# STA2020 ANOVA Notes

Ané Cloete

# Table of contents

# Preface

*Welcome to the Experimental Design and ANOVA section of STA2020.*

This book is not an exhaustive guide to designing experiments or conducting ANOVA. Instead, it has been tailored specifically to align with the learning outcomes and methods covered in STA2020.

This module consists of four main sections:

1. Experimental Design
2. Completely Randomized Designs
3. Randomized Complete Block Designs
4. Factorial Experiments

The first two chapters lay the groundwork for the module. Once you grasp these concepts, the remaining sections should be easier to follow. Before diving into these topics, there are two preliminary sections:

1. A brief introduction to statistical modeling
2. A guide to hypothesis testing

I encourage you to read through these first, as they provide essential context for the rest of the material.

Throughout the book, you will find R code presented in chunks like this:

```r
x <- c(1,2,3,4,5)
mean(x) # Computes the mean of a set of numbers
```

```
[1] 3
```

R is consistently used to visualize, illustrate, and demonstrate key methods and concepts. Running the code yourself will greatly enhance your understanding, so I encourage you to do so.

> Some parts of these notes have been adapted from the STA1007 notes, authored by Dr. Res Altwegg and Dr. Birgit Erni, as well as from various textbooks.

# Statistical Modelling

## What is a Model?

A **statistical model** is a mathematical representation of how data is generated. It describes the relationship between observed data and underlying factors (parameters) while accounting for random variation. Suppose that we are interested in estimating the age of a tree from its stem diameter. To do this we need to know by how much the stem diameter increases per year. We could describe this relationship or process as follows:

$$D = \alpha + \beta \times Age$$

describing a linear increase of diameter with age. Once we have a good idea of how fast diameter increases with age ( ) we can predict diameter from age. The (mathematical) model above is a very simple representation of this process with only two parameters, the intercept and the growth rate.

With the chosen parameter values, diameter increases linearly with age. Of course, this model is not realistic except for special situations but it gives us powerful insights. In reality we don't know $\beta$, but usually need to estimate it from data. Also, not every tree grows equally fast, because of environmental and individual differences between trees. We can accept that the above is a simple model for the average behaviour of a tree, but to capture variability between trees (because of variability between environmental conditions from tree to tree, variability between individual trees, measurement error), we add an error term.

$$D = \alpha + \beta \times Age_i + e_i$$

The response that we observe is then described by an average behaviour, but the actual observed value will vary around this average. To summarise, the statistical model has a stochastic component which captures variability in the response that cannot be explained by the deterministic part of the model. Another distinguishing feature of statistical modelling is that we obtain estimates of the parameter values from the data, e.g. by fitting a line to the observations, i.e. we learn from data.

## More generally

Statistical models are not perfect predictors of the data, rather they attempt to describe the "central tendency" of the observations. To get to the actual observed value some deviation from the central tendency needs to added (i.e. error). Such models typically have the following the form:

$$\text{Observed Response} = \text{Model Predicted Response} + \text{Error}$$

Mathematically this can be stated as:

$$Y = \hat{Y} + e$$

A simple example of a statistical model you may have encountered is the **mean** as a predictor. Suppose you measure the number of customers entering two stores over 20 days. The observed counts for each store fluctuate daily, but you may want to summarize the data using the average number of customers.

For each store $i$, a basic statistical model for these observations would be:

$$Y_{ij} = \mu_i + e_{ij}$$

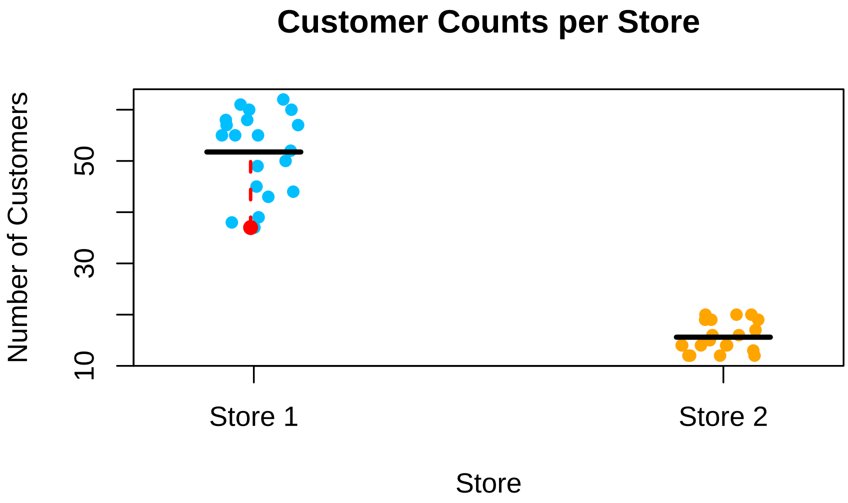
where:

- $Y_{ij}$ is the number of customers observed on day $j$ at store 1,
- $\mu_i$ is the true mean number of customers at store $i$,
- $e_{ij}$ is the error term, representing deviations from the mean.

The error term $e_{ij}$ accounts for day-to-day fluctuations that cause the actual number of customers to vary around the mean. Below this data is simulated and plotted, with the model overlain. The black line is the mean and the red dashed line represents the error for one observation, i.e. deviation from the fitted model response, in this case the mean.

```r
store1 <- rpois(20, 50)
store2 <- rpois(20, 15)
storedata <- data.frame(numcust = c(store1, store2),
                        store = factor(rep(c("Store 1", "Store 2"), each = 20)))

stripchart(numcust ~ store, data = storedata,
           method = "jitter", pch = 16, col = c("deepskyblue", "orange"),
           vertical = TRUE, main = "Customer Counts per Store",
           xlab = "Store", ylab = "Number of Customers")
means <- tapply(storedata$numcust, storedata$store, mean)
segments(x0 = 1:2- 0.1, x1 = 1:2 + 0.1, y0 = means, y1 = means, lwd = 3, col = "black"
min_count <- min(storedata$numcust[storedata$store == "Store 1"])
```

```
min_x <- jitter(rep(1, sum(storedata$numcust == min_count)))
points(min_x, min_count, col = "red", pch = 16, cex = 1.2)
segments(x0 = min_x, x1 = min_x, y0 = min_count, y1 = means["Store 1"], col = "red", lwd = 2, lty
```



Another basic example of this structure is a **linear regression model**:

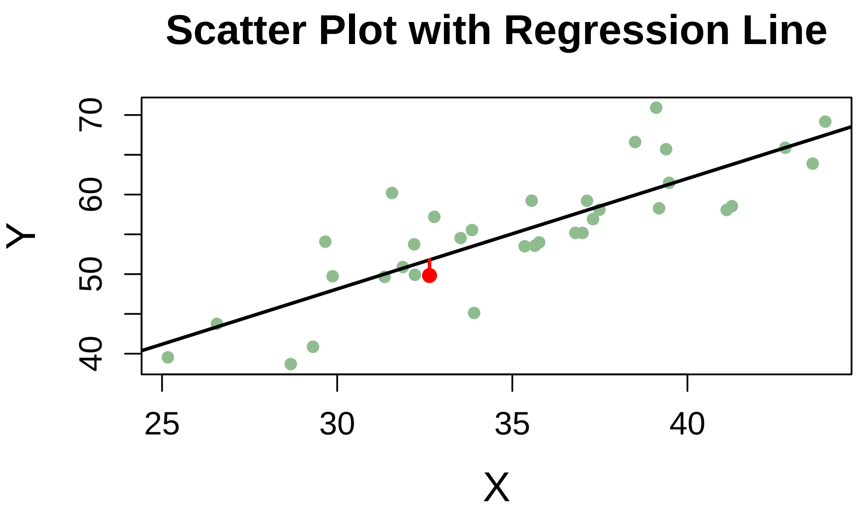
$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

where:

- $Y_i$ is the observed response,
- $\beta_0$ and $\beta_1$ are unknown parameters representing the intercept and slope,
- $X_i$ is the predictor variable,
- $e_i$ is the random error term.

```
# Generate random x values and error term
set.seed(123)  # Ensures reproducibility
x <- rnorm(35, mean = 35, sd = 5)
error <- rnorm(35, mean = 0, sd = 5)

# Define true model parameters
beta0 <- 2
beta1 <- 1.5

# Generate y values based on the regression model
y <- beta0 + beta1 * x + error
```

```
# Fit a linear regression model
model <- lm(y ~ x)  # This was missing!

# Select an observation to highlight
obs_index <- 20
x_obs <- x[obs_index]
y_obs <- y[obs_index]
y_pred <- predict(model, newdata = data.frame(x = x_obs))

# Scatter plot of data points
plot(x, y, pch = 16, col = "darkseagreen",
     xlab = "X", ylab = "Y",
     main = "Scatter Plot with Regression Line",
     cex.lab = 1.5, cex.axis = 1.2, cex.main = 1.5)

# Add regression line
abline(model, col = "black", lwd = 2)

# Highlight the observed point
points(x_obs, y_obs, col = "red", pch = 16, cex = 1.2)

# Draw a dashed vertical line from the predicted value to the observed value
segments(x0 = x_obs, x1 = x_obs, y0 = y_pred, y1 = y_obs, col = "red", lwd = 2, lty =
```



# Notation

When we fit the model to our data, we **estimate** the unknown parameters using observed data. We denote these estimates using **hat notation** to distinguish them from the true (but unknown) population parameters:

$$\hat{\beta}_0, \quad \hat{\beta}_1$$

Similarly, the **fitted values** (model-predicted responses) are denoted as:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

Thus, after fitting the model, the observed response can be rewritten as:

$$Y_i = (\hat{\beta}_0 + \hat{\beta}_1 X_i) + e_i = \hat{Y}_i + e_i$$

where:

- $\hat{Y}_i$ is the **fitted (predicted) value**, and
- $e_i = Y_i - \hat{Y}_i$ is the **residual**, representing the difference between the observed and predicted values.

# A brief guideline to hypothesis testing

> These notes have been adapted from the STA1007 notes (authored by Dr Res Altwegg and Dr Greg Distiller and some other textbooks.

Hypothesis testing is a statistical procedure of using sample data to make inferences about populations. Unlike estimation, where the goal is to quantify a parameter, hypothesis testing assesses whether an observed effect is statistically significant. More specifically, a hypothesis test evaluates two mutually exclusive statements about the population and determines which statement the data supports.

## The General Framework

Hypothesis testing follows a structured process:

1. **State the Hypotheses**: Define the null hypothesis (H ) and the alternative hypothesis (H ).

The basic idea of hypothesis testing is that we set up a so-called null hypothesis and then ask how likely our data are if the null hypothesis were true. If they are unlikely, we conclude that we have found evidence against the null hypothesis, i.e. the null hypothesis is probably not true.

The alternative hypothesis covers all the possibilities not covered by the null hypothesis. If we conclude that the null hypothesis is probably not true, that means that the alternative hypothesis is probably true. These two hypotheses are not equal in how we treat them:

- We start by assuming the null is true and check if the data gives enough evidence to reject it.
- If the data strongly contradicts the null, we lean toward the alternative hypothesis.

But we never prove the alternative hypothesis outright—we only show that the null is unlikely based on the evidence. You can think of the null hypothesis as representing a baseline against which the data are compared, whereas the alternative hypothesis is what we really care about, worry about or want to demonstrate. This is an important asymmetry and will need some careful reflection.

Below is an example:

Null Hypothesis ($H_0$): “The average weight of chocolate bars is 100g.”

Alternative Hypothesis ($H_A$): “The average weight of chocolate bars is less than 100g.”

Lack of evidence against $H_0$ is not the same as evidence for $H_0$. We never say that we have evidence for $H_0$ or that we accept $H_0$ as true.

2. **Choose a Test Statistic**: Select an appropriate statistic to measure the observed effect.

A numerical function of the data that quantifies the strength of the observed effect, whose value determines the result of the test. Examples include the mean difference, proportion difference, or z-score.

3. **Determine the Null Distribution:** Establish what the test statistic would look like if H  were true.

We have a test statistic and to say something about how likely this test statistic (or more extreme is) under the null hypothesis, we need the null distribution of the test statistic (that is the sampling distribution of the test statistic as if the null hypothesis were true). We then compared the observed value of the test statistic to that null distribution and asked ourselves how unusual it is in light of that distribution.

4. **Compute the P-value:** Calculate the probability of obtaining a test statistic as extreme as the observed one under H .

The probability of obtaining a result as extreme as the observed one if H  is true. A small P-value (typically <0.05) suggests strong evidence against ($H_0$).

5. **Make a Decision:**

In the approach you have been taught, we compare the P-value to a predefined significance level ( ) and conclude whether to reject $H_0$. Here we would like to emphasise that the p-value is a measure of evidence against $H_0$ - see below!

# One-Sided vs. Two-Sided Tests

Two-sided test: Tests for deviations in both directions. Example: “The average human body temperature is different from 37°C.”

One-sided test: Tests for deviations in a single direction. Example: “Students who study more than an hour score higher.”

## Decision Making in Hypothesis Testing

A small P-value constitutes evidence against $H_0$. But how small is small enough? Sometimes, we want to make a firm decision about whether we can believe that the observed pattern is real or not. This requires us to choose a threshold for P. This threshold is called the significance level and denoted by $\alpha$. If we obtain a P-value that is smaller than $\alpha$, we say that we have obtained a "statistically significant result" or that "$H_0$ is rejected". If our P-value is larger than $\alpha$, we say that our result is "not significant" or that "$H_0$ is not rejected". In most situations, researchers choose a significance level of $\alpha = 0.05$, which roughly corresponds to the probability of obtaining five heads in a row when tossing a fair coin, not a very likely event! Different values for $\alpha$ are also sometimes used; the next most common significance level is $\alpha = 0.01$.

Before we go further, we want to emphasize that there is nothing magic about a specific value of $\alpha$. This threshold is an arbitrary choice and should not be taken too seriously. There is not much difference between a P-value of 0.051 and 0.049. Both constitute about the same strength of evidence against $H_0$. Yet, when we apply $\alpha = 0.05$, we would reach opposite conclusions in the two cases. It is always better to report the exact P-value rather than just state $\mathrm{P} > 0.05$ or $\mathrm{P} < 0.05$ or state that a result is "not significant" or "significant". And it is particularly important not to imply that a "non-significant" result means that there is no effect (that would be saying $H_0$ is true when we might in fact have some evidence against it)!

Alas, dividing results into "significant" vs "not significant" is very entrenched in many fields and you will encounter these terms a lot. And used wisely, this distinction can have its merits. So we'll stick with it.

# Part I

# Experimental Design

# Chapter 1

# Experiments and experimental design

There are two fundamental ways to obtain information in research: by *observation* or by *experimentation*. In an observational study the observer watches and records information about the subject of interest. In an experiment, the experimenter actively manipulates variables hypothesized to affect the response (insert small example). Although both are important ways of understanding the world around us, only through experiments can we **infer causality**.

That is, by designing and conducting an experiment properly, if we observe a result such as a change in variable A leads to a change in our response (say variable B), we can confidently conclude that A **caused** this change in B. If we were to merely study variable B and observe that as variable A changes, B also changes without conducting an experiment, then we can only say that variable A and B are associated. We could not easily conclude that any change in B is due to A. It could be some other factor that is correlated with A or it could be that B caused the change in A! The key is that a well-designed experiment controls and holds constant (as best we can) all other factors that might affect the response, so we can be sure the result is caused by the variable we manipulated.

Imagine a company wants to determine whether their voluntary employee training program (the explanatory variable) increases productivity (the response). They decide to track the productivity of employees who chose to complete the training and those who did not. They note that, on average, trained employees are more productive. Can we confidently conclude that the training program caused increased productivity?

This is an observational study since no variable was actively manipulated, they merely observed and recorded the productivity of two groups of employees. So, we cannot conclude that completing the training program increases productivity - we cannot infer causality. It could be due to many other factors, either observed or unobserved, such as maybe employees who choose to do the training program are inherently more motivated and thus productive. Can you think of any other factors?

If they actively manipulate the explanatory variable, training program, by randomly assigning employees to complete the training program or not and control other factors by ensuring the employees are as similar as possible accross the groups (i.e. conducted an experiment). Any differences in productivity between the two groups could then be ascribed to the training program. If they happen to find that the employees who were assigned the training program are more productive, they can confidently say that the program caused increased productivity (and perhaps make it compulsory for all employees!).

Experimental studies are extremely important in research and in practice. They are almost the only way in which one can control all factors to such an extent as to eliminate any other possible explanation for a change in a response other than the variable actively manipulated. In this course, we only consider experimental studies and those which aim to compare the effects of a number of **treatments** (comparative experiments).

Here are some other reasons for conducting experiments:

1. They are easy to analyse. A well designed experiment results in independent estimates of treatment effects which allow us to easily interpret the effects.

2. Experiments are frequently used to find optimal levels of variables which will maximise (or minimise) the response. Such experiments can save enormous amounts of time and money. Imagine trying to find the optimal settings for producing electricity from coal without proper experimentation. Such a trial and error process would be extremely costly, wasteful and time consuming. In a similar vein, what if the fictional company in our previous example decided to invest a bunch of money in fine-tuning their training program based solely on the results of an observational study. In reality though, it turns out that adjusting their hiring process to identify more keen candidates would have been much more efficient and inexpensive.

3. In an experiment we can choose exactly those settings or **treatment levels** we are interested in, e.g. we can investigate the effect of different shift lengths (6, 8 or 9 hours) on employee productivity or test specific price points (R100, R150, R200) to determine which price maximizes sales or revenue. We can actively manipulate the variable(s) to the levels we are interested in.

Experimental studies and their design are fundamental to science, allowing us to further knowledge and test theories. So lets define them more rigorously. We'll start by introducing some terminology.

Establishing causality through observation is possible, but a bit more difficult.

Experiments are the most reliable way to establish causation because they involve direct manipulation of variables and control for other factors that might influence the outcome. By ensuring that differences in results are due to the specific factor being studied, experiments help avoid misleading conclusions caused by external influences or chance associations.
However, in some cases, causation can still be inferred from observational studies, especially when there is a well-understood relationship between cause and effect, consistent patterns across different settings, and no plausible alternative explanations. For example, the link between smoking and lung cancer was established through observational data, where researchers accounted for other possible influences and found strong, consistent evidence that smoking increases cancer risk. While experiments are preferred, careful analysis and logical reasoning can sometimes provide enough evidence for causal claims without direct intervention.

## Key points

1. Two ways of doing research: observation and expermentation.
2. Experimentation is the path to causality.
3. Experiments actively manipulate variables to isolate their effects on a response while controlling everything else.
4. We consider comparative experiments where the aim is to compare treatments.

# Chapter 2

# Terminology

## Treatment factors, treatment levels and treatments:

The **treatment factor** is the factor or variable that the experimenter actively manipulates to measure its effect on the response. All factors/variables that are investigated, controlled, manipulated, thought to influence the response, are called the treatment factors. They become the **explanatory variables** (mostly categorical) in the model. For each treatment factor, we actively choose a set of **levels**. For example, the treatment factor "temperature" can have levels 10, 20, and 50°C. If temperature is the only treatment factor in the experiment, the **treatments**[1] will also be 10, 20, and 50°C.

