)
```

9.4 Examples

9.4.1 Split-Plot Design, ANOVA Skeleton

Three new types of pizzas in six different packings are investigated by 90 consumers on a 0-10 scale. Each person rates the six packings of just one type of pizza, that is pizzas are randomized to persons and each person tastes the different packings in random order.

This is a split-plot design with persons as whole plots and rating orders (or time slots) as split plots. Pizza type is the whole-plot factor, packing the split-plot factor. We have:

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

Where $\eta_{k(i)}$ is the whole-plot error. The ANOVA skeleton is given by:

Plot level	Source	df	MS	F
Whole plots	pizza	2	MS_B	$\frac{MS_B}{MS_\eta}$
	residual	87	MS_η	
Split plots	packing	5	MS_A	$\frac{MS_A}{MS_E}$
	packing:pizza	10	MS_{AB}	$\frac{MS_{AB}}{MS_E}$
	residual	435	MS_E	
	total	539		

9.4.2 Split-Plot Design with Blocking

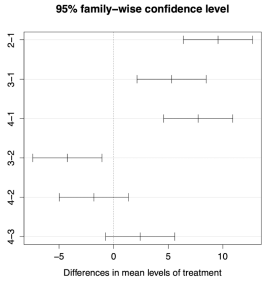
A soil scientist wanted to investigate the effects of nitrogen supplied in four different forms and later evaluate those effects combined with those of thatch accumulation (two, five or eight years of accumulation) on the quality of an established turf. A golf green had been constructed and seeded with grass on the experimental plots. The nitrogen treatment plots were arranged on the golf green in a randomized complete block design with two block levels. Each of the eight experimental plots was split into three subplots to which the levels of the second treatment factor were randomly assigned.

This is a split-plot design with whole-plot factor nitrogen, split-plot factor thatch and a block factor block.

$$Y_{ijkl} = \mu + \gamma_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \eta_{l(ij)} + \epsilon_{l(ijk)}$$

Where $l = 1$, γ_i fixed effect of the block, α_j fixed main effect of nitrogen, β_k fixed main effect of thatch, $(\alpha\beta)_{jk}$ interaction, $\eta_{l(ij)}$ error on the whole-plot level and $\epsilon_{l(ijk)}$ the error on the split-plot level.

Ex. We fitted the RCB experiment and plot `plot(TukeyHSD(x = "fit", "treatment", conf.level = 0.95))`.



Which treatment differences are significant?

- ☐ $2-1, 3-1, 4-1$
- ☐ $3-2$
- ☐ $4-2, 4-3$
- ☐ $2-1$
- ☒ $2-1, 3-1, 4-1, 3-2$

Since a treatment difference is significant if the corresponding confidence interval does not cover 0.

Ex. Given a two-way ANOVA model with a significant factor A, an insignificant factor B and a significant interaction term. Can we remove any terms from the model?

No, we should not remove a significant term and since the interaction is significant we cannot remove the main effect B.

Ex. When fitting two ANOVA models with `fit1 <- aov(y age * educ, data = survey)` and `fit2 <- aov(y educ * age, data = survey)`, we might encounter different p-values for both `age` and `educ`, what might be the reason?

If there is a difference, the design must be unbalanced. In that case, the type I sum of squares used in `aov()` is dependent on the order of how the variables appear in the model. The interaction has the same p-value because it is adjusted for `age` and `educ` in both cases.

Ex. The mixed effect model $\text{Rating}_{ij} = \mu + \text{Type}_i + \text{Subject}_j + \epsilon_{ij}$ is used. Type is a fixed effect and Subject a random (block) factor. For fixed j, are the Rating_{ij} uncorrelated?

No, they are not as subjects have an effect since $\text{Var}(\text{Subject}_i)) \neq 0$.

Ex. Which statement about RCB is true:

- ☐ A RCB tries to ensure heterogeneity within blocks
- ☒ The error with r blocks hard $r-1$ less DF than the error of a completely randomized design with the same number of experimental units.