If we manipulate more than one factor (e.g., temperature and pressure), we have two treatment factors. When several treatment factors are manipulated, the experiment is called factorial and the **treatments** are all possible combinations of the factor levels. If we have pressure levels "low" and "high," there are 6 treatments in total:

[1] The terminology of treatments can be traced back to 1920's when it was first applied by Ronald Fisher in the agricultural sciences. He is often refered to as the Founder of Statistics! Have a look at the very first application of ANOVA here and also a nice article describing the history of statistics and his contribution to the field.

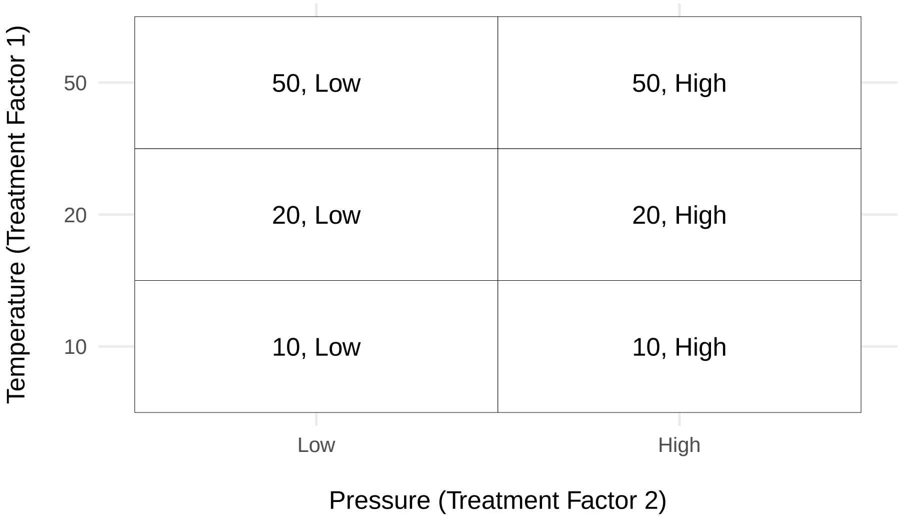

Temperature (Treatment Factor 1)
50
20
10
50, Low
50, High
20, Low
20, High
10, Low
10, High
Low
High
Pressure (Treatment Factor 2)

Figure 2.1: Visualization of how treatments are formed as combinations of treatment levels.

In the figure above, there are two treatment factors: Temperature (on the y-axis) and Pressure (on the x-axis). The axis ticks represent the levels of each treatment factor, and the blocks within the grid represent the treatments, which are specific combinations of the levels of Temperature and Pressure. Each treatment is labeled with the corresponding combination of levels (e.g., '50, Low' or '10, High').

Example 1

Three groups of students, 5 in each group, were receiving therapy for severe test anxiety. Group 1 received 5 hours, group 2 received 10 hours and group 3 received 15 hours. At the end of therapy each subject completed an evaluation of test anxiety. Did the amount of therapy have an effect on the level of test anxiety?
The three groups of students received the scores on the Test Anxiety index (TAI) at the end of treatment shown in the table below.

| Group 1 | Group 2 | Group 3 |
|---|---|---|
| 48 | 55 | 51 |
| 50 | 52 | 52 |
| 53 | 53 | 50 |
| 52 | 55 | 53 |

| | | |
|---|---|---|
| 50 | 53 | 50 |

When faced with a text like this, it is useful to identify the treatment factors, their levels and the treatments, as well the response. Clearly, from the question, we are interested in the effect of therapy on test anxiety. A statement like this can generally be read as the effect of the treatment factor on the response. Nowhere is another treatment factor mentioned, so we only have one in this example. What are the levels of therapy we set? The levels are 5, 10 and 15 hours of therapy and since we only have one factor these are also the treatments. Let's summarise this as follows:

- **Response:** Test Anxiety
- **Treatment Factor:** Therapy
- **Treatment Levels:** 5, 10, and 15 hours of therapy
- **Treatments:** 5, 10, and 15

## Experimental and observational unit

The **experimental unit** is the entity (e.g. material, object, or individual) to which a treatment is assigned or that receives the treatment. By contrast, the **observational unit** is the entity from which the response is recorded. This distinction is very important because it is the experimental units which determine how often the treatment has been replicated and therefore the precision with which we can measure the treatment effect. In the methods that we cover in this course, we require that in the end there is only one 'observation' (response value) per experimental unit. If several measurements have been taken on an experimental unit, we will combine these into one observation, typically by taking the mean. Very often, the experimental unit is also the observational unit.

What are the experimental units? To determine this, revisit the text of Example 1 and ask yourself: what entity received the treatments or to what were treatments applied? Most of you, will probably answer the students and this is correct. Each student received the respective treatment (number of hours in therapy) assigned to their group and so there are $5 \times 3 = 15$ experimental units.

There is an argument to be made that it is not clear whether the students received therapy on their own or that the groups of students received therapy together. In that case, treatments were applied to groups of students and so there would be three experimental units. This will usually be clear from the text, but we'll use this scenario to illustrate some concepts as we go.

We also need to know what the observational units are. The text states that at the end of therapy, each student completed an evaluation to determine their level of test anxiety. So the response, test anxiety, was measured on the student level which means students are the observational units. In the first scenario, the students are both the experimental units and observational units. But this would not be the case if groups are the experimental unit.

We also require that there is only one observation per experimental unit, the first scenario meets this requirement. For the second scenario, we have 5 observations per group and so we would have to take the mean of these values to end up with one response value per group.

Let's add to the summary assuming students are the experimental units:

- **Experimental unit (no):** Student (15)
- **Observational unit (no):** Student (15)

## Homogeneity of experimental units

When the set of experimental units are as similar as possible such that there are no distinguishable differences between them, they are said to be **homogeneous** (a fancy word for saying they are of the same kind). The more homogeneous the units are, the smaller the experimental error variance (natural variation between between observations of the same treatments) will be. It is super important to have fairly homogeneous units because it allows us to detect differences between treatments more easily.

## Blocking

If the experimental units are not fairly similar but are heterogeneous (the opposite of homogeneous), we can group them into sets of similar units. This process is called **blocking** and the groups are considered "blocks". We compare the treatments within each block as if each block is its own mini-experiment. This way we account for the differences between blocks and can better isolate the effect of the treatments.

> Example 2
>
> Imagine you're testing the effectiveness of two marketing strategies (A and B) to increase sales at a chain of coffee shops. The coffee shops are located in different neighborhoods, where factors like income levels might influence sales. To prevent these differences from skewing the results, you

group the coffee shops into "blocks" based on neighborhood characteristics such as income level (e.g., low, medium, high).
Within each block, you randomly assign coffee shops to either Strategy A or Strategy B. This approach allows you to compare the strategies while controlling for variability caused by differences in neighborhood features.
Without blocking, would you be able to confidently attribute differences in sales to the strategies alone? Likely not, as any observed differences could be due to neighborhood-specific factors rather than the strategies themselves.

## Replication and pseudoreplication

If a treatment is applied independently to more than one experimental unit it is said to be **replicated**. Treatments must be replicated! Making more than one observation on the same experimental unit is not replication, but *pseudoreplication*. Pseudoreplication is a common fallacy. The problem is that without true replication, we don't have an estimate of uncertainty, of how repeatable, or how variable the result is if the same treatment were to be applied repeatedly.

In Example 1, if experimental units were the groups and we didn't take the average of the observations per group, we would have pseudoreplication as each student would not be an independent replicate of a treatment - effectively, we have only applied each treatment once. You might notice that we then only have one true replicate per treatment group and this is problematic. To get an estimate of uncertainty, we would have to repeat this experiment a few more times to get more than one proper replicate.

The first scenario, however, did not have this problem and each treatment was replicated five times. After going through all this, we have the following summary:

- **Response:** Test Anxiety
- **Treatment Factor:** Therapy
- **Treatment Levels:** 5, 10, and 15 hours of therapy
- **Treatments:** 5, 10, and 15
- **Experimental unit (no):** Student (15)
- **Observational unit (no):** Student (15)
- **Replicates:** 5

Tip

Creating a summary like this, is a handy exercise for any experiment you come across, and we'll keep doing it for every experiment in this book. As we go along, we'll also add information about the type of experiment that was conducted.

# The three R's of experimental design

**Experimental Design** is a detailed procedure for grouping, if blocking is necessary, experimental units and for how treatments are assigned to the experimental units. There are three fundamental principles, known as the 'three R's of experimental design' which are at the core of a good experiment. The following section might feel a bit repetitive, but these concepts cannot be emphasised enough.

## Replication

Let's define it again: replication is when each treatment is applied to several experimental units. This ensures that the variation between two or more units receiving the same treatment can be estimated and valid comparisons can be made between treatments. In other words, replication allows us to separate variation due to differences between treatments from variation within treatments. For true replication, each treatment should be **independently** applied to several experimental units. If this is not the case, treatment effects become confounded with other factors.

Confounding means that is not possible to separate the effects of two (or more) factors on the response, i.e. it is not possible to say which of the two factors is responsible for any changes in the response. This is what happened in the Example 1 when groups are the experimental units. With only one replicate per treatment, the effect of therapy is confounded with the experimental unit or the effect of group on test anxiety. The reason why this is a problem is that any difference between the treatments could be due to any differences between the groups and not just the number of therapy hours. The same would be true if we only had one student per group. Why? Take a moment to think about this.

Consider the first row of the data from Example 1. It looks like the student in group 2 scored the highest, followed by group 3 and then group 1. So does longer therapy sessions lead to higher test anxiety? Likely not! With only one student per treatment, we are not able to say that any differences in the response are due to the treatments. It could be due to any differences between the individuals. Maybe the student in group 3 tends to score higher on anxiety tests regardless of the treatment, or perhaps the student in group 1 was unusually calm that day. Without replication, these individual differences could mask (or mimic) the true effects of the treatments.

By replicating the treatments across multiple students, we can average out these individual differences and gain a clearer picture of whether therapy duration truly impacts test anxiety. With five students per group, we might observe that group 1 consistently scores lower than group 3. This consistency would provide stronger evidence that the treatments, and not just individual variation, are responsible for the observed differences. So by replication, we can compare within treatment variation to variation between treatments.

| Treatment 1 | Treatment 2 | Treatment 3 |
|---|---|---|
| 48 | 55 | 51 |
| 50 | 52 | 52 |
| 53 | 53 | 50 |
| 52 | 55 | 53 |
| 50 | 53 | 50 |

## Randomisation

Randomisation refers to the process of randomly assigning treatments to experimental units such that each experimental unit has equal chance of receiving a specific treatment. Randomisation ensures that:

1. There is no bias on the part of the experimenter, either conscious or unconscious, when assigning treatments to experimental units.
2. No experimental unit is favored to receive a particular treatment.
3. Possible differences between units are equally distributed among treatments. If there are clear differences between units, then blocking should be performed and randomisation occurs within blocks. We'll talk more about this when we encounter Randomised Block Designs.
4. We can assume independence between observations.

Randomisation is not haphazard. In statistics (and here in the context of experimental design), randomisation has a specific meaning: namely that each experimental unit has the same chance of being allocated any of the treatments. This can be done using random number generators such as with software packages, dice or drawing number from a hat (provided the number have been shuffled adequately and have equal chance to be picked).

Let's have a look at randomisation in R. Suppose we have 4 treatments (`A`, `B`, `C`, and `D`) and 32 experimental units. There are no differences between the units, so we don't have to block, and we can equally split the units across the treatments, which means we have 8 units per treatment, i.e., 8 replicates. In R, we first create a long vector of 8 `A`s, 8 `B`s, 8 `C`s, and 8 `D`s called `all.treat`. Then shuffle the vector to obtain a randomisation using the function `sample`.

```
# repeat the vector A, B, C, D 8 times
all.treats <- rep(c("A","B","C","D"), times = 8)

# permutation of all.treats (sample withut replacement)
rand1 <- sample(all.treats)

# example output
rand1
```

```
 [1] "C" "D" "A" "B" "B" "C" "A" "B" "A" "D" "C" "C" "A" "D" "D" "C" "C" "B" "D"
[20] "C" "C" "B" "B" "A" "B" "D" "D" "B" "A" "A" "A" "D"
```

Experimental unit 1 recipes the first treatment that appears as the first element in the shuffled vector, experimental unit 2 receives the second and so on.

## Reduction of Unexplained Variation (Blocking)

Unexplained variation (or experimental error variance or within treatment variance) is largely due to inherent differences between experimental units. The larger this unexplained variation, the more difficult it becomes to detect treatment differences (a treatment signal). To minimise experimental error variance we can control extraneous factors (i.e. keeping all else constant) and by choosing homogeneous experimental units. Otherwise, we can block experimental units to reduce the variation.

Blocking variables are nuisance factors that might affect your response or introduce systematic variation in the response and we are typically, not interested in these. Often, they are factors that cannot be randomised, e.g. biological sex of a person, time of day, location of a warehouse etc. We control the effect of such variables on the response by blocking for them so that we can investigate the possible effect of a variable that we are interested in. Usually, in a complete block experiment, there are as many experimental units per block as there are treatments, so that each treatment is applied once in every block. Treatments are randomized to the experimental units in the blocks. We can then compare the effects of treatments on similar experimental units, and we can estimate the variation induced in the response due to the differences between blocks. This variation due to blocks can then be removed from the unexplained variation.

Blocking also offers the opportunity to test treatments over a wider range of conditions, e.g. if I only use people of one age in my experiment (say students)

I cannot generalize my results to older people. However, if i use different age blocks I will be able to tell whether the treatments have similar effects in all age groups or not.

Lastly, if blocking is not feasible, randomization will ensure that at least treatments and nuisance factors are not confounded.

> "Block what you can, randomize what you cannot."
>
> — Box, Hunter & Hunter (1978)

# Chapter 3

# Designing an Experiment

When planning an experiment we need to decide on:

- treatment factors and their levels
- the response
- experimental material / units
- blocking factors
- number of replicates

Some of these will be determined by the research question and how experimental units are assigned to treatments are determined by the design. The design that will be chosen for a particular experiment depends on the **treatment structure** (determined by the research question) and the **blocking structure** (determined by the available experimental units).

Here are two ways the treatments can be structured:

1. **Single factor**: the treatments are the levels of a single treatment factor.
2. **Factorial**: when more than one factor are of interest, then the experiment is said to be a factorial experiment. The treatments are constructed by crossing the treatment factors like we did in Figure 2.1 such that the treatments are all possible combinations of the treatment levels. For example, if factor A has $a$ levels and factor B has $b$ levels, there are $a \times b$ treatments. Such an experiment would then be called an $a \times b$ factorial experiment.

The blocking structure is determined the set of experimental units chosen or available for the experiment.are there any structures/differences that need to be blocked? Do I want to include experimental units of different types to make the results more general? How many experimental units are available in each block? For the simplest design in this course, the number of experimental units in each block corresponds to the number of treatments. This is called a complete block experiment. There are several other blocking structures, such as incomplete blocks and blocks with missing values, all with specific analysis which we will not cover here.

In this course, we cover two basic designs: Completely Randomized Designs (CRD) and Randomized Block Designs (RBD). For both designs, the treatment structure can be single or factorial. Where they differ is in terms of the experimental units and how randomization occurs.

### *Completely Randomized Designs (CRD)*

When all experimental units are fairly homogeneous, a CRD is used. Treatments are randomized to all experimental units.

### *Randomized Block Design*

This design is used when all experimental units are not homogeneous or blocking is required to control a nuisance factor. The treatments are randomized to the units within blocks.

# Part II

# Single Factor Completely Randomised Designs

# Chapter 4

# Introduction

Completely Randomized Designs (CRDs) are the simplest experimental designs. They are used when experimental units are uniform enough and we expect them to react similar to a given treatment. In other words, we have no reason to suspect that a group of experimental units might react differently to the treatments. We also don't expect any effects (besides possibly a treatment effect) to cause any systematic changes in the response. So, we don't have to block for differing experimental units or any nuisance factors.

Remember experimental design is the procedure for how experimental units are grouped and treatments are applied. We have already said that there are no blocks in CRDs. So randomisation occurs without restriction and to all experimental units. More generally, each of the $a$ treatments are randomly assigned to $r$ experimental units, such that each experimental unit is equally likely to receive any of the treatments. This means that there are $N = r \times a$ experimental units in total. We only consider designs that are *balanced* meaning that there an equal number of experimental units per treatment, i.e. a treatment is applied to $r$ units. The experiment is then said to have $r$ replicates.

The aim when analysing CRDs is to determine whether there is an effect of the treatment factor. We accomplish this by testing for differences in the treatment means (mean of response values in each treatment) through analyses different sources of variation in the response. This will become clear as we progress.

## 4.1 Example: The effect of social media multitasking on classroom performance.

As a student, I used to believe I could multitask effectively. I would scroll through my phone during lectures, study while texting friends, or listen to podcast while driving. It felt like I was paying attention to everything, but in hindsight, I can barely recall the details of those podcasts. I often had to revisit lectures or restart study sessions because my focus wasn't truly there. This tendency extends beyond student life. In the average workplace, tasks are frequently interrupted by social media, email checks, or notifications. Many of us feel the constant pull of our phones when trying to concentrate, whether we're working, studying, or even relaxing.

In an era of perceived multitasking, where devices and distractions dominate our attention, it's worth asking: Does social media multitasking impact academic performance of students?

Example 5.1

Two researchers from Turkey, Demirbilek and Talan (2018), conducted a study to try and answer this question. Specifically, they examined the impact of social media multitasking during live lectures on students' academic performance.
A total of 120 first-year undergraduate students from the same Turkish University were randomly assigned to one of three groups:

1. **Control Group:** Students used traditional pen-and-paper notetaking.
2. **Experimental Group 1 (Exp 1):** Students engaged in SMS texting during the lecture.
3. **Experimental Group 2 (Exp 2):** Students used Facebook during the lecture.

Over a three-week period, participants attended the same lectures on Microsoft Excel. To measure academic performance, a standardised test was administered.

**The analysis of experimental data is determined by the design.** This is the first thing we need to investigate. The design dictates the terms that we will include in our statistical model and so it is crucial to be able to identify the design and all factors included (blocking and treatment). It is also important to check that randomisation has been done correctly and determine the number of replicates used. In the previous chapter we started doing this by creating a summary of the design and we do the same here. From the description of the study, it is clear that:

- **Response Variable:** Academic performance, as measured by test scores.
- **Treatment Factor:** Level of social media multitasking.
- **Treatment Levels (Groups):** Control, Exp 1, and Exp 2.

Students were randomly assigned to one of the three groups, and performance was measured for each individual. Although this may seem obvious, they only took one measurement per student, so we don't have to worry about pseudoreplication. This setup indicates that the students are both the experimental units and the observational units in this study. With a total of 120 experimental units and three treatments, the experiment has 40 replicates. Since only one treatment factor was investigated, and no blocking was performed, this is classified as a **single-factor Completely Randomized Design (CRD).** Here is the study breakdown:

- **Response Variable:** Academic Performance
- **Treatment Factor:** Level of Social Media Multitasking
- **Treatment Levels:** Control, Experimental 1 (SMS), Experimental 2 (Facebook)
- **Treatments:** Control, Experiment 1, Experiment 2
- **Experimental Unit:** Student (120)
- **Observational Unit:** Student (120)
- **Replicates:** 40 students per group
- **Design Type:** Single-Factor Completely Randomized Design (CRD)

Before we continue, now is the time to note that we won't be using the real data collected in this experiment. It wasn't available but I have simulated data to match their results. I've also made some other modifications such as the original study included 122 students but to ensure a balanced design I include only 120.

## 4.2 Exploratory data analysis (EDA)

Before we start any analyses, we have to conduct some exploratory data analysis to get a feel for our data. We start by checking whether it has been read in correctly and then look at some descriptive statistics.

In R, we read in the data set and then use some commands to inspect the data set:

```
multitask <- read.csv("Datasets/multitask_performance.csv")
nrow(multitask) # check number of rows
```

```
[1] 120
```

```
head(multitask) # check first 5 rows
```

```
    Group Posttest
1    Exp1 86.39427
2    Exp1 64.19996
3    Exp2 52.75394
4 Control 67.81147
```

```
5    Exp1 52.39911
6    Exp1 56.58150
```

```r
tail(multitask) # check last 5 rows
```

```
      Group Posttest
115 Control 77.94344
116 Control 63.58444
117    Exp1 55.17758
118    Exp2 67.16150
119    Exp2 32.58373
120    Exp2 49.58119
```

```r
summary(multitask)
```

```
    Group              Posttest    
 Length:120         Min.   :23.38  
 Class :character   1st Qu.:52.67  
 Mode  :character   Median :65.01  
                    Mean   :63.59  
                    3rd Qu.:76.32  
                    Max.   :98.78  
```

The data set consists of 120 rows (each row representing a student) and two columns (`Group` and `Posttest`). The first column, `Groups`, contains the treatment the student was assigned and the `Posttest` column contains the response measure. Using the functions `head` and `tail`, we can look at the first and last 5 rows and the function `summary` provides us with a description of each column. We do this to check that R has read in our data correctly (you can view the whole data set by running `view(multitask)` as well). The summary tells us that the `Group` column is of the class "character". For our analysis, we want it to be read as a factor:

```r
multitask$Group <- as.factor(multitask$Group)
summary(multitask)
```

```
     Group       Posttest    
 Control:40   Min.   :23.38  
 Exp1   :40   1st Qu.:52.67  
 Exp2   :40   Median :65.01  
              Mean   :63.59  
              3rd Qu.:76.32  
              Max.   :98.78  
```

Now, we can see that there are 40 replicates per treatment group, confirming that the experiment is balanced. I have assumed that, based on the results shown, that the `Posttest` scores were recorded as percentages and using the summary we can quickly check whether there are any observations that are not on the appropriate scale or might be outliers. Looks good so far!

## 4.3 Checking assumptions

Demirbilek and Talan (2018) had several research questions, but here we only consider the following:

Are there any differences in mean academic performance between the three groups?

You might think that we could perform three t-tests (Control vs Exp 1, Control vs Exp 3, Exp 1 vs Exp 2). We could, but the problem with this approach is what we call multiple testing. When conducting many tests, there is an increased risk of making a Type 1 Error (rejecting the null hypothesis when it is in fact true) [1].

When we have more than two groups, we can use a one-way analysis of variance (ANOVA) which can be seen as an extension of a *t*-test and is called "one-way" because there is a single factor being considered. In the next section, we will see that ANOVA is a linear model and some of the assumptions are about the model errors (just like regression):

1. There are no outliers.
2. The errors are independent.
3. The errors are normally distributed.
4. All groups have equal population variances.

We need to check the validity of these assumptions. There are both formal and informal techniques. Formal techniques (i.e. hypothesis tests) are not always appropriate for several reasons such as small data sets or that testing one assumption usually requires that the other two hold, complicating the order of tests. Informal techniques are more than sufficient and in this course, we stick with them.

### Outliers

Outliers are unusual observations (response values) that deviate substantially from the remaining data points. They can have a large influence on the estimates of our model. Think of statistics such as means and variances, outlying observations will shift the mean towards them and distort the variability of the data.

If we're lucky, outliers are artefacts of data recording or entering issues, such as a missing decimal points or incorrect scaling (called error outliers). These types of outliers can be corrected and the analysis can be done as usual. If,

[1] Can't remember what a *t*-test is and/or need a refresher on hypothesis testing? Have a look this video on t-tests and document for a brief reminder. **Also, a quick (and cool) sidenote:** This study by Chen et al. (2024) used a Completely Randomized Design (CRD), randomly assigning undergraduate students to playback speed groups (1x, 1.5x, 2x, and 2.5x) to measure the effect on comprehension of recorded lectures. Using ANOVA they found that comprehension was preserved up to 2x speed. I personally like to increase the playback speed to 1.5px if I just need to revise something quickly.

however, there are freak observations that are not clearly due to anything like data inputting, then they are likely genuine unusual responses (called interesting outliers) and should not be discarded. There are many ways of identifying and dealing with outliers (Aguinis, Gottfredson, and Joo (2013) found 29 different ways in the literature). Here, it is recommended that the analysis should be run with and without the outliers to see whether the conclusion depends on their inclusion. When dealing with outliers, it is best to be transparent and clear about how they were handled. Simply removing outliers with no explanation is questionable research practice.

A good way to check for outliers, is to inspect the data visually with a box-plot of your data grouped by treatment.

```
boxplot(Posttest ~ Group, data = multitask, col = c("skyblue", "lightgreen", "pink"),
        main = "Posttest Scores by Group",
        xlab = "Group",
        ylab = "Posttest Scores")

stripchart(Posttest~Group, data = multitask, vertical = TRUE, add = TRUE, method = "ji
```

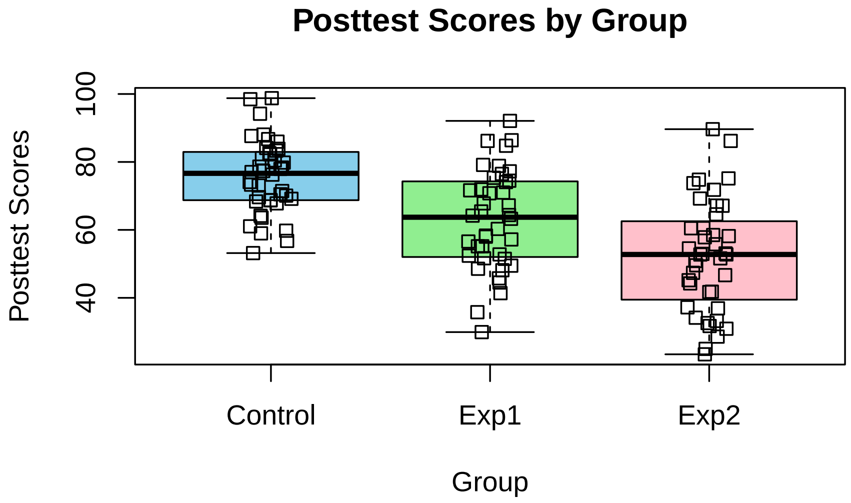


Figure 4.1: Box-plots of Post treatment scores by group.

The first line of code plots the box-plot and by inputting `Posttest~Groups` as the first argument we are say plot the values of `Posttest` by `Groups`. There are extra graphical parameters specified to make the plot look a bit nicer. The function `stripchart` is used to overlay the data points. Based on these plots, there aren't any obvious outlying observations.

## Equal population variance

The model assumes that population variances in different levels of the treatment factor are equal. That is, it is assumed in ANOVA that the variance of the response within each treatment is a separate estimate of the same population variance.

Since we only have sample data, we would not expect that the sample variances to be exactly the same. If they are different it does not mean the assumption is not met. We expect them to differ a bit due to chance simply because we are sampling. Every time we sample from a population, the data set will be different and so will it's variability. The sample variances need to be similar enough so that our assumption of equal population variance is reasonable.

To check this assumption, we can inspect the box-plots again and compare the heights. More specifically, we look at the interquartile ranges (IQR). From looking at the plot, the IQRs do not vary widely. If you prefer to look at the actual values, we can use R to obtain them:

```r
sort(tapply(multitask$Posttest,multitask$Group,IQR))
```

```
 Control     Exp2     Exp1 
14.01068 20.94529 21.97001 
```

Another measure of variability we can look at, are the standard deviations (sd's). With the same line of code but just replacing the function we want to apply, we obtain the sd of each group:

```r
sort(tapply(multitask$Posttest,multitask$Group,sd))
```

```
 Control     Exp1     Exp2 
10.82887 14.60601 16.42678 
```

The rule of thumb is to use the ratio of the smallest to largest standard deviation and check whether it is smaller than five. In our case, the smallest sd (of the Control group) is about 1.5 times smaller than the largest sd (of the Exp 2 group) which is acceptable.

## Normally distributed errors

We can check this assumption by looking at the residuals after model fitting. A common misconception is to think that the response needs to be normally distributed. However, it is only the unexplained variation, i.e. the errors or residuals (estimates of errors), that we assume to be normally distributed. Of course, if the response has a clearly non-normal distribution (e.g. Binomial), then the residuals are likely to be non-normal as well. So, we can check our response values before hand for obvious deviation from normality, but we have to check this assumption again after fitting our model. Things to look for are asymmetric box-plots which indicate skew distributions. We also want to check that the data points tend to cluster around the median. In Figure 4.1, there are no signs of any clear deviation from normality. Other graphs we could look at are histograms or Quantile-Quantile (Q-Q) plots. Q-Q plots show the theoretical quantiles of the standard normal distribution against the actual quantiles of our data. We want our data to be as close to the xy line as possible (deviations in the tails are expected).

```
par(mfrow = c(1,3))

# First we subset the data for each group
control <- multitask$Posttest[multitask$Group == "Control"]
exp1 <- multitask$Posttest[multitask$Group == "Exp1"]
exp2 <- multitask$Posttest[multitask$Group == "Exp2"]


qqnorm(control, pty = 4, col ="blue", main = "Control")
qqline(control, col = "red")

qqnorm(exp1, pty = 4, col ="blue", main = "Exp 1")
qqline(exp1, col = "red")

qqnorm(exp2, pty = 4, col ="blue", main = "Exp 2")
qqline(exp2, col = "red")
```

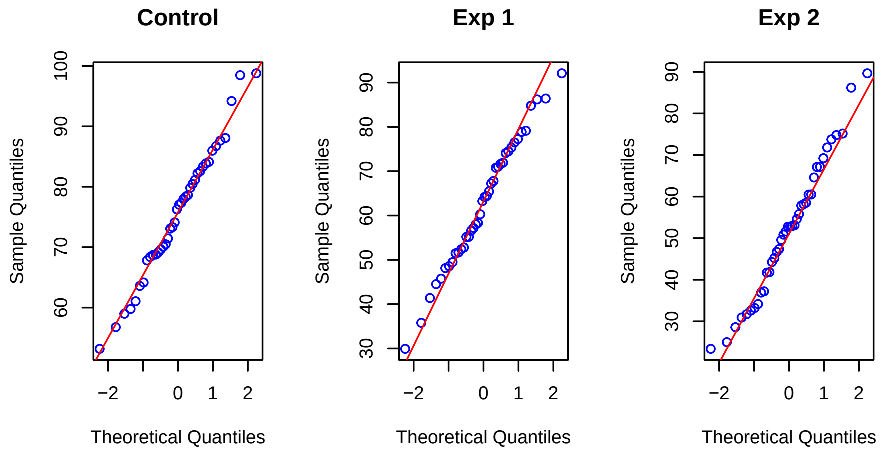


Figure 4.2: Q-Q plots of response per treatment group.

The `qqnorm` function plots the theoretical quantiles on the x-axis and the sample quantile son the y-axis. So each point on the plot corresponds to a quantile from the sample plotted against the expected quantile from the standard normal distribution. As a reference we add a straight 45-degree line (in red) using the `qqline` function to indicate what perfect normality would look like.

## Independent errors

The assumption is that the **errors** are independent. While we can check for certain types of dependence in the residuals after fitting the ANOVA (as we will see later), dependence among observations generally results in dependent residuals. Therefore, before fitting any models, we examine the observations and the experimental design to identify potential violations of independence.

In statistics, if one observation influences another in some way or another, they are said to be dependent. For the type of data considered here, there are two types of independence we require. Firstly, observations within treatments should be independent and second, observations between samples should be independent. Another way of saying this, is **there should be independence within and among treatments.** Depending on the direction of any violations, the within treatment variance or among treatment variance can either be deflated or inflated and treatment effects can be biased. This has considerable impact on the test statistic (F-ratio for ANOVA, more on this later) which could lead to misleading results. [2]

Violations of independence typically occur when the experimental units within or among treatments are connected in some way. Dependence within a sample can occurs when they are taken in a non-random sequence. Doing so typically allows some other variable to introduce dependence between successive observations. For example, measurement drift (when a tool's reading gradually changes over time), physical effects (e.g. temperature) of the location of experimental units or the experimenter might become better (or worse) at taking the measurement as they move along. If these variables are not taken into account (by including them as factors in the model), it leads to a lack of independence in the errors of our model. Specifically, they lead to auto-correlated residuals; observations made closer together in time or space are more similar to each other than expected (this is what we check after model fitting).

An informal check we could do, is to plot the data in the order in which they were collected (if this information is available) whether that is temporally or spatially to see if any patterns emerge. To do this in R, we can create a Cleveland dot plot.

```r
dotchart(multitask$Posttest, ylab = "Order of observation", xlab ="Post treatment test score")
```

[2] Underwood (1996) has a very detailed explanation of the independence assumption (and the others) in the context of ANOVA. The book is for ecological experiments, but much of it pertains to all types of experiments.

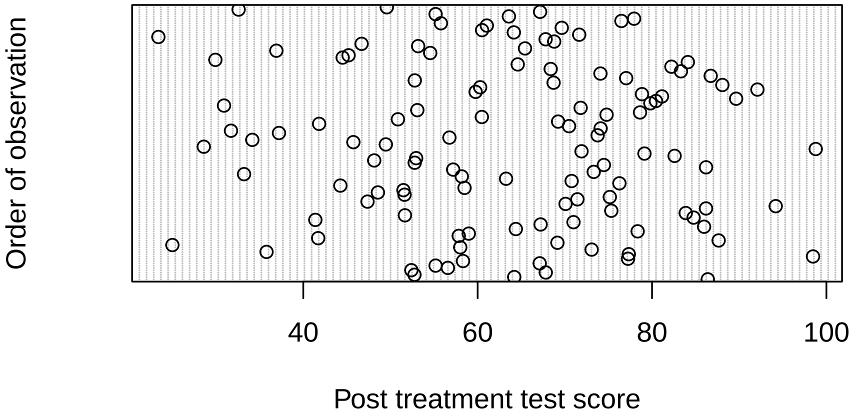


Figure 4.3: Cleveland dot chart of response values in the order in which they appear in the data set.

We have assumed that the order in which the observations appear in the data set are the order in which they were recorded. If there were any factors that caused systematic trends, (i.e. dependence) in the observations, then there would be some kind of pattern in the dot chart. For our example, there is no clear pattern. After fitting the model, we can also plot the residuals against spatial coordinate or against order to check for obvious patterns. This method, however, only detects violations of independence if observations are related to time or space.

Dependence between treatments can occur if we apply the treatments to the same group of experimental units or if experimental units from different treatments are able to interact in some way during the experiment. These types of violations including those mentioned above, are ones that we can mostly prevent or control by properly designing the experiment. When we control for factors that might induce dependence, we can include them in our model.

Other reasons for dependence may not be as obvious or easy to eliminate as we will see below. In the end, they may not have a strong impact on our estimates but it is important to carefully scrutinize your design and the system you are studying to identify possible sources of dependence so that these can be addressed and dealt with properly.

In our example, within and among group dependence could be caused by the students interacting or influencing each other in some way (by sharing notes for example). During the lectures, this can be controlled by careful monitoring and randomising their position in the lecture theater, but outside of lectures, it is less easy to control. Here we can argue that if students interacted outside of lectures the impact on their academic performance (as measured by the test) would likely be negligible. The integrity of the students is at play. It is not really possible to diagnose this type of dependence after the fact, only with careful design and implementation can these be avoided.

**It is the onus of the experimenter to design and conduct experiments that ensure independence.** With more thought (and if we're lucky, funding) all well-designed experiments should lead to independent data. If violations are found after the fact, they cannot typically be corrected and then methods that deal specifically with dependent data (if appropriate) should be used[3].

## A quick note on the robustness of ANOVA

A statistical procedure is said to be robust to departures from a model assumption if the results remain unbiased even when the assumption is not met. The robustness of ANOVA is as follows:

1. The assumption of normality is not super crucial. Only severe departures from normality such as long-tailed distributions or skewed distributions when sample sizes are unequal and/or small are particularly problematic.

2. Independence within and among groups is extremely important. ANOVA does not handle dependent data and other analyses should be attempted if there is dependence.

3. ANOVA is relatively robust to violations of the equal variance assumption as long as there are no outliers, sample sizes are large and fairly equal (in the case of unbalanced designs which we do not cover here), and the sample variances are relatively equal.

4. ANOVA is not very resistant to severely outlying observations either.

Note

1. In this course, you will always encounter data that has already been collected and the description of the experiment will likely not be very exhaustive. You might be task then with thinking about how the assumption of independence could have been violated, but for the most part we will assume the data are independent, both within and among samples (unless otherwise stated or you are asked if the assumption holds).
2. No real data set ever meets all assumptions of a model perfectly. As the famous (at least in the world of statistics) quote by George

[3] A few of these methods are repeated measures ANOVA, mixed-models or hierarchical models.

> Box goes: “All models are wrong but some are useful.” Judging whether a particular data set meets our assumptions reasonably well is therefore a bit of an art. You will likely read and hear that being able to identify violations comes from **experience**. The best way to get experience is to look at lots of data sets where you know how well they meet the assumptions. That’s best done via simulation. We therefore encourage you to use the attached R code to simulate data where various assumptions are violated. Run the code a number of times to get a feeling for how variable your actual sample can be even if the data generating mechanism doesn’t change. You may also want to play around with the sample sizes and you can change the degree to which the assumptions are violated to get a feeling for how these violations show up in the plots.

## 4.4 Summary

Completely Randomized Designs (**CRDs**) are the simplest experimental designs, used when experimental units are **uniform** and expected to react similarly to treatments. Since no nuisance factors are controlled, randomization occurs **without restriction**, and treatments are **evenly assigned** across experimental units (**balanced design**).

The social media multitasking study served as an example, where 120 students were randomly assigned to three groups (Control, SMS, Facebook) to measure their academic performance. This setup represents a single-factor CRD, where students are both the experimental and observational units with 40 replicates per group.

Before conducting ANOVA, we:

- Checked the data set for correct structure (120 observations, treatment groups as factors).
- Inspected summary statistics and visualized distributions (box-plots, histograms, Q-Q plots).

For ANOVA, the following assumptions were examined:

1. **Outliers**: Check via box-plots.
2. **Equal variance**: Assess using interquartile ranges and ratio of sample standard deviations.
3. **Normality of errors**: Verified using Q-Q plots.
4. **Independence within and between treatment groups**: Considered through study design.

Proper experimental design ensures valid conclusions. Identifying violations of assumptions early helps prevent biased results.