- ☐ The number of experimental units in a block can be less than the number of treatments
- ☐ Can be constructed from a completely randomized design by imposing blocking after randomization

Ex. Given three levels of a treatment, which design should be preferred if interested in difference of treatments (assuming the variance of the units is the same for both designs)?

D1: 2 pieces of land with 3 sub-fields each and form 2 complete blocks

D2: 3 pieces of land with 2 sub-fields each and form 3 blocks

- ☒ D1 should be preferred

☐ D2 should be preferred

☐ There is no important difference between D1 and D2

The incomplete block design D2 has more variance in the estimates of treatment differences than the complete block design D1. Therefore D1 should be preferred.

Ex. Consider a balanced one-way ANOVA model. What happens to the 95%-quantile of the F-distribution of the global test if we increase the number of observations?

☐ The quantile gets larger

☒ The quantile gets smaller

☐ The quantile stays the same

Ex. Consider a test with 4 types of toothpaste and 3 types of packaging. 60 participants have been selected and each participant is supposed to test and rate every packaging of exactly one toothpaste type. Which type of design is this?

☒ Split-plot design with participants as whole-plots

☐ Split-plot design with toothpaste as whole-plots

☐ Split-plot design with packagings as whole-plots

Ex. Consider the following experimental design with a treatment factor having levels A, B, C, D, E, F. Which of the statements is true?

Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
A	A	B	D	D	B
B	C	C	E	F	F

☐ It is not possible to estimate all treatment differences

☐ The design is balanced

☐ This is an unreduced BIBD

☒ If estimating each of the treatment differences $A-B, F-B$, and $D-F$ has variance $2\sigma^2$, then estimating the treatment difference $A-D$ via the decomposition $A-D = (A-B)-(F-B)-(D-F)$ has variance $6\sigma^2$.

Ex. Which of the statements is **wrong**?

☐ Blocking is used to reduce variance by choosing blocks such that each has homogeneous experimental units.

☒ In an RCB, we can never test interactions between block and treatment.

☐ In an RCB with a factorial treatment structure, we can test interactions between treatment factors even if we only observe every treatment combination once in each block.

☐ Blocking can increase precision, even if the p-value corresponding to the block factor is not significant.

Ex. Which of the statements about multiple testing is true?

☐ The more different tests we perform (each at a fixed level α), the less likely we are committing a type I error.

☐ Tukey Honest Significant Difference is less powerful than the Bonferroni correction.

☒ The family-wise error rate (the probability of committing a type I error) can always be controlled by using the Bonferroni procedure.

Ex. Given 3 p-values $(0.12, 0.48, 0.09)$, applying Bonferroni correction to the p-values. What is the 2nd largest p value?

☐ 0.03 ☐ 0.12 ☒ 0.27 ☐ 0.36

The corrected p-values are $(0.36, 1.44, 0.27)$.

Ex. Assuming a fixed effect model with crossed factors A, B. Factor A has a fixed effect on the response and factor B has a random effect on the response. There is also a random effect specifically for the interaction between A and B. The fixed effect and confidence intervals of A are interpreted as...

☒ the population average of the effect of A (across all levels of B).

☐ the average effect of A for the observed levels of factor B.

Ex. Select the correct statement about Latin Squares.

☐ In a Latin Square design, the treatments are randomly assigned to the combination of levels of the block factors (without any restriction).

☐ In a Latin Square design, every treatment appears exactly once for each combination of the levels of the block factors.

☒ In a Latin Square design, every treatment appears exactly once for each level of any of the two block factors.

☐ A Youden square design is an extension of a Latin Square design to three block factors.

Ex. Conduct an experiment by recording the exam scores of students and whether they attend the exercise courses or not. A student is randomly assigned to go to the exercise session or not. There is a significant, positive effect of attendance on exam scores. Is this a well-designed and valid experiment?