# Chapter 5

# A Simple Model for a CRD

To analyse data collected from a Completely Randomised Design we could use $t$-tests and compare the samples two at a time. This approach is problematic for two reasons. Firstly, the test statistic of a $t$-test is calculated with a standard deviation based only on the two samples it considers. We want our test statistic to consider the variability in all samples collected. Second, when we conduct multiple tests the overall Type 1 Error rate increases. That is, when doing many tests, the chance of making *at least one wrong conclusion* increases with the number of tests (if you want to know more see the box below). To avoid this, we will use the ANOVA method which was specifically developed for comparing multiple means.

Multiple Testing / Comparisons

When we conduct a test, there is always a possibility that a significant result is due to chance and not actually a real difference. In first year, you were taught the Neyman-Pearson approach to hypothesis testing, which entails setting a significance level ($\alpha$) for the test you will conduct. This significance level is the Type 1 error rate (probability of falsely rejecting $H_0$). A common $\alpha$ is 0.05, meaning that 5% of the time we will reject the null hypothesis even if it is true. That means when we find a significant result, one of two things have happened:

1. Either we genuinely found a significant result or,
2. We were that unlucky, that our result is one of those 5% cases.

We will never know, this is the basis of statistical testing. We accept that we cannot tell which of our conclusions are Type 1 Errors. When we conduct many tests, the overall Type 1 Error rate increases. That is the overall chance of *at least one wrong conclusion* increases with the number of tests conducted. This is not good! We already might be wrong 5% and

> we don't want to increase that risk even further when conducting multiple tests.

## 5.1 The model

When we collect samples, we usually want to learn something about the populations from which they were drawn. To do this, we can develop a model for the observations that reflects the different sources of variation believed to be at play.

For Completely Randomised Designs, we have $a$ treatments which implies $a$ population means $\mu_1, \mu_2, \mu_3, \dots, \mu_a$. We are interested in modelling the means of the treatments and the differences between them. Ultimately we want to test whether they are equal which we'll get to in the next section. First, we construct a simple model for each observation $Y_{ij}$:

$$Y_{ij} = \mu_i + e_{ij},$$

where

$$\begin{aligned}
i &= 1, \dots, a \quad (a = \text{number of treatments}) \\
j &= 1, \dots, r \quad (r = \text{number of replicates}) \\
Y_{ij} &= \text{observation of the } j^{th} \text{ unit receiving treatment } i \\
\mu_i &= \text{mean of treatment } i \\
e_{ij} &= \text{random error with } e_{ij} \sim N(0, \sigma^2)
\end{aligned}$$

That is, each observation is modeled as the sum of its population mean and some random variation, $e_{ij}$. This random variation represents unexplained differences between individual observations within the same group and we assume that these differences follow a normal distribution with mean 0 and constant variance across all treatment groups. [1]

We can change the notation slightly by arbitrarily dividing each mean into a sum of two components: the overall mean $\mu$ (the mean of the entire data set, which is the same as the mean of the $a$ means[2]) and the difference between the population mean and the overall mean. In symbols, this translates to:

[1] As opposed to non-constant variance across all treatment groups: $e_{ij} \sim N(0, \sigma_i^2)$ where the $\sigma_i^2$'s are different.

[2] $\mu = \frac{\sum \mu_i}{a}$

$$\begin{aligned}
\mu_1 &= \mu + (\mu_1 - \mu) \\
\mu_2 &= \mu + (\mu_2 - \mu) \\
&\vdots \\
\mu_a &= \mu + (\mu_a - \mu)
\end{aligned}$$

The difference $(\mu_i - \mu)$ is the **effect of treatment** $i$, denoted by $A_i$. So each population mean is the sum of the overall mean and the part that we attribute to the particular treatment $(A_i)$:

$$\mu_i = \mu + A_i, \quad i = 1, 2, \ldots, a,$$

where $\sum A_i = 0$.

Why the $\sum A_i = 0$ constraint?

This constraint ensures that the treatment effects are expressed as deviations from the overall mean. To see why this holds, take the sum of both sides of the equation:

$$\sum_{i=1}^{a} \mu_i = \sum_{i=1}^{a} (\mu + A_i).$$

Expanding the right-hand side:

$$\sum_{i=1}^{a} \mu_i = a\mu + \sum_{i=1}^{a} A_i.$$

By definition, the overall mean $\mu$ is the mean of the treatment means:

$$\mu = \frac{1}{a} \sum_{i=1}^{a} \mu_i.$$

Multiplying both sides by $a$ gives:

$$\sum_{i=1}^{a} \mu_i = a\mu.$$

Comparing this with our earlier equation:

$$a\mu = a\mu + \sum_{i=1}^{a} A_i.$$

Subtracting $a\mu$ from both sides, we get:

$$\sum_{i=1}^{a} A_i = 0.$$

This constraint is standard in ANOVA models to ensure that the treatment effects are relative to the overall mean rather than being arbitrarily defined. It is not an additional assumption; any $a$ means can be written in this way.

Replacing $\mu_i$ in the model above leads to the common parameterisation of a single-factor ANOVA model[3]:

$$Y_{ij} = \mu + A_i + e_{ij}$$

where

$$\begin{aligned}
i &= 1, \dots, a \quad (a = \text{number of treatments}) \\
j &= 1, \dots, r \quad (r = \text{number of replicates}) \\
Y_{ij} &= \text{observation of the } j^{th} \text{ unit receiving treatment } i \\
\mu &= \text{overall or general mean} \\
A_i &= \text{effect of the } i^{th} \text{ level of treatment factor A} \\
e_{ij} &= \text{random error with } e_{ij} \sim N(0, \sigma^2)
\end{aligned}$$

Comparison to regression

If you wanted to, you could rewrite this with the regression notation you've encountered before as a regression model with a single categorical explanatory variable:

$$Y_i = \beta_0 + \beta_1 T2_i + \beta_2 T3_i + e_i$$

where $T2$ and $T3$ are indicator variables (i.e. $T2 = 1$ if observation $i$ is from treatment 2 and 0 otherwise). The intercept estimates the mean of the baseline category, here it is $T1$.

These two models are equivalent. The data are exactly the same: in both situations we have $a$ groups and we are interested in the mean response of these groups and the difference between them. The model notation is just slightly different. In the ANOVA model we use $\mu$ and $A_i$ instead of $\beta_0$ and $\beta_i$ which have different meanings.

| Regression | ANOVA |
|---|---|
| $\beta_0$ is the mean of the baseline category | $\mu$ is the overall mean |

[3] Often called **Model I**.

| | |
|---|---|
| $\beta_1$ is the difference between the means of category 2 and the baseline category. | $A_i$ is the effect of treatment $i$, i.e. change in mean response relative to the overall mean. |

When all the explanatory variables are categorical, which is mostly the case in comparative experimental data, it is more convenient to write the model in the ANOVA form, for two reasons:

1. The $A_i$ notation is more concise, because we don't have to add all the dummy variables. This makes it easier to read and understand because there is only one term per factor.
2. Mathematically it is more convenient. In this format all terms are deviations from a mean. This leads directly to sums of squares[4] (squared deviations from a mean) and analysis of variance. We will see later that we can partition the total sum of squares into one part for every factor in the model. This allows us to investigate the variability in the response contributed by every model term (or factor).

The model can be interpreted as follows:

Each observation, $Y_{ij}$, is the sum of the overall mean ($\mu$), plus the effect of the treatment it belongs to ($A_i$), and some random error ($e_{ij}$). We use two subscripts on the $Y$. One to identify the group (treatment) and the other to identify the subject (experimental unit) within the group:

$$
\begin{aligned}
Y_{1j} &= \mu + A_1 + e_{1j} \\
Y_{2j} &= \mu + A_2 + e_{2j} \\
Y_{3j} &= \mu + A_3 + e_{3j} \\
&\vdots \\
Y_{aj} &= \mu + A_a + e_{aj}
\end{aligned}
$$

## 5.2 Estimation

Okay, so we have a model which we now need to **fit to our data**. When we do this, we estimate the model parameters using our data. The parameters we want to estimate are $\mu$ (the overall mean), the treatment effects ($A_i$) and $\sigma^2$ (the error variance). As for regression, we find **least squares estimates** for the parameters which minimise the residual or error sum of squares[5]:

[4] In statistics, sums of squares is a measure of variability and refers to squared deviations from a mean or expected value. For example, the residual sums of squares (sum of squared deviations of the observations from the fitted values).

[5] error = observed - fitted.

$$\text{SSE} = \sum_i \sum_j e_{ij}^2 = \sum_i \sum_j (Y_{ij} - \hat{Y}_{ij})^2 = \sum_i \sum_j (Y_{ij} - \mu - A_i)^2$$

It turns out when we solve for the estimates that minimise the SSE[6], we obtain the following estimators:

$$\hat{\mu} = \bar{Y}_{..}$$
$$\hat{\mu}_i = \bar{Y}_{i.}$$

and

$$\hat{A}_i = \bar{Y}_{i.} - \bar{Y}_{..}$$

From linear model theory we know that the above are unbiased estimates[7] of $\mu$ and the $A_i$'s. What does this tell you? It tells you that we can use the sample means as estimates for the true means. The estimated mean response for treatment $i$ is the observed sample mean of treatment $i$ and the observed overall mean is the estimated grand mean.

For the last parameter, the error variance, an unbiased estimator is found by dividing the minimised SSE (i.e. calculated with the least squares estimates) by its degrees of freedom:

$$s^2 = \frac{1}{N-a} \sum_{ij} (Y_{ij} - \bar{Y}_{i.})^2$$

This quantity is called the Mean Squares for Error (MSE) or residual mean square. It has $(N - a)$ degrees of freedom since we have $N$ observations and have estimated $a$ means. If you look at the formula you'll notice that it is an average of the observed variability from the different treatment groups.

Compare this with regression

Compare this with the equations you saw in the regression section. Barring the extra subscript, the only difference is the equation for calculating the fitted/predicted value.
In regression, the fitted value is:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

and here it is:

[6] Another name for this is the residual sums of squares (RSS).

[7] Unbiased means that the expected value of these statistics equals the parameter being estimated. In other words, the statistic equals the true parameter on average.

$$\hat{Y}_{ij} = \bar{Y}_{i.} = \hat{\mu} + \hat{A}_i$$

## 5.3 In context of the social media multitasking example

Let's take what we've learned so far and apply it to our example. We had $a = 3$ treatments each with $r = 40$ replicates. The model equation is:

$$Y_{ij} = \mu + A_i + e_{ij}$$

where

$$\begin{aligned} i &= 1, \dots, 3 \\ j &= 1, \dots, 40 \end{aligned}$$

If we write the model out for each treatment, we get:

$$\begin{aligned} Y_{Cj} &= \mu + A_C + e_{Cj} \\ Y_{E1j} &= \mu + A_{E1} + e_{E1j} \\ Y_{E2j} &= \mu + A_{E2} + e_{E2j} \end{aligned}$$

and when we fit the model to the data, the predicted means for the treatments are:

$$\begin{aligned} \hat{Y}_C &= \hat{\mu} + \hat{A}_C = \bar{Y}_{C.} \\ \hat{Y}_{E1} &= \hat{\mu} + \hat{A}_{E1} = \bar{Y}_{E1.} \\ \hat{Y}_{E2} &= \hat{\mu} + \hat{A}_{E2} = \bar{Y}_{E2.} \end{aligned}$$

To fit this model in R, we use the `aov` function and then use another function to extract the estimated parameters. By specifying type = "effects", the function returns the $\hat{A}_i$'s

This tells us that the average score for students in the control group is roughly 12% higher than the overall average[8]. Both experimental groups performed worse, with students in the second group scoring, on average, about 11% less than the mean across all groups. We can also extract the overall mean and the treatment means by specifying type = "means":

[8] Remember: $\mu_i = \mu + A_i$

The grand mean (i.e. average of all test scores) was 64% in this experiment. The control group scored on average 76% which is 12% higher than the overall mean and so on. So we have the estimates for the effects, grand mean and treatment means.

The last parameter we need to estimate is the error variance $\sigma^2$. Have a look at the formula again:

$$s^2 = \frac{1}{N-a}\sum_{ij}(Y_{ij} - \bar{Y}_{i.})^2$$

If we focus on the sum and break into sums of squares for each treatment $i$, we get for the first treatment (let's say that is the control group):

$$\sum_{j}(Y_{1j} - \bar{Y}_{1.})^2$$

Which is the sum of the squared differences between the observations in the control group and the mean score of the control group. We can easily calculated that in R:

First, we subset the data set for the scores in the control group. Then we find the mean and calculate the squared differences, which is all summed together to give the sums of squares for treatment group 1. We can repeat this for the remaining treatments and sum the three sum of squares together and divide by $N - a$ to get the MSE.

Later we will see that we can extract this quantity easily from the ANOVA table. But for now, this is a useful exercise to make sure you understand the formula. So, $\widehat{\sigma^2} = s^2 = 200$ (rounded off to the nearest integer) and $\hat{\sigma} = s = 14$. This is the estimate of variance we will use to conduct an hypothesis to determine if there are any difference in the treatment means. Now you can see that it takes into account the variability of all our samples.

## 5.4 Standard errors and confidence intervals

In the previous section we saw how the parameters of the ANOVA model are estimated. We also need a measure of uncertainty for each of these estimates (in the form of a standard error, variance, or confidence interval). Let's start with the variance of a treatment mean estimate:

$$Var(\mu_i) = \frac{\sigma^2}{n_i}$$

Remember that the sampling distribution of the mean is $N(\mu, \frac{\sigma^2}{n})$ and here we assumed that the groups have equal population variances.

**Variance, Standard Deviation and Standard Error: what's all this again?** The variance (Var) is a good way of measuring variability. The Standard Deviation (SD) is the square root of the variance of a sample or population. The Standard Error (SE) is the SD of an estimate (read that again).

If we assume that two treatment means are independent, the variance of the difference between two means is:

$$Var(\hat{\mu}_i - \hat{\mu}_j) = Var(\hat{\mu}_i) + Var(\hat{\mu}_j) = \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_j}$$

To estimate these variances we substitute the MSE for $\sigma^2$ as it is an unbiased estimate of the error variance (the variability within each group). The standard errors of the estimates are found by taking the square root of the variances. The standard error is the standard deviation of an estimated quantity, and is a measure of its precision (uncertainty); how much it would vary in repeated sampling.

We can assume normal distributions for our estimates because we have assumed a normal linear model and because they are means (or differences between means). This means that confidence intervals for the population treatment means are of the form:

$$\text{estimate} \pm t_v^{\alpha/2} \times \text{SE(estimate)}$$

where $t_v^{\alpha/2}$ is the $\alpha/2^{th}$ percentile of the Student's $t$ distribution with $v$ degrees of freedom. The degrees of freedom are the error degrees of freedom, $N - a$ for CRD.

What are the standard errors associated with the parameter estimates in the social media example? We can easily extract this by specifying an extra argument to the `model.tables` function.

Standard error of the effects:

and for the treatment means:

So, now we have parameter estimates and their standard errors. Equipped with these, we are closer to answering the original question: Does social media multitasking impact academic performance of students? Based on the model we fitted and the parameters we estimated, how do we test this? The answer is with an ANOVA table.

## 5.5 Summary

This chapter introduces the Completely Randomized Design (CRD) model and explains why ANOVA is preferred over multiple t-tests, which inflate the Type 1 Error rate.

In ANOVA, each observation is modeled as:

$$Y_{ij} = \mu + A_i + e_{ij}$$

where $\mu$ is the overall mean, $A_i$ is the treatment effect (difference between treatment mean $\mu_i$ and the overall mean), and $e_{ij}$ is random error which normally distributed with mean 0 and variance ($\sigma^2$).

Parameters are estimated using least squares, with the mean squares error (MSE) providing an estimate of variance ($\sigma^2$).

Applying ANOVA to the social media multitasking study, we estimated treatment means and effects together with their standard errors, setting the stage for hypothesis testing using an ANOVA table.

# Chapter 6

# Analysis of Variance

The ANOVA model we have introduced is identical to a regression model with categorical variables, it is just parameterised differently. So why the different names and emphasis on variance - **AN**alysis **O**f **VA**riance? A well designed experiment allows us to estimate the within-treatment variability and between treatment variability. More specifically, it enables the partitioning of the total sum of squares into independent parts, one for each factor in the model (treatment and/or blocking factors). This allows us unambiguously to estimate the variability in the response contributed by each factor and the experimental error variance! We can then use this partitioning to perform hypothesis tests. In other words: by looking at the variation we can find out if the response differs due to the treatments.

An ANOVA applied to a single factor CRD is called a one-way ANOVA or between-subjects ANOVA or an independent factor ANOVA. It is a generalization of the 'two-sample t-test assuming equal variances' to the case of more than two populations.

## 6.1 An Intuitive Explanation

Before we consider real data, we first want to look at a constructed example to explain the main ideas behind ANOVA. Assume that we carried out two experiments on plants removing nitrate ($NO_3$) from storm water. In both experiments, we consider three plant species (un-creatively called 'A', 'B', and 'C'). In both experiments, we have three replicates per treatment. We are only interested in comparing the species so there is no control treatment. We obtained the following data:

If you look at these data sets carefully, you will see that each of the three species had the same mean in the two experiments. However, the measurements were much more variable in Experiment 2 than in Experiment 1.

Table 6.1: Hypothetical Experiment

(a) Experiment 1

| Species | A | B | C |
|---|---|---|---|
| | 40 | 48 | 58 |
| | 42 | 50 | 62 |
| | 38 | 52 | 60 |
| Average | 40 | 50 | 60 |

(b) Experiment 2

| Species | A | B | C |
|---|---|---|---|
| | 40 | 65 | 45 |
| | 25 | 35 | 75 |
| | 55 | 50 | 60 |
| Average | 40 | 50 | 60 |

Which experiment has better evidence that the true mean $NO_3$ removal rate differs between species? **Pause and think about this before reading on.**

Intuitively, we would say that Experiment 1 shows much stronger evidence for a true effect than Experiment 2. Why? Both experiments show the same differences among the treatment (species) means. So the variability in the treatment means is the same. However, **the variability among the observations within treatments** differs between the two experiments. In Experiment 1, the variability within treatments is much less than the variability among treatments. In Experiment 2, the variability within treatments is about the same as the variability among treatments.

The basic idea of ANOVA relies on the ratio of the among-treatment-means variation to the within-treatment variation. This is the F-ratio. The F-ratio can be thought of as a signal-to-noise ratio:

- Large ratios imply the signal (difference among the means) is large relative to the noise (variation within groups), providing evidence of a difference in the means.
- Small ratios imply the signal (difference among the means) is small relative to the noise, indicating no evidence that the means differ.

## 6.2 The F-test

When we take the ratio of two variances, it can be shown that the ratio follows an F-distribution with degrees of freedom equal to those of the two variances.

So, for example, say we want to compare the variability between two independent groups, each with normally distributed observations. We define the test statistic as the ratio of the two sample variances:

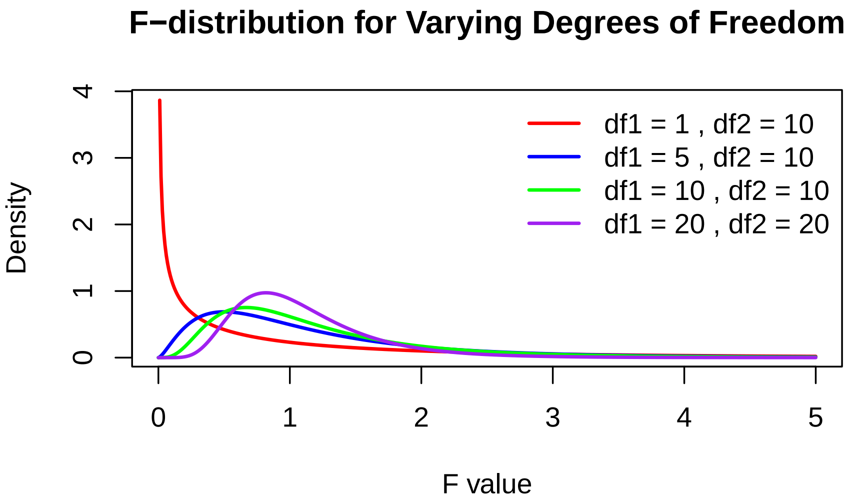
$$F = \frac{s_1^2}{s_2^2}$$

where $s_1^2$ and $s_2^2$ are the sample variances of the two groups. The resulting statistic follows an F-distribution with degrees of freedom:

- $df_1 = n_1 - 1$ for the numerator (corresponding to variance $s_1^2$)
- $df_2 = n_2 - 1$ for the denominator (corresponding to variance $s_2^2$)

The F-distribution is a probability distribution that arises frequently, particularly in ANOVA and regression analysis.

```r
# Define the range of F-values
x <- seq(0, 5, length.out = 500)

# Define degrees of freedom pairs
df_pairs <- list(
  c(1, 10),
  c(5, 10),
  c(10, 10),
  c(20, 20)
)

# Define colors for different lines
colors <- c("red", "blue", "green", "purple")

# Create an empty plot
plot(x, df(x, df_pairs[[1]][1], df_pairs[[1]][2]), type="n",
     xlab="F value", ylab="Density",
     main="F-distribution for Varying Degrees of Freedom")

# Loop through df pairs and add lines
for (i in seq_along(df_pairs)) {
  lines(x, df(x, df_pairs[[i]][1], df_pairs[[i]][2]), col=colors[i], lwd=2)
}

# Add a legend
legend("topright", legend=paste("df1 =", sapply(df_pairs, `[[`, 1), ", df2 =", sapply(df_pairs, `
       col=colors, lwd=2, bty="n")
```



Key properties of the F-distribution:

- It is always non-negative: $F \geq 0$.
- It is asymmetric and skewed to the right, particularly for small degrees of freedom.
- As the degrees of freedom increase, the F-distribution approaches a normal shape.

## 6.3 Analysis of Variance for CRD

Let's go back to the linear model for the single-factor CRD that we examined earlier:

$$Y_{ij} = \mu + A_i + e_{ij}$$

where $\mu$ is the overall mean, $A_i$ are the treatment effects (that is the difference between treatment means and the overall mean), and $e_{ij}$ are the error terms (the differences between the observation and the fitted value, i.e. treatment mean). Remember that the estimated values for these parameters are the observed values:

$$\begin{aligned}\hat{\mu} &= \bar{Y}_{..}\\ \hat{A}_i &= \bar{Y}_{i.} - \bar{Y}_{..}\\ \hat{e}_{ij} &= Y_{ij} - \bar{Y}_{i.}\end{aligned}$$

By taking $\mu$ over to the left-hand-side in the equation, and substituting the above observed values we obtain:

$$\begin{aligned} Y_{ij} - \mu &= (\mu_i - \mu) + (Y_{ij} - \mu) \\ Y_{ij} - \bar{Y} &= (\bar{Y}_i - \bar{Y}) + (Y_{ij} - \bar{Y}_i) \end{aligned}$$

Squaring and summing both sides gives the decomposition:

$$\sum_i \sum_j (Y_{ij} - \bar{Y})^2 = \sum_i \sum_j (\bar{Y}_i - \bar{Y})^2 + \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2$$

Each term represents squared deviations:

- The first term is of observations around the overall mean representing the total variation in the response.
- The second is of the group means around the overall mean representing the explained variation or variation between treatments and,
- The last term represents the deviations of observations from their treatment means (unexplained or within treatment variation).

We could also call these:

$$SS_{\text{total}} = SS_{\text{between groups}} + SS_{\text{within groups}}$$

or

$$SS_{\text{total}} = SS_{\text{treatment}} + SS_{\text{error}}$$

**The analysis of variance is based on this identity**[1]. The total sums of squares equals the sum of squares between groups plus the sum of squares within groups.

Back to our constructed example. What are the different sums of squares? For Experiment 1, we get: $SS_{\text{total}} = 624; SS_{\text{between groups}} = 600; SS_{\text{within groups}} = 24$. Verify these numbers and do the same for Experiment 2.

## 6.4 ANOVA Table

This division of the total sums of squares is typically summarised in an analysis of variance table. The first column contains the "source" of the variability with the first entry (the order is not important, although this is the typical order) representing the between-treatment variability (explained variation), second is

[1] In mathematics, an identity is an equation that is always true, regardless of the values of it's variables. In other words, the identity is true for all observations.

the error (unexplained variation, variation of experimental units within treatments) and lastly the total variation. Here we have used the notation $SS_A$ to represent the sums of squares for treatment factor A. The second column gives the sums of squares of each source. The third column contains the degrees of freedom.

| Source | Sums of<br>Squares (SS) | df | Means<br>Squares (MS) | F |
|---|---|---|---|---|
| Treatment | $\sum_i n_i(\bar{Y}_i - \bar{Y})^2$ | $a-1$ | $MS_A = SS_A/(a-1)$ | $MS_A/MSE$ |
| Residuals (Error) | $\sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2$ | $N-a$ | $MSE = SSE/(N-a)$ | |
| Total | $\sum_i \sum_j (Y_{ij} - \bar{Y})^2$ | $N-1$ | | |

The fourth column contains the Mean squares. This is what we get when we divide sums of squares by the appropriate degrees of freedom.

$$\text{MS} = \frac{SS}{df}$$

This is simply an average and may be seen as an estimate of variance. So when we divide the treatment SS by its degrees of freedom, we get an estimate of the variation due to treatments and similarly, for the the residual SS, we get an estimate of the error variance. You've seen this before!

$$\text{MSE} = \hat{\sigma}^2 = \frac{1}{N-a}\sum_i \sum_j (Y_{ij} - Y_{i.})^2$$

### 6.4.1 What Are Degrees of Freedom?

Degrees of freedom (df) represent the number of independent pieces of information available for estimating a parameter. When making statistical calculations, we typically lose one degree of freedom for every estimated parameter before the current calculation.

For example, when estimating the standard deviation of a data set, we first estimate the mean, thereby reducing the number of independent observations available to calculate variability. This is why the denominator in the variance formula is $N-1$:

$$s^2 = \frac{\sum(Y_i - \bar{Y})^2}{N-1}$$

You can think of degrees of freedom as the number of independent deviations around a mean. If we have $n$ observations and their mean, once we know $n-1$ of the values, the last one is fixed—it must take on a specific value to satisfy the mean equation. Therefore, only $n-1$ observations are truly free to vary.

**Example: Three Numbers Summing to a Fixed Mean**

Say we have three ($n = 3$) numbers: (4, 6, 8). The mean of these three numbers is 6. If we only knew the first two numbers (4,6) and the mean, the third number must be 8:

$$\begin{aligned}\bar{x} &= \frac{\sum x_i}{n}\\ 6 &= \frac{4+6+x_3}{3}\\ 18 &= 10 + x_3\\ x_3 &= 8\end{aligned}$$

Since the third number is uniquely determined by the first two and the mean, we only have $n-1$ (i.e., 2) degrees of freedom.

**Another Intuitive Analogy**

Imagine you are distributing a fixed amount of money among friends. If you have R100 and four friends, you can freely allocate money to three friends, but whatever is left must go to the fourth friend to ensure the total remains R100. Similarly, once the first $n-1$ values are chosen, the last value is determined, limiting the degrees of freedom.

**In ANOVA**

If you look at the treatment sums of squares: $\sum_i n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$. We have $a$ deviations around the grand mean. But once we know $a-1$ of the treatment means and the grand mean[2], the last mean is fixed. So we have $a-1$ independent deviations around the overall mean.

If you look at the treatment sums of squares: $\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$. We are using $N$ observations and calculating the deviations of these observations around the overall mean. So, only $N-1$ observations are free to vary, the last observation is fixed for the calculated mean to hold true.

## 6.5 Back to the constructed example

What does the ANOVA table look like for our constructed example? You've already worked out the sums of squares. What are the df's and Mean squares?

Let's have a look at Experiment 1 first.

[2] Remember, $\mu = \frac{\sum \mu_i}{a}$.

```
# Experiment 1 data
exp1data <- data.frame(species = rep(c("A","B","C"), each = 3),
                      response = c(40,42,38,48,50,52,58,62,60))

exp1_anova <- aov(response~species, data = exp1data)
summary(exp1_anova)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)
species      2    600     300      75 5.69e-05 ***
Residuals    6     24       4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And then Experiment 2:

```
# Experiment 2 data
exp2data <- data.frame(species = rep(c("A","B","C"), each = 3),
                      response = c(40,25,55,65,35,50,45,75,60))

exp2_anova <- aov(response~species, data = exp2data)
summary(exp2_anova)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
species      2    600     300   1.333  0.332
Residuals    6   1350     225
```

Since the overall mean and the treatment means were the same in both experiment, we expected the $SS_{\text{treatment}}$ to be the same in both experiments. This was indeed the case – they are 600 in both experiments. The sample sizes were also the same in both experiments, so we would expect the df to be the same. With 9 observations, we have 8 df in total. Three treatments (Species) leads to 2 treatment df and 6 df remain for the residuals. The difference between the two experiments is that the observations were much more variable in Experiment 2 than in Experiment 1. Accordingly, we find that $SS_{\text{error}}$ was much larger in Experiment 2, and this led to larger MSE in Experiment 2. How does this affect the conclusions we draw from each of the experiments? This is where the F-ratio comes in.

## 6.6 The F-test in ANOVA

We first set up the null and alternate hypothesis. The null hypothesis is that all treatments have the same mean, or equivalently, that all treatment effects are zero.

$$H_0 : \mu_1 = \mu_2 = ... = \mu_a$$
$$H_0 : A_1 = A_2 = ... = A_a = 0$$

And the alternative hypothesis is the opposite of that:

$$
\begin{aligned}
H_A &: \text{At least one } \mu_i \text{ is different.} \\
H_A &: \text{At least one } A_i \neq 0
\end{aligned}
$$

If $H_0$ is true, the among-treatment-means variation should equal the within-treatment variation. We can use the F-ratio to test $H_0$:

Read that again. The alternative is that at least one treatment is different, there is a difference somewhere. It is not that all treatment means are different.

$$F^* = \frac{MS_A}{MSE}$$

This ratio has an F-distribution with $a-1$ numerator degrees of freedom and $N-a$ denominator degrees of freedom.

You can think of the *F-ratio* as a signal-to-noise ratio. If $H_0$ is true, $F$ is expected to be close to 1. If $H_0$ is false, $F$ is expected to be much larger than 1. This means that the F-test we conduct is a **one-sided upper tailed test**. If $H_0$ is false, the means squares for treatment will be much larger than the MSE, resulting in large F-values. We are only interested in this one side of possible outcomes therefore, a one-sided test.

In Experiment 1, $F = \frac{300}{4} = 75$, which leads to a very small $p$-value ($< 0.001$). The signal was much larger than the noise, and our data are very unlikely if $H_0$ were true. So we have good evidence that the treatments differ.

In Experiment 2, $F = \frac{300}{225} = 1.33$, which leads to a large $p$-value (0.33). Signal and noise were of similar magnitude, and our data are not unlikely if $H_0$ were true. So we have no evidence against $H_0$, i.e., no evidence that nitrate extraction differs between species.

How did we get these p-values? This is the same as in any hypothesis test. We have a test statistic and to say something about how likely this test statistic (or more extreme is) under the null hypothesis, we need the null distribution of the test statistic (that is the sampling distribution of the test statistic as if the null hypothesis were true). We then compared the observed value of the test statistic to that null distribution and asked ourselves how unusual it is in light of that distribution. Does our test statistic belong to this null distribution?

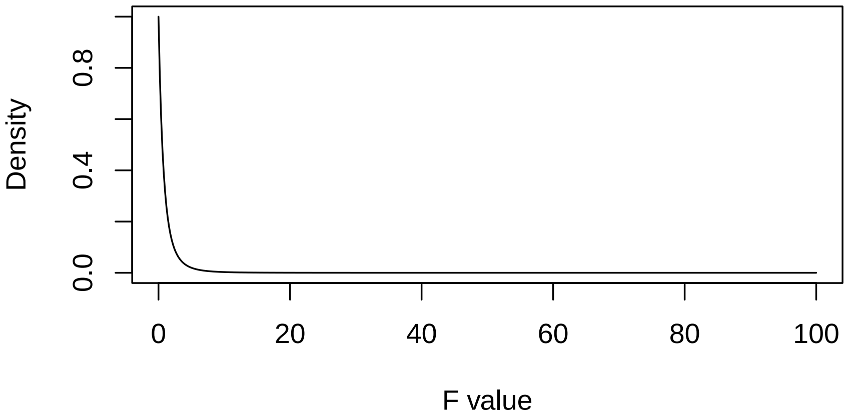
The $F$ test statistic follows an F distribution as specified above.

$$\mathrm{F}^* \sim \mathrm{F}_{(a-1),\,(N-a)}$$

For both experiment, this equates to an F distribution with 2 numerator and 6 denominator degrees of freedom which looks like this:

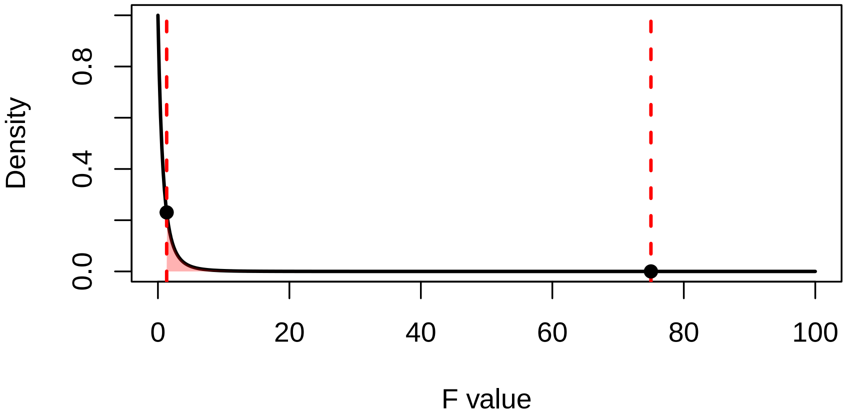
```r
# Define the range of F-values
x <- seq(0, 100, length.out = 500)
y <- df(x, df1 = 2, df2 = 6)
```

```
plot(x, y, type="l",
     xlab="F value", ylab="Density",
     main="")
```



We can plot the test statistics on the graph as well and highlight the area under the curve to the right of each of these test statistics:

```
# Define x values
x <- seq(0, 100, length.out = 500)
y <- df(x, df1 = 2, df2 = 6)

# Define test_stats
test_stats <- c(75, 1.33)

# Plot the F-distribution density curve
plot(x, y, type = "l", col = "black", lwd = 2,
     xlab = "F value", ylab = "Density",
     main = "")

# Add vertical lines at test_stats
abline(v = test_stats, col = "red", lty = 2, lwd = 2)

# Shade the areas to the right of the test_stats
polygon(c(test_stats[1], x[x >= test_stats[1]], max(x)),
        c(0, y[x >= test_stats[1]], 0), col = rgb(0, 0, 1, 0.3), border = NA)

polygon(c(test_stats[2], x[x >= test_stats[2]], max(x)),
        c(0, y[x >= test_stats[2]], 0), col = rgb(1, 0, 0, 0.3), border = NA)

# Add points at the critical values
points(test_stats, df(test_stats, df1 = 2, df2 = 6), pch = 19, col = "black")
```



Remember sampling distributions are probability distributions. For continuous random variables, the area under the curve represents probability. Specifically, the probability of a random variable taking on a specific value or larger, is the area under the curve to the right of that value. For test statistics and their probability distribution, that probability is the p-value. The p-value is the probability of observing a test statistic at least as extreme as we did if the null hypothesis was in fact true. The smaller the p-value, the stronger the evidence against $H_0$.

We can obtain the p-value in two ways (you will need to be able to do both):

1. Using Software.

In R, there are several built-in functions for certain probability distributions. These functions typically follow a naming convention:

- `d<dist>()` for density functions
- `p<dist>()` for cumulative probability functions
- `q<dist>()` for quantile functions
- `r<dist>()` for random sampling

For example, when working with the F-distribution, we use:

- `df(x, df1, df2)` for the probability density function (PDF)

- `pf(x, df1, df2)` for the cumulative distribution function (CDF)
- `qf(p, df1, df2)` for quantiles
- `rf(n, df1, df2)` for random sampling

To obtain a p-value, we often use the cumulative probability functions (`p<dist>()`) with returns $Pr[X < x]$ so $Pr[X > x] = 1 - Pr[X < x]$. Below is how to obtain the p-value for the second experiment:

```r
f_statistic <- 1.33
df1 <- 2  # Numerator degrees of freedom
df2 <- 6  # Denominator degrees of freedom