☐ Yes ☒ No

There is only one experimental unit since there is only one exercise session and not multiple.

Ex. Andrew and George want to fit a One-Way Anova model. For the encoding scheme of factors, Andrew uses the `contr.sum` constraint, whereas George uses the `contr.treatment` constraint.

☐ The estimated coefficients and predicted values will be the same.

☐ The estimated coefficients will be the same, but the predicted values per treatment group will be different.

☒ The estimated coefficients will be different, but the predicted values per treatment group will be the same.

☐ The models will be different and might even yield to different statistical inference.

Ex. You perform a test of 10 hypotheses and get the following 10 p-values: 0.001, 0.004, 0.012, 0.024, 0.043, 0.048, 0.051, 0.089, 0.212, 0.762. You want to control the FWER at 5% and use a Bonferroni correction. How many hypotheses do you reject?

☐ 1 ☒ 2 ☐ 3 ☐ 4

Ex. The TAs want to fit a two-way ANOVA model but the data is unbalanced. The first model is `aov(y ~ A*B)` and the second one is `aov(y ~ B*A)`. What do we expect from the results?

☐ All the p-values will be the same.

☒ Only the p-value of the interaction term is guaranteed to be the same in both cases.

☐ Only the p-values of the main effects are guaranteed to be the same.

☐ The fitted coefficients are different.

Ex. The denominator mean square for F-tests in random effects models will always be the MS_E, like in fixed effects models.

☐ True ☒ False

Ex. Suppose we are given data coming from a Split-Plot design experiment and we decide to fit a two-way ANOVA instead of the correct model. What can we say in general about the resulting p-values?

☐ The p-values will be the same.

☐ The p-values will pessimistic.

☒ The p-values will optimistic.

Ex. Given the random effect model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1...4,\ j = 1...5$, what is the number of parameters?

☐ 2 ☒ 3 ☐ 4 ☐ 5

The parameters are μ, σ^2_α and σ^2.

Ex. You run a random effects model and get the following R-output. Can you reject the null-hypothesis that the true variance for random effect of batch is equal to 0.3?

```
Computing profile confidence intervals...
sd_(Intercept)|batch       0.50143  1.19412
sd_sigma.                  0.58312  0.97491
(Intercept)               10.28499 11.24959
```

☐ Yes, can reject H_0

☒ No, cannot reject H_0

We have to square the confint to get the confint for the variance.

Ex. We are testing some hypothesis while controlling the Type I error at level alpha = 0.05. Is the following statement True or False? If we increase the sample size N, the Power increase and the Type I error probability decreases.

☐ True ☒ False

The Type II error decreases while the Type I error is fixed at level alpha!

Ex. The denominator mean square for F-tests in random effects models will always be the MS_E, like in fixed effects models.

☐ True ☒ False

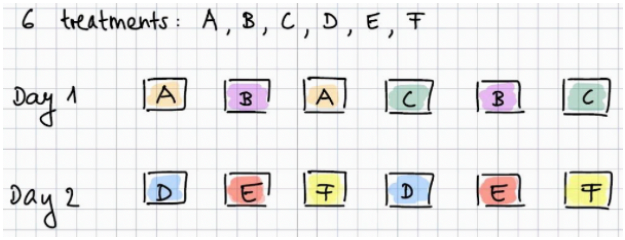
Ex. Assume we perform 100 statistical tests. We use the rule to reject the H_0 if the corresponding p-value is less than 0.05. If all 100 null hypotheses are true, then we expect to get ...

☒ 5 significant results.

☐ 5 significant results but only if the tests are independent.

☐ Cannot be judged with the available information.

Ex. Consider the following block design and the statements: 1) we can estimate the difference between treatment A and C, 2) we can estimate the difference between treatment A and D.



☐ True / True ☒ True / False ☐ False / True ☐ False / False

The day component is disconnected, therefore we cannot decide if a difference comes from the different days and we can only make statement about the results within the same day.