# Upper-tail probability (right-tailed test)
p_value <- 1 - pf(f_statistic, df1, df2)
p_value
```

```
[1] 0.332583
```

This value is quite large and corresponds to the area to the right of an F value of 1.33 for the distribution above. We interpret this p-value as the test statistic is quite likely to have come from this null distribution, there is a 33% chance of observing this test statistic or more extreme if the null hypothesis is true. We do not have strong evidence against the null hypothesis of equal means.

Caution

A large p-value does not mean that $H_0$ is true!

- The p-value is not the probability that the null hypothesis is true.
- The p-value is not the probability that the alternative hypothesis is false.
- The p-value is a statement about the relation of the data to the null hypothesis.
- The p-value does not indicate the size or biological importance of the observed pattern.

Tip

You can round the p-value if you need to enter the value to a certain number of decimals in a quiz or test using the function `round`.

2. Using tables.

Before the days of widespread programming, statisticians used tables to find critical values and p-values for various probability distributions. These tables were pre-computed for different significance levels (e.g., 0.05, 0.01) and degrees of freedom. In modern statistical analysis, we no longer rely on static tables, as software like R can compute exact probabilities. But since we have written examinations, we have to learn how to do this and it is a useful exercise to make sure you understand what you are doing and not just spitting out a value.

F-tables look like this:

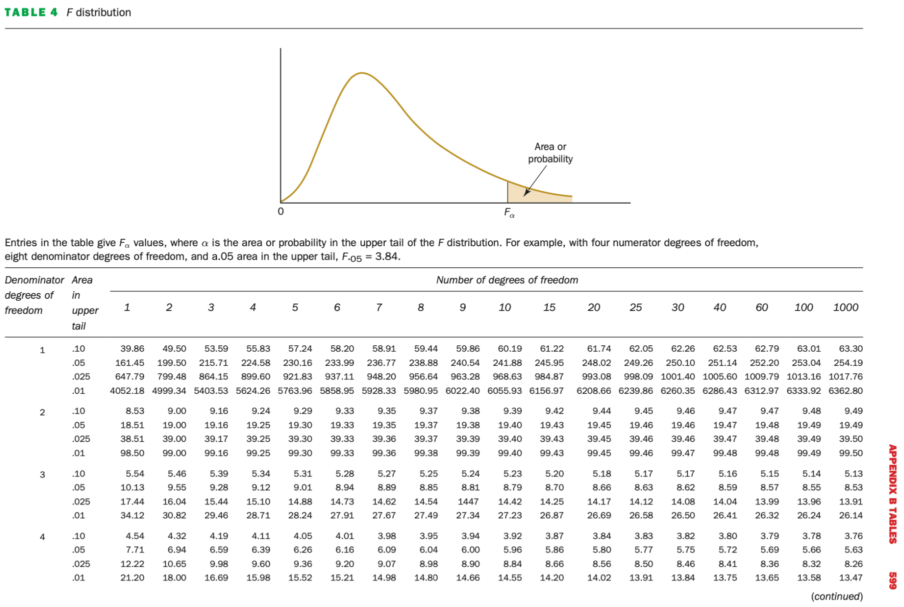

Area or probability
0
$F_\alpha$
APPENDIX B TABLES 599

**TABLE 4** *F* distribution

Entries in the table give $F_\alpha$ values, where $\alpha$ is the area or probability in the upper tail of the *F* distribution. For example, with four numerator degrees of freedom, eight denominator degrees of freedom, and a.05 area in the upper tail, $F_{.05}$ = 3.84.

| Denominator degrees of freedom | Area in upper tail | Number of degrees of freedom | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 15 | 20 | 25 | 30 | 40 | 60 | 100 | 1000 |
| 1 | .10 | 39.86 | 49.50 | 53.59 | 55.83 | 57.24 | 58.20 | 58.91 | 59.44 | 59.86 | 60.19 | 61.22 | 61.74 | 62.05 | 62.26 | 62.53 | 62.79 | 63.01 | 63.30 |
| | .05 | 161.45 | 199.50 | 215.71 | 224.58 | 230.16 | 233.99 | 236.77 | 238.88 | 240.54 | 241.88 | 245.95 | 248.02 | 249.26 | 250.10 | 251.14 | 252.20 | 253.04 | 254.19 |
| | .025 | 647.79 | 799.48 | 864.15 | 899.60 | 921.83 | 937.11 | 948.20 | 956.64 | 963.28 | 968.63 | 984.87 | 993.08 | 998.09 | 1001.40 | 1005.60 | 1009.79 | 1013.16 | 1017.76 |
| | .01 | 4052.18 | 4999.34 | 5403.53 | 5624.26 | 5763.96 | 5858.95 | 5928.33 | 5980.95 | 6022.40 | 6055.93 | 6156.97 | 6208.66 | 6239.86 | 6260.35 | 6286.43 | 6312.97 | 6333.92 | 6362.80 |
| 2 | .10 | 8.53 | 9.00 | 9.16 | 9.24 | 9.29 | 9.33 | 9.35 | 9.37 | 9.38 | 9.39 | 9.42 | 9.44 | 9.45 | 9.46 | 9.47 | 9.47 | 9.48 | 9.49 |
| | .05 | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 | 19.40 | 19.43 | 19.45 | 19.46 | 19.46 | 19.47 | 19.48 | 19.49 | 19.49 |
| | .025 | 38.51 | 39.00 | 39.17 | 39.25 | 39.30 | 39.33 | 39.36 | 39.37 | 39.39 | 39.40 | 39.43 | 39.45 | 39.46 | 39.46 | 39.47 | 39.48 | 39.49 | 39.50 |
| | .01 | 98.50 | 99.00 | 99.16 | 99.25 | 99.30 | 99.33 | 99.36 | 99.38 | 99.39 | 99.40 | 99.43 | 99.45 | 99.46 | 99.47 | 99.48 | 99.48 | 99.49 | 99.50 |
| 3 | .10 | 5.54 | 5.46 | 5.39 | 5.34 | 5.31 | 5.28 | 5.27 | 5.25 | 5.24 | 5.23 | 5.20 | 5.18 | 5.17 | 5.17 | 5.16 | 5.15 | 5.14 | 5.13 |
| | .05 | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 | 8.70 | 8.66 | 8.63 | 8.62 | 8.59 | 8.57 | 8.55 | 8.53 |
| | .025 | 17.44 | 16.04 | 15.44 | 15.10 | 14.88 | 14.73 | 14.62 | 14.54 | 1447 | 14.42 | 14.25 | 14.17 | 14.12 | 14.08 | 14.04 | 13.99 | 13.96 | 13.91 |
| | .01 | 34.12 | 30.82 | 29.46 | 28.71 | 28.24 | 27.91 | 27.67 | 27.49 | 27.34 | 27.23 | 26.87 | 26.69 | 26.58 | 26.50 | 26.41 | 26.32 | 26.24 | 26.14 |
| 4 | .10 | 4.54 | 4.32 | 4.19 | 4.11 | 4.05 | 4.01 | 3.98 | 3.95 | 3.94 | 3.92 | 3.87 | 3.84 | 3.83 | 3.82 | 3.80 | 3.79 | 3.78 | 3.76 |
| | .05 | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 | 5.86 | 5.80 | 5.77 | 5.75 | 5.72 | 5.69 | 5.66 | 5.63 |
| | .025 | 12.22 | 10.65 | 9.98 | 9.60 | 9.36 | 9.20 | 9.07 | 8.98 | 8.90 | 8.84 | 8.66 | 8.56 | 8.50 | 8.46 | 8.41 | 8.36 | 8.32 | 8.26 |
| | .01 | 21.20 | 18.00 | 16.69 | 15.98 | 15.52 | 15.21 | 14.98 | 14.80 | 14.66 | 14.55 | 14.20 | 14.02 | 13.91 | 13.84 | 13.75 | 13.65 | 13.58 | 13.47 |

(*continued*)

Read it carefully. The table says: "Entries in the table give $F_\alpha$ values, where $\alpha$ is the area or probability in the upper tail of the F distribution. For example, with four numerator degrees of freedom, eight denominator degrees of freedom, and 0.05 area in the upper tail, F.05 = 3.84." This is important, not all tables look like this. See if you can find the F-value mentioned.

The numerator df is in the column and the denominator df is in the row. In the row dimension are different $\alpha$ values as well. To find an F-value, locate the df in the column and row. Can you find the following:

- $F_{4,4}^{0.05} = 6.39$
- $F_{10,2}^{0.1} = 9.39$
- $F_{1,1}^{0.025} = 647.79$
- $F_{7,3}^{0.01} = 27.67$

This is how we find critical values of F-distributions. If you are asked to compare a test statistic with a critical value at a specific significance level, you will find the value with a table like this. To find the critical values in R, we use the `fq` function:

```
# F4,4 0.05

qf(p = 0.05, df1 = 4, df2 = 4, lower.tail = FALSE) # if lower.tail = TRUE which is the default,
[1] 6.388233
```

```
# F10,2 0.11
qf(p = 0.1,   df1 = 10, df2 = 2, lower.tail = FALSE)
```

```
[1] 9.391573
```

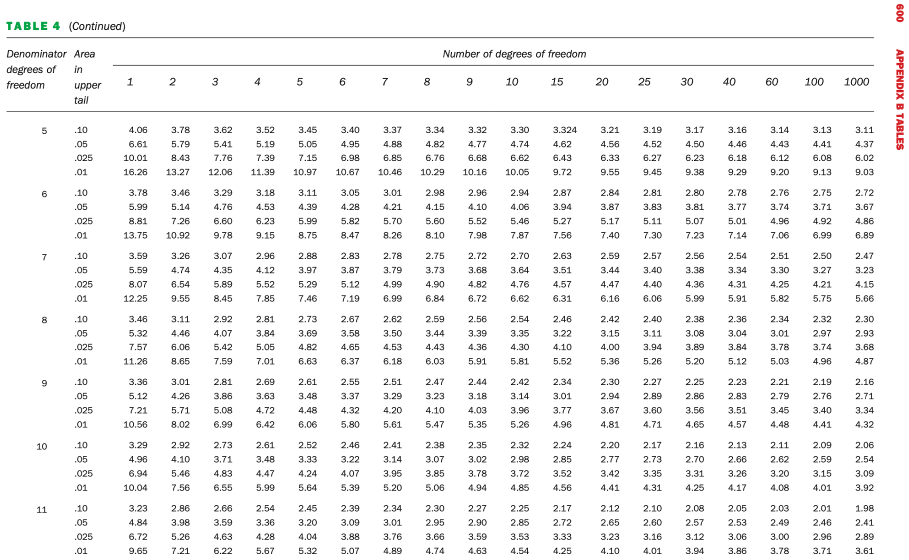
```
# F1,1 0.025
qf(p = 0.025, df1 = 1,  df2 = 1, lower.tail = FALSE)
```

```
[1] 647.789
```

```
# F7,3 0.01
qf(p = 0.01,  df1 = 7,  df2 = 3, lower.tail = FALSE)
```

```
[1] 27.6717
```

Now, how do we use the tables to obtain p-values? The test statistic for the first Experiment was 1.33 and the df's were 2 (num) and 6 (denom). If we look at the table above, it only goes to 4 denominator degrees of freedom, so we need the continuation of the table.



600 APPENDIX B TABLES

**TABLE 4** (*Continued*)

| *Denominator degrees of freedom* | *Area in upper tail* | *Number of degrees of freedom* | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *15* | *20* | *25* | *30* | *40* | *60* | *100* | *1000* |
| 5 | .10 | 4.06 | 3.78 | 3.62 | 3.52 | 3.45 | 3.40 | 3.37 | 3.34 | 3.32 | 3.30 | 3.324 | 3.21 | 3.19 | 3.17 | 3.16 | 3.14 | 3.13 | 3.11 |
| | .05 | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 | 4.62 | 4.56 | 4.52 | 4.50 | 4.46 | 4.43 | 4.41 | 4.37 |
| | .025 | 10.01 | 8.43 | 7.76 | 7.39 | 7.15 | 6.98 | 6.85 | 6.76 | 6.68 | 6.62 | 6.43 | 6.33 | 6.27 | 6.23 | 6.18 | 6.12 | 6.08 | 6.02 |
| | .01 | 16.26 | 13.27 | 12.06 | 11.39 | 10.97 | 10.67 | 10.46 | 10.29 | 10.16 | 10.05 | 9.72 | 9.55 | 9.45 | 9.38 | 9.29 | 9.20 | 9.13 | 9.03 |
| 6 | .10 | 3.78 | 3.46 | 3.29 | 3.18 | 3.11 | 3.05 | 3.01 | 2.98 | 2.96 | 2.94 | 2.87 | 2.84 | 2.81 | 2.80 | 2.78 | 2.76 | 2.75 | 2.72 |
| | .05 | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 | 3.94 | 3.87 | 3.83 | 3.81 | 3.77 | 3.74 | 3.71 | 3.67 |
| | .025 | 8.81 | 7.26 | 6.60 | 6.23 | 5.99 | 5.82 | 5.70 | 5.60 | 5.52 | 5.46 | 5.27 | 5.17 | 5.11 | 5.07 | 5.01 | 4.96 | 4.92 | 4.86 |
| | .01 | 13.75 | 10.92 | 9.78 | 9.15 | 8.75 | 8.47 | 8.26 | 8.10 | 7.98 | 7.87 | 7.56 | 7.40 | 7.30 | 7.23 | 7.14 | 7.06 | 6.99 | 6.89 |
| 7 | .10 | 3.59 | 3.26 | 3.07 | 2.96 | 2.88 | 2.83 | 2.78 | 2.75 | 2.72 | 2.70 | 2.63 | 2.59 | 2.57 | 2.56 | 2.54 | 2.51 | 2.50 | 2.47 |
| | .05 | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 | 3.51 | 3.44 | 3.40 | 3.38 | 3.34 | 3.30 | 3.27 | 3.23 |
| | .025 | 8.07 | 6.54 | 5.89 | 5.52 | 5.29 | 5.12 | 4.99 | 4.90 | 4.82 | 4.76 | 4.57 | 4.47 | 4.40 | 4.36 | 4.31 | 4.25 | 4.21 | 4.15 |
| | .01 | 12.25 | 9.55 | 8.45 | 7.85 | 7.46 | 7.19 | 6.99 | 6.84 | 6.72 | 6.62 | 6.31 | 6.16 | 6.06 | 5.99 | 5.91 | 5.82 | 5.75 | 5.66 |
| 8 | .10 | 3.46 | 3.11 | 2.92 | 2.81 | 2.73 | 2.67 | 2.62 | 2.59 | 2.56 | 2.54 | 2.46 | 2.42 | 2.40 | 2.38 | 2.36 | 2.34 | 2.32 | 2.30 |
| | .05 | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 | 3.22 | 3.15 | 3.11 | 3.08 | 3.04 | 3.01 | 2.97 | 2.93 |
| | .025 | 7.57 | 6.06 | 5.42 | 5.05 | 4.82 | 4.65 | 4.53 | 4.43 | 4.36 | 4.30 | 4.10 | 4.00 | 3.94 | 3.89 | 3.84 | 3.78 | 3.74 | 3.68 |
| | .01 | 11.26 | 8.65 | 7.59 | 7.01 | 6.63 | 6.37 | 6.18 | 6.03 | 5.91 | 5.81 | 5.52 | 5.36 | 5.26 | 5.20 | 5.12 | 5.03 | 4.96 | 4.87 |
| 9 | .10 | 3.36 | 3.01 | 2.81 | 2.69 | 2.61 | 2.55 | 2.51 | 2.47 | 2.44 | 2.42 | 2.34 | 2.30 | 2.27 | 2.25 | 2.23 | 2.21 | 2.19 | 2.16 |
| | .05 | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 | 3.01 | 2.94 | 2.89 | 2.86 | 2.83 | 2.79 | 2.76 | 2.71 |
| | .025 | 7.21 | 5.71 | 5.08 | 4.72 | 4.48 | 4.32 | 4.20 | 4.10 | 4.03 | 3.96 | 3.77 | 3.67 | 3.60 | 3.56 | 3.51 | 3.45 | 3.40 | 3.34 |
| | .01 | 10.56 | 8.02 | 6.99 | 6.42 | 6.06 | 5.80 | 5.61 | 5.47 | 5.35 | 5.26 | 4.96 | 4.81 | 4.71 | 4.65 | 4.57 | 4.48 | 4.41 | 4.32 |
| 10 | .10 | 3.29 | 2.92 | 2.73 | 2.61 | 2.52 | 2.46 | 2.41 | 2.38 | 2.35 | 2.32 | 2.24 | 2.20 | 2.17 | 2.16 | 2.13 | 2.11 | 2.09 | 2.06 |
| | .05 | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 | 2.85 | 2.77 | 2.73 | 2.70 | 2.66 | 2.62 | 2.59 | 2.54 |
| | .025 | 6.94 | 5.46 | 4.83 | 4.47 | 4.24 | 4.07 | 3.95 | 3.85 | 3.78 | 3.72 | 3.52 | 3.42 | 3.35 | 3.31 | 3.26 | 3.20 | 3.15 | 3.09 |
| | .01 | 10.04 | 7.56 | 6.55 | 5.99 | 5.64 | 5.39 | 5.20 | 5.06 | 4.94 | 4.85 | 4.56 | 4.41 | 4.31 | 4.25 | 4.17 | 4.08 | 4.01 | 3.92 |
| 11 | .10 | 3.23 | 2.86 | 2.66 | 2.54 | 2.45 | 2.39 | 2.34 | 2.30 | 2.27 | 2.25 | 2.17 | 2.12 | 2.10 | 2.08 | 2.05 | 2.03 | 2.01 | 1.98 |
| | .05 | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 | 2.72 | 2.65 | 2.60 | 2.57 | 2.53 | 2.49 | 2.46 | 2.41 |
| | .025 | 6.72 | 5.26 | 4.63 | 4.28 | 4.04 | 3.88 | 3.76 | 3.66 | 3.59 | 3.53 | 3.33 | 3.23 | 3.16 | 3.12 | 3.06 | 3.00 | 2.96 | 2.89 |
| | .01 | 9.65 | 7.21 | 6.22 | 5.67 | 5.32 | 5.07 | 4.89 | 4.74 | 4.63 | 4.54 | 4.25 | 4.10 | 4.01 | 3.94 | 3.86 | 3.78 | 3.71 | 3.61 |

Now we locate the F-values with 2 and 6 degrees of freedom and compare the test statistic of the second experiment (1.33) to them. The smallest value is 3.46 where the probability to the right of that value is 0.1. Our test statistic is much smaller than this, so lies further to the right and so logically, the right-hand-side probability of this value with be greater than 0.1. So we conclude that the p-value that our p-value is $> 0.1$ (which it is, we calculated it to be 0.32). With tables we cannot get exact probabilities, but we can say something about the magnitude of the p-value. Try it for the first experiment which had an F-value of 75.

## 6.7 Conclusion: Does social media multitasking impact academic performance of students?

Let's revisit the real experiment we started this section with. I repeat the experiment description below.

> Example 5.1
>
> Two researchers from Turkey, Demirbilek and Talan (2018), conducted a study to try and answer this question. Specifically, they examined the impact of social media multitasking during live lectures on students' academic performance.
> A total of 120 undergraduate students were randomly assigned to one of three groups:
>
> 1. **Control Group:** Students used traditional pen-and-paper note-taking.
> 2. **Experimental Group 1 (Exp 1):** Students engaged in SMS texting during the lecture.
> 3. **Experimental Group 2 (Exp 2):** Students used Facebook during the lecture.
>
> Over a three-week period, participants attended the same lectures on Microsoft Excel. To measure academic performance, a standardised test was administered.

In the previous sections we introduced this study, checked the model assumptions and obtained estimates of the model parameters. Now equipped with that information and all that you have learnt, we are ready to fit to conduct the ANOVA hypothesis test to finally answer our question:

Does social media multitasking impact academic performance of students?

We start with the hypotheses:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_a$$

In words we say that the average academic performance of students did not differ across the treatments (levels of social media multitasking).

And the alternative hypothesis is the opposite of that:

$$H_A : \text{At least one } \mu_i \text{ is different.}$$

At least one of the social media multitasking treatments resulted in a different mean academic performance, they are not all equal.

We have fit the model already (called `m1`) and call the `summary` function to obtain the ANOVA table:

```
# m1 <- aov(Posttest ~ Group, data = multitask)

summary(m1)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)    
Group         2  10975    5488   27.42 1.72e-10 ***
Residuals   117  23417     200                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Violà! We have our ANOVA table. Inspect the results and make sure you understand how each value is obtained and what they represent. By just looking at the table, you should be able to answer the following questions:

1. How many treatments were there?
2. How many observations in total?
3. Is there evidence for a treatment effect?

The first two you can answer with the degrees of freedom and the third is answered by conducting the hypothesis test. With three treatments, we have 2 treatment degrees of freedom. We had 40 students per group, the sample size is then 120 which means there are 117 degrees of freedom for the residuals. The treatment MS (5488) was much larger that the MSE (200). This leads to an F-ratio of 27.42 with a p-value of $1.72 \times e^{-10}$ (that's extremely small). We have strong evidence that the treatments did result different academic performances across students At least one treatment resulted in a different mean academic performance. In a report, you would write:

"The manipulation of social media multitasking affected the academic performance of students in this experiment ($F_{2,117} = 27.42$, $p = 1.72 \times e^{-10}$)."

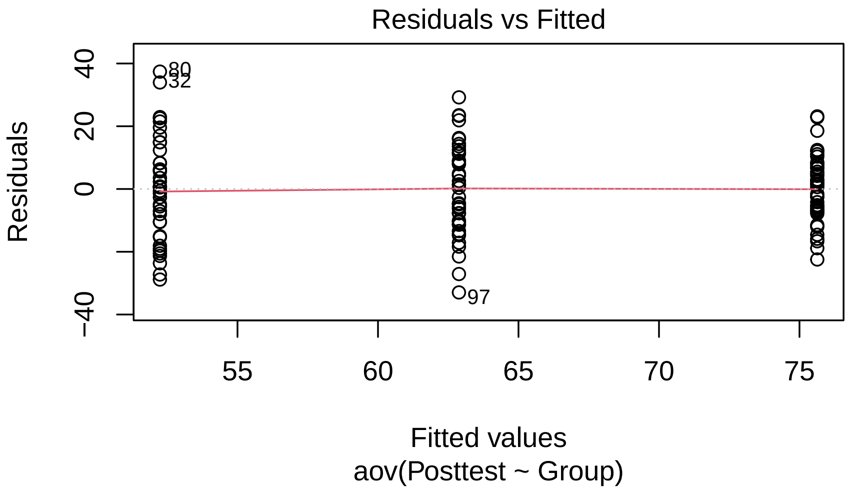
But which treatments differed? **We cannot answer that question with this hypothesis.** It only tells us that there is a difference, there is a treatment effect. **It does not tell us where the difference or possible differences lie.** To determine this, we need to use treatment contrasts. Before we do this or present any results, we need to do one last thing.

## 6.8 Model Checking

Remember that we said some of our assumptions need to be checked after the model is fitted. Our model specifies the error terms are (1) normally distributed, (2) all with the same variance (homoscedastic), and (3) that they are independent. The residuals are estimates of these error terms and we can therefore use them to check the model assumptions. Normally distributed, equal variance and independent really means that there is no discernible pattern or structure left in the residuals. If there is, then the model has failed to pick up an important structure in the data.[3]

We call the function `plot` on our model object. For our purposes we are only going to look at two of the plots and we inspect them one by one by specifying the plot number with the argument `which`:

```
plot(m1, which = 1)
```

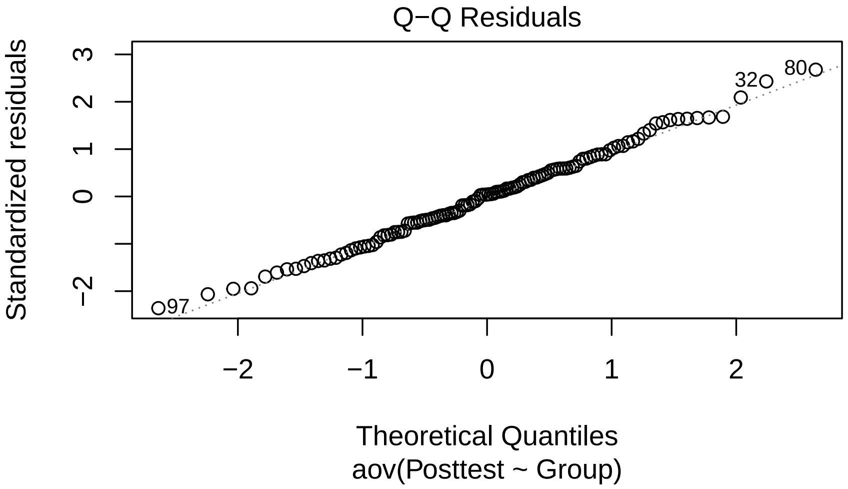


This is a plot of the residuals (obs - fitted) against the fitted values and we are hoping to see no patterns. We have three lines, one for each treatment group and we want to check that our residuals are centered around zero and have constant variance across the groups. [4]

```
plot(m1, which = 2)
```

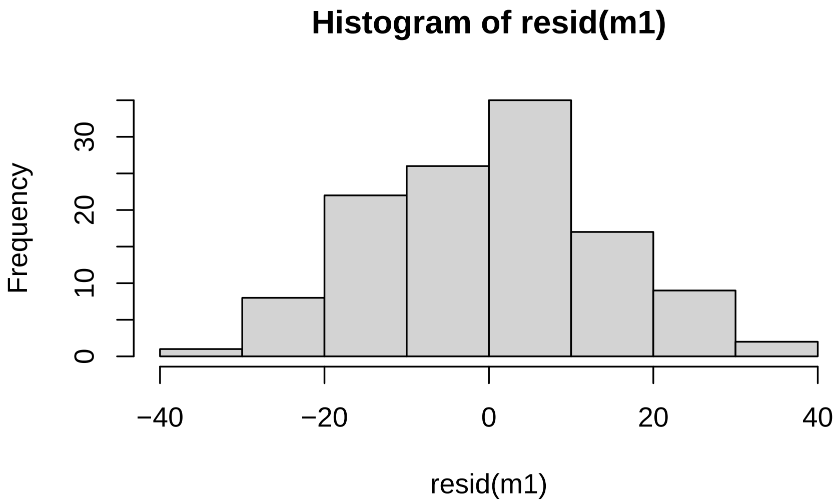
[3] The same concepts apply to linear regression models.

[4] Remember we assumed $e_{ij} \sim N(0, \sigma^2)$ and residuals are estimates of the errors.



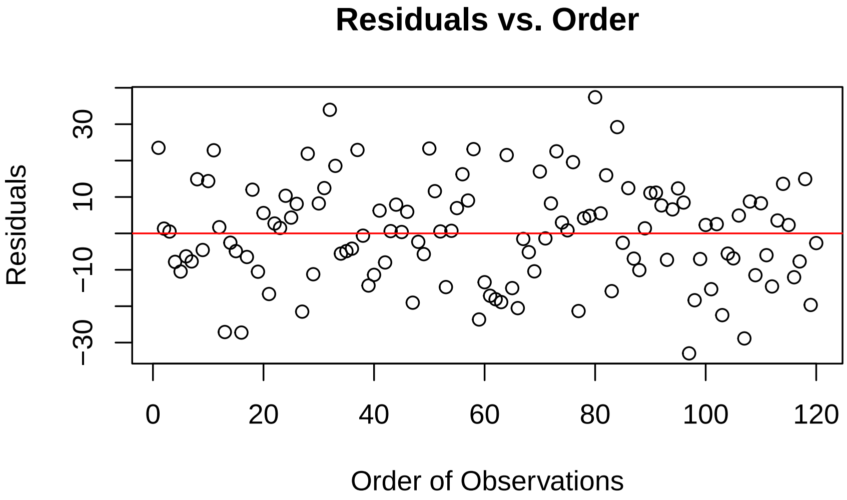
The second plot is a Q-Q plot which we have seen before when we checked the assumption of normality before model fitting. Now, we plot the standardised residuals against the theoretical quantiles of a standard normal distribution. We are looking for the same pattern as before, that the points fall close to the dotted line. As usual, there many be some deviations at the tails but for the most part, there are no serious problems with this plot. If there is some doubt, we can also look at a histogram of the residuals:

```
hist(resid(m1))
```



The assumption of independent errors is mostly checked before model fitting and by consideration of the experimental design. If we suspected auto-correlated residuals, we could plot the residuals against order:

```r
plot(resid(m1) ~ seq_along(resid(m1)),
     xlab = "Order of Observations",
     ylab = "Residuals",
     main = "Residuals vs. Order")
abline(h = 0, col = "red")
```



There are no patterns at all, the residuals appear randomly distributed. So no indications of dependence.

## 6.9 Summary

That's a lot. So let's summarise what we did in this chapter:

We introduced Analysis of Variance (ANOVA), which is fundamentally the same as a regression model with categorical variables but parameterised differently. ANOVA allows us to partition total variance into between-treatment and within-treatment variability, helping us determine whether observed differences in the response variable are due to the treatments and not just sampling error.

We explored ANOVA through a constructed experiment on nitrate removal by plants, demonstrating that variation within treatments influences our ability to detect true treatment effects. The F-ratio, a measure of the signal-to-noise ratio, is central to ANOVA. A large F-ratio suggests that between-group variability is greater than within-group variability, providing evidence that at least one treatment differs.

The F-test determines statistical significance, and its p-value is derived from the F-distribution. A small p-value suggests strong evidence against the null hypothesis ($H_0$), indicating at least one group mean differs. The ANOVA table summarises the calculations of the hypothesis test, including sums of squares (SS), degrees of freedom (df), mean squares (MS), and the F-statistic.

Applying ANOVA to real experimental data, we analysed the impact of social media multitasking on student performance. With three treatment groups (control, SMS, Facebook), we found a statistically significant effect ($F_{2,117} = 27.42$, $p = 1.72 \times 10^{-10}$), confirming that at least one treatment influenced academic performance. However, ANOVA does not specify which groups differ and how they differ — this requires post-hoc tests.

Finally, we validated model assumptions:

- Normality: Checked via a Q-Q plot and histogram of residuals.
- Homoscedasticity (equal variance): Examined using a residuals vs. fitted plot.
- Independence: Considered in the experimental design and checked by plotting residuals against observation order.

# Chapter 7

# Contrasts

The aim in many experiments is to compare treatments. To do this we contrast one group of means with another, i.e. we compare means, or groups of means, to see if treatments differ, and by how much they differ. A comparison of treatments (or groups of treatments) is called a contrast. If the experiment has been conducted as a result of specific research hypotheses, these will already define the contrasts we should construct first.

For a single-factor CRD with only two treatments, we could conduct a t-test to compare the two means or construct a confidence interval to estimate the difference. But we know that with more than two treatments, we encounter problems of multiple testing. How do we contrast treatments when we have a factor with more than two levels?

## Contrasting pairs of treatment means

We wrote the ANOVA model as:

$$Y_{ij} = \mu + A_i + e_{ij}$$

with overall mean and the treatment effects as the parameters (as well as the error variance). Because the effects are constrained to sum to zero, i.e. $\sum_i^a A_i = 0$ we call this ANOVA model the *sum-to-zero* parameterisation.

The above parameterisation is useful for constructing ANOVA tables. For estimating differences between treatments, however, a different parameterisation is more useful:

$$Y_{ij} = A_i + e_{ij}$$

In this version, we no longer have the overall mean as a parameter but use only the treatment effects $A_i$. Remember that any model ultimately needs to describe the treatment means. There are a number of different ways in which to do this. One is the so-called *treatment contrast* parameterisation, which R uses as default for regression models. In this parameterisation, $A_1$ estimates the mean of the baseline treatment (by default, R orders the treatments alphabetically and takes the first one as baseline). The other parameters then estimate the difference between each treatment and the baseline treatment: $A_2$ estimates the difference between the second and the first treatment, $A_3$ estimates the difference between the third and the first, etc.

To get a better understanding of this, let's fit the model to the social media data with this parameterisation. In R, this is done by using the `lm` function.

Construction of treatmenat means under the *treatment contrast* parameterisation:

$$\begin{aligned} \mu_1 &= A_1 \\ \mu_2 &= A_1 + A_2 \\ &\vdots \\ \mu_a &= A_1 + A_a \end{aligned}$$

and under the *sum-to-zero* parameterisation:

$$\begin{aligned} \mu_1 &= \mu + A_1 \\ \mu_2 &= \mu + A_2 \\ &\vdots \\ \mu_a &= \mu + A_a \end{aligned}$$

```r
m1.tc <- lm(Posttest ~ Group, data = multitask)
summary(m1.tc)
```

```

Call:
lm(formula = Posttest ~ Group, data = multitask)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.964 -10.175   0.583   8.550  37.408 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   75.634      2.237  33.812  < 2e-16 ***
GroupExp1    -12.752      3.163  -4.031 9.92e-05 ***
GroupExp2    -23.394      3.163  -7.395 2.32e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.15 on 117 degrees of freedom
Multiple R-squared:  0.3191,	Adjusted R-squared:  0.3075 
F-statistic: 27.42 on 2 and 117 DF,  p-value: 1.717e-10
```

This is exactly the same output as you have seen before in the regression section! The intercept measures the mean of the baseline treatment (here it is the Control group). The next estimate `GroupExp1` is the difference between the mean of Experiment 1 and the mean of the Control group. Similarly, the last one is the difference between the mean of the Control Group and that of Experiment 2. You can verify this by using the mean estimates we obtain when we fit the model previously:

```r
model.tables(m1, type = "means")
```

```

Tables of means
Grand mean
```

```
63.58527

 Group
Group
Control    Exp1    Exp2
  75.63   62.88   52.24
```

```
62.88 - 75.63 #GroupExp1
```

```
[1] -12.75
```

```
52.24 - 75.63 #GroupExp2
```

```
[1] -23.39
```

Why is this useful? Now, we can formally test whether these differences are statistically significant using a hypothesis test!

Think back to regression—what was the null hypothesis for the coefficients in the output?

It was:

$$\beta_i = 0$$

.

The same principle applies here. We test whether the treatment effects ($A_i$) are equal to zero:

$$H_0 : A_i = 0$$

Since we are interested in testing differences between groups, and the control group serves as the baseline, we are specifically testing:

$$\begin{gathered} H_0 : A_2 = 0 \\ H_0 : A_3 = 0 \end{gathered}$$

This is the test that R conducts in the output above. It tests, for the last two parameters, the hypothesis that the difference between Experiment 1 and the Control is zero an that the difference between Experiment 2 and the Control is zero. In both cases, the p-values are extremely small which suggest that there are differences (the effects are not equal to zero).

What about the intercept? This is testing that the mean of the Control group is zero. So, it tests whether the students in the control group scored zero on average. This doesn't really make sense and it is not a useful test. So not all tests that R carries out are necessarily useful or informative! Very often testing whether the intercept is different from zero is not interesting.

What if we aren't interested in the contrast R perform by default? We wanted to know whether there is a difference between the other two groups? We simply need to change the baseline treatment that R uses and we can do this easily using the `relevel` command:

```
m1.tc <- lm(Posttest ~ relevel(Group, ref ="Exp1"), data = multitask)
summary(m1.tc)
```

```
Call:
lm(formula = Posttest ~ relevel(Group, ref = "Exp1"), data = multitask)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.964 -10.175   0.583   8.550  37.408 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           62.882      2.237  28.111  < 2e-16 ***
relevel(Group, ref = "Exp1")Control   12.752      3.163   4.031 9.92e-05 ***
relevel(Group, ref = "Exp1")Exp2     -10.642      3.163  -3.364  0.00104 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.15 on 117 degrees of freedom
Multiple R-squared:  0.3191,    Adjusted R-squared:  0.3075 
F-statistic: 27.42 on 2 and 117 DF,  p-value: 1.717e-10
```

Notice that the residual standard error, F-value and other statistics at the end of the output are exactly the same as for Model m1.tc above. The two models are equivalent and provide the same fit to the data. The only difference is that the parameters have different interpretations.

To conclude this section, we present the final results of the social media multitasking experiment. The ANOVA revealed a significant treatment effect on academic performance ($F = 27.42$, $p = 1.72 \times e^{-10}$). Specifically, students in both experimental conditions performed worse than those in the control group. On average, students in Experiment 1 scored 12% lower ($t = -4.031$, $p = 9.92 \times 10^{-5}$), while those in Experiment 2 scored 24% lower ($t = -7.395$, $p = 2.32 \times 10^{-11}$), with a standard error of 3.163. Students in Experiment 2 scored on average 10% less than those in Experiment 1 ($t = -3.364$, $p = 0.001$). This confirms that multitasking with social media during lectures negatively impacted academic performance in this experiment.

I think the message is clear, going on social media during lectures is probably not going to help you learn. In general reducing the time you spend on social media will probably help you. You certainly don't have to delete all social media apps, but taking intentional breaks and trying to give your full attention when it is required, will certainly make a difference. Here are some videos that have motivated me to improve my focu and decrease my time spent on social media!

- Why we can't focus https://www.youtube.com/watch?v=6QltxZ-vPMc
- Quit social media https://www.youtube.com/watch?v=3E7hkPZ-HTk

# Part III

# Randomised Block Designs

# Chapter 8

# Introduction

So far, we have examined completely randomised designs where randomisation of experimental units to treatments was completely unrestricted. With complete randomisation, all other variables (the environment that we can never control completely) that might affect the response are, on average, equal in all treatment groups. This allows us to be confident that differences in group means are due to the treatments.

However, there is often important variation in additional variables that we are not directly interested in. If we can group our experimental units with respect to these variables to make them more similar, we achieve a more powerful design. This is the idea of blocking.

If blocks are used effectively, we can separate variability due to treatments, blocks, and errors, reducing unexplained variability. That is, variability between blocks can be estimated and removed from the residual error. Essentially, we compare treatments over more similar experimental units than in a completely randomised design. With reduced error variance, our test becomes more powerful.

Blocking is also useful when we want to demonstrate that treatment differences hold over a wider range of conditions. For example, in the social media multitasking example, the experiment was conducted on first year students. Strictly speaking, the results then only apply to first year students and extrapolation to students in different years of their degree is limited. Alternatively, we could choose students from first, second and third year (for example) and apply one replicate of each treatment within year. In this case, year of study would be the blocking factor.

More generally, we often want to show that our results hold for different species, age groups, or biological sexes. In such cases, we could use species, age, or sex as blocks. While blocks are typically used to control for variation in variables we are not directly interested in, sometimes these factors may also be of interest in their own right.

## 8.1 Treatments vs. Blocks

When is a factor a treatment, and when is it a block?

A good way to distinguish between them is by asking whether we can manipulate the factor and randomly assign experimental units to its levels.

- We generally cannot manipulate the age or sex of an individual, but we can manipulate, for example, the food they receive. So, age and sex are blocking factors, whereas food type is a treatment.

- We can manipulate the level of social media multitasking, but we cannot manipulate the year of study of students. So, level of social media multitasking is a treatment, while year of study is a block.

Although we can always estimate differences between blocks, we need to be much more cautious when inferring causality from block-level differences or from any factor that we cannot randomise (as is the case in observational studies).

Example of observational study

Suppose we are studying whether different music streaming platforms (e.g., Spotify, Apple Music, YouTube Music) influence a song's popularity. We cannot randomly assign a song to a particular streaming platform because artists typically release their music on multiple platforms simultaneously. However, platform choice is still the main factor of interest.

We would analyze differences in song popularity (e.g., number of streams, chart position) across platforms as we would for any treatment factor. However, we must be cautious when attributing differences solely to the platform itself because other factors could also play a role. For instance:

- Artist popularity: A well-known artist might naturally attract more streams, regardless of the platform.
- Marketing strategies: Some platforms might promote certain songs more aggressively.
- Release timing: Songs released during peak listening hours or days may perform better.
- Platform demographics: Different platforms cater to different audiences, which might influence engagement.

Since we cannot randomly assign songs to platforms, we cannot be certain that observed differences in popularity are only due to the platform. Instead, they may be influenced by a combination of these external factors. This is an observational study.

Suppose we study whether different teaching methods (interactive vs. traditional) affect student performance, conducted in public and private schools.

- Teaching method is the treatment (assigned to students).

- School type is a blocking factor (cannot be randomly assigned).

If private school students perform better, we cannot conclude school type caused the difference due to potential confounders such as socioeconomic background or teacher quality.

Even though we control for school type, observed differences may be due to these external factors, not just the school itself. We cannot be sure that the observed differences are really only due to school type.

Sometimes, however, blocking variables can also be randomised. Suppose a study is testing two medications (A vs. B) for blood pressure, experiments are conducted in two labs (Lab 1 & Lab 2).

- Medication is the treatment (randomly assigned).

- Lab is a blocking factor (controls lab-related variability).

Patients could have been randomly assigned to labs, but if logistical constraints prevent this, lab is used as a block. Since we only care about medication effects, lab differences are treated as a nuisance variable. The real difference is interest. We are not interested block effects on the response, only treatment effects. Blocking factors are used to control for known sources of variation that might obscure the treatment effect.

## 8.2 Choosing Blocking Factors

Any variable that might affect the response besides treatment factor should be considered for blocking. Common blocking factors include:

- Geographic location: field, site, regions or cities that share similar economic conditions.
- Time: experimental replication over different days or weeks. Blocking for economic cycles or seasonal effects.
- Subject: person, plant, businesses, phenotype.
- Demographic groups: age, gender, income or education level, consumer behavior segments.

- Equipment: container types, growth chambers.

For example, if we are testing the effectiveness of a new advertising campaign, it would be useful to block by city or region to control for differences in local economies, purchasing behavior, or media consumption. Similarly, if an experiment measures the impact of dynamic pricing on sales, it is a good practice to replicate the price changes across multiple days, blocking for daily or weekly variations in consumer spending habits. This way time accounts for these differences rather inflating the error variance.

Likewise, if we are studying the effect of sports training programs on player performance, and athletes train in different facilities with varying equipment, we could assign a block to each training center to ensure that facility-related differences are accounted for. This prevents training location from being mistaken as a treatment effect, allowing a clearer evaluation of the actual program's impact.

The key takeaway is that reducing error variance increases the power of the experiment. Thoughtful blocking design helps achieve this by accounting for known sources of variation.

## 8.3 Randomised Complete Block Design

There are a few different types of randomised block designs depending on the availability of experimental units and size of the blocks. Here we will consider the best case scenario, where blocks are big enough to contain an equal amount of experimental units such that *each treatment occcurs exactly once within a block*. If we have a single treatment factor with $a$ levels, then we have $a$ experimental units per block. This design is said to be *balanced*, each block is the same with respect to treatments. In balanced block designs, the treatment and block effects can be completely separated (are independent) . This greatly simplifies the interpretation of results.

As in CRD, randomisation is still a crucial component of the design. The difference is that now $a$ treatments are assigned randomly to the $a$ experimental units within a block, i.e. randomisation is not complete over ALL experimental units but restricted within each block. Within each block, the experimental units are equally likely to receive any of the $a$ treatments. You can see this as CRD within each block!

Let's see how we could randomise treatment within blocks using R. Imagine we had four treatments (A,B,C and D). We randomise the treatments to the units within one block like this:

```r
units <- 1:4
rbind(sample(units,4), rep(c("A","B","C","D")))
```

```
     [,1] [,2] [,3] [,4]
[1,] "3"  "2"  "4"  "1"
[2,] "A"  "B"  "C"  "D"
```

The third unit receive treatment A, the second receives B and so on. We then repeat this for every block.

## 8.4 The Pygmalion Effect

The **Pygmalion effect** is a psychological phenomenon that suggests when people are held to high expectations, they tend to perform better. This applies to, for example, teachers and students, managers and employees or coaches and athletes. It is named after a mythological king of Cyprus, Pygmalion, who fell in love with a sculpture he created of his ideal woman.

Many experiments have found results to support this type of self-fulfilling prophecy. Typically, they involve putting someone in charge of a group of people, then privately telling the leader that say a few of these people are exceptional (these people were randomly selected though). Then, later the performance of the group is measured and if the Pygmalion effect is present, the individuals who were marked as exceptional should have performed better.

This article explains the concept nicely and also briefly discusses the study we will use later. https://thedecisionlab.com/biases/the-pygmalion-effect

Back in 1990, one researcher in this field, noticed that experiments like these might involve something that is called interpersonal contrasts. When some individuals are singled out for high expectations, others might feel neglected. This could potentially skew the results by making the others look good even though it was just the rest that performed poorly. The researcher wanted to conduct an experiment to test the Pygmalion effect without interpersonal contrasts.

They achieved this by applying the high expectation to an entire group and not selected individuals within a group. Let's have a look the exact experiment!

Example: The Pygmalion Effect in Military Training

A study conducted by Eden (1990) examined whether raising leaders' expectations of their trainees would enhance performance, without creating interpersonal contrast effects.
A total of 10 army companies consisting of 2 platoons each were used in the study. Within each company, one randomly assigned platoon received the Pygmalion treatment, while the other two served as controls. The idea is that the assignment of the Pygmalion treatment to an entire platoon prevents interpersonal contrasts.

1. **Pygmalion Group:** Platoon leaders were informed that their trainees had exceptionally high command potential based on pre-existing evaluations.

2. **Control Group:** Platoon leaders received no expectation-enhancing information.

Over the training period, leaders in both conditions met biweekly with a psychologist to reinforce expectations. At the end of the program, soldiers took multiple tests which measured their performance in four areas:

- Theoretical specialty knowledge (taught by platoon leaders)

- Practical specialty skills (taught by platoon leaders)
- Physical fitness (assessed independently)
- Target shooting (assessed independently)

A platoon is a military unit typically consisting of 30 to 50 soldiers, led by a platoon leader (usually a lieutenant). Several platoons form a company, which is a larger military unit consisting of three to five platoons, commanded by a company leader (usually a captain).

First things first! We need to determine the design so we can use the appropriate analysis. The researcher was interested in determining the effect of the Pygmalion effect on performance. This indicates to use that there is a single treatment factor and that it is whether or not the Pygmalion effect was applied (we'll call this the Pygmalion Treatment) and the response is some measure of performance. The text gives four possible responses! The four areas in which the performance was tested. We'll start with the first one as our response. So far we know:

The name of the treatment factor is not always obvious. It is usually something that describes the collection of similar treatments created in response to some research hypothesis or what has been manipulated. In biological or ecological studies, it can be quite clear. For example, if we had treatments high, medium and low rainfall, "Rainfall" is the variable we manipulated.

Also, as before, I've modified the example slightly. In the original study, there three platoons per company with two serving as control and one company only had two. So we have simplified the design so that it is balanced.

- **Response Variable:** Theoretical specialty knowledge.
- **Treatment Factor:** Pygmalion Treatment.
- **Treatment Levels (Groups):** Control, Pygmalion
- **Treatments:** Control, Pygmalion

Now, the treatments were randomly assigned to platoons within a company. This gives away two things, (1) the experimental unit is an entire platoon and (2) treatments were assigned randomly within a company, i.e. a block! They were not interested in the effect of company on performance but merely wanted to account for possible differences between platoons in companies. So here Company is a blocking variable and the 10 companies are the blocks. Finally, on what was the response measured? The soldiers! They are then the observational units. The paper doesn't state how many soldiers were in each platoon and it doesn't really matter since the scores have to be combine to have one measurement per experimental unit.

Here is the final summary of the design:

- **Response Variable:** Theoretical specialty knowledge.
- **Treatment Factor:** Pygmalion Treatment.
- **Treatment Levels (Groups):** Control, Pygmalion
- **Treatments:** Control, Experiment 1, Experiment 2
- **Experimental Unit:** Platoon (20)
- **Observational Unit:** Soldier
- **Replicates:** 10 platoons received each treatment
- **Randomisation:** To platoons within companies, i.e. restricted to within blocks.

- **Design Type:** Randomised Complete Block Design (CRD)

You will typically have access to the data as well to help you identify some aspects of the design. There are two ways the data might be represented, in long format:

| Company | Treat | Score |
|---|---|---|
| C1 | Pygmalion | 80.0 |
| C1 | Control | 63.2 |
| C2 | Pygmalion | 83.9 |
| C2 | Control | 63.1 |
| C3 | Pygmalion | 68.2 |
| C3 | Control | 76.2 |
| C4 | Pygmalion | 76.5 |
| C4 | Control | 59.5 |
| C5 | Pygmalion | 87.8 |
| C5 | Control | 73.9 |
| C6 | Pygmalion | 89.8 |
| C6 | Control | 78.9 |
| C7 | Pygmalion | 76.1 |
| C7 | Control | 60.6 |
| C8 | Pygmalion | 71.5 |
| C8 | Control | 67.8 |
| C9 | Pygmalion | 69.5 |
| C9 | Control | 72.3 |
| C10 | Pygmalion | 83.7 |
| C10 | Control | 63.7 |

or in wide format:

| Company | Pygmalion | Control |
|---|---|---|
| C1 | 80.0 | 63.2 |
| C2 | 83.9 | 63.1 |
| C3 | 68.2 | 76.2 |
| C4 | 76.5 | 59.5 |
| C5 | 87.8 | 73.9 |
| C6 | 89.8 | 78.9 |
| C7 | 76.1 | 60.6 |

| | | |
|---|---|---|
| C8 | 71.5 | 67.8 |
| C9 | 69.5 | 72.3 |
| C10 | 83.7 | 63.7 |

The **long format** represents each observation as a separate row, with treatments recorded in a single column. This format is useful for statistical modeling and visualization since it keeps data structured for comparisons across treatments. You might have noticed that so far all the data sets we have used when fitting models using `aov` have been in long format. You will struggle to fit the model with a data set in wide format!

The **wide format** organizes data so that each unit (e.g., a company) appears in a single row, with treatments as separate columns. This format is often preferred for paired comparisons and summary tables.

Both formats contain the same information but serve different purposes depending on the type of analysis being performed.

# Chapter 9

# Assumptions

Data from randomised block designs are analysed with two-way ANOVAs. The assumptions of a two-way ANOVA are the same as a one-way. That is,

1. Equal population variance.
2. Normal errors.
3. Independent errors.
4. No outliers.

With the addition of a block variable comes a new assumption:

5. The effects of the blocks and treatments are additive.

Simply put, it means that the we assume the treatment effects are similar in all blocks. That if a treatment is applied in one block, the effect is the same as in another block. For example, if we applied had a treatment factor: marketing strategy with two treatments A and B, and we want to apply it to stores in different economic regions (blocks), the effect of, for example, marketing strategy A should be the same in both regions. We will check this assumption visually as well. One way, is to plot the response against the block for each treatment. But first! Some exploratory data analysis.

```r
# read in data
pyg_data <- read.csv("Datasets/pygmalion_data.csv")

# look at first and last few rows
head(pyg_data); tail(pyg_data)
```

```
  Company     Treat Score
1      C1 Pygmalion  80.0
2      C1   Control  63.2
3      C2 Pygmalion  83.9
4      C2   Control  63.1
```

```
5      C3 Pygmalion  68.2
6      C3   Control  76.2
```

```
   Company     Treat Score
15      C8 Pygmalion  71.5
16      C8   Control  67.8
17      C9 Pygmalion  69.5
18      C9   Control  72.3
19     C10 Pygmalion  83.7
20     C10   Control  63.7
```

```
summary(pyg_data)
```

```
   Company              Treat               Score      
 Length:20          Length:20          Min.   :59.50  
 Class :character   Class :character   1st Qu.:66.78  
 Mode  :character   Mode  :character   Median :73.10  
                                       Mean   :73.31  
                                       3rd Qu.:79.17  
                                       Max.   :89.80  
```

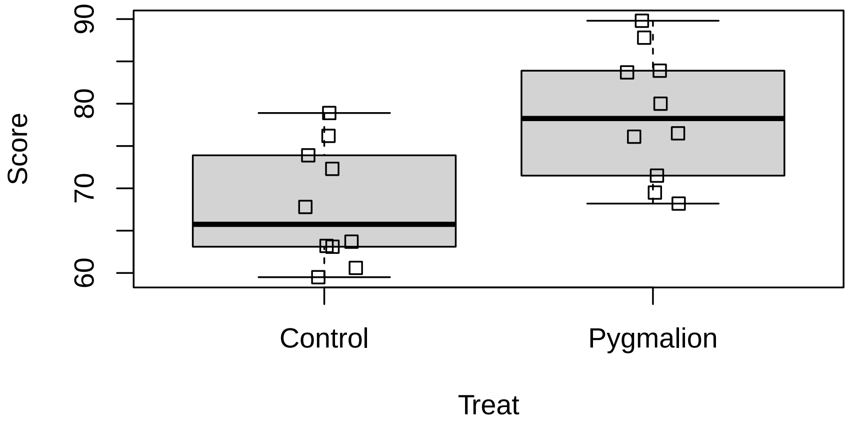
Ah! We need to convert both the company and Treat variable to factors.

```
pyg_data$Company <- as.factor(pyg_data$Company)
pyg_data$Treat <- as.factor(pyg_data$Treat)

summary(pyg_data)
```

```
    Company         Treat        Score      
 C1     :2   Control  :10   Min.   :59.50  
 C10    :2   Pygmalion:10   1st Qu.:66.78  
 C2     :2                  Median :73.10  
 C3     :2                  Mean   :73.31  
 C4     :2                  3rd Qu.:79.17  
 C5     :2                  Max.   :89.80  
 (Other):8                                 
```

Nice, now we can see that we had ten replicates per treatment, two observations per block which means 20 observations in total. Let's go ahead and check the the first four assumptions.

```
boxplot(Score~Treat, data = pyg_data)

stripchart(Score~Treat, data = pyg_data, add = TRUE, vertical = TRUE, method = "jitter"
```



Okay, the boxplots look relatively symmetric, there are no clear signs of non-normality. They also look very similar in terms of height, the assumption of homogeneity seems reasonable as well. Let's have a look at the sample standard deviations.

```
sd(pyg_data$Score[pyg_data$Treat == "Pygmalion"])
```

```
[1] 7.587124
```

```
sd(pyg_data$Score[pyg_data$Treat == "Control"])
```

```
[1] 6.927209
```

Then, we need to check the independence assumption. This is often the hardest assumption to verify because it requires knowledge about how the data were collected. In practice, you will need to assess independence in one of two ways:

1. Before conducting an experiment – Ideally, you would discuss the study design with the researchers before data collection to ensure that independence is maintained.
2. When analyzing existing data – If you are reviewing a published study, you must rely on the authors' description of the experimental setup to determine whether independence is reasonable.

In this study, the researchers assumed platoons operated independently and took steps to prevent treatment contamination:

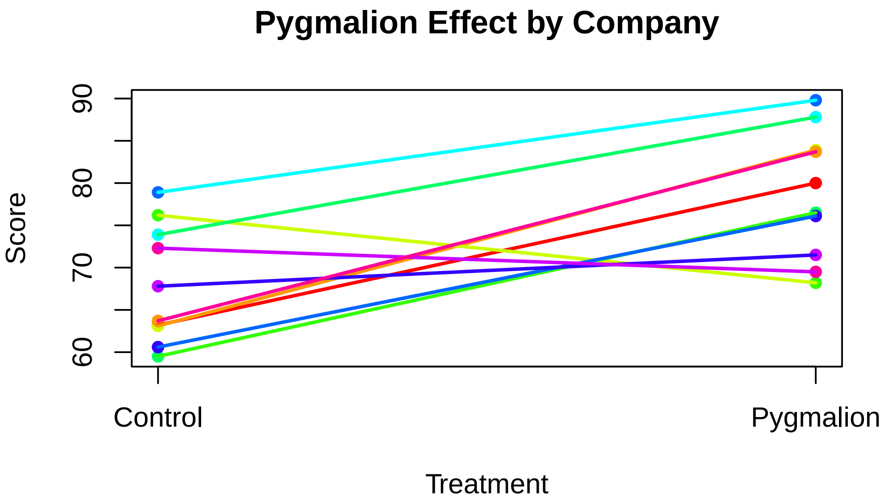
- Randomization ensured that each platoon was independently assigned to the Pygmalion or control condition, reducing bias.

- Leaders were instructed not to discuss their treatment condition, preventing expectation spillover.
- Each platoon was analyzed separately, ensuring observations within treatments were treated as independent.

After fitting the model, and assuming the order in which the response was measured is the order in which it appears in the data set, we can check for any pattern in the residuals that may indicate dependence.

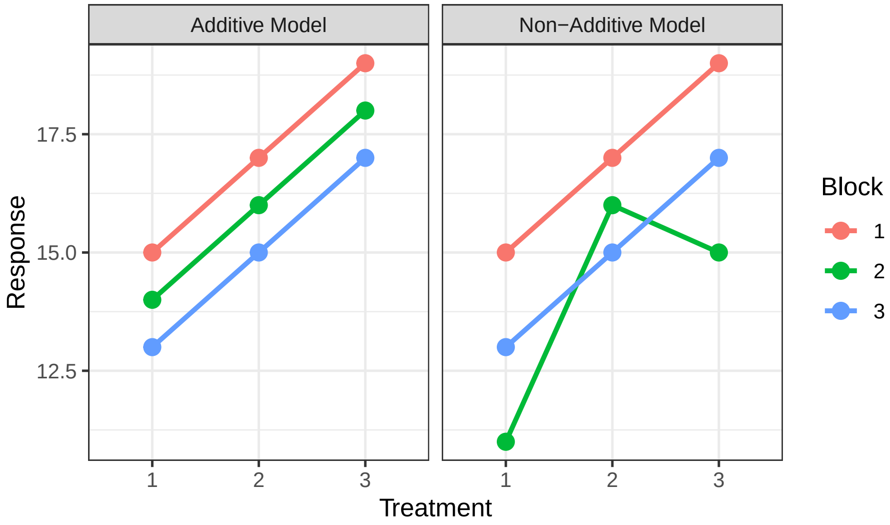
Now, let's check the new assumption of additivity. We can plot the response against treatment and add colour-coded lines connecting the experimental units from the same block. This is a bit tedious to do with base R (don't worry, we won't expect you to code this manually) but you have to understand and interpret the plot it produces.

```r
# Ensure Treat is a factor with proper order
pyg_data$Treat <- factor(pyg_data$Treat, levels = c("Control", "Pygmalion"))

# Convert Treat to numeric for plotting (1 = Control, 2 = Pygmalion)
pyg_data$Treat_numeric <- as.numeric(pyg_data$Treat)

# Define colors for companies
company_colors <- rainbow(length(unique(pyg_data$Company)))
names(company_colors) <- unique(pyg_data$Company)

# Create base plot
plot(pyg_data$Treat_numeric, pyg_data$Score,
     xlab = "Treatment", ylab = "Score",
     main = "Pygmalion Effect by Company",
     pch = 16, col = company_colors[pyg_data$Company], xaxt = "n")

# Add custom x-axis labels
axis(1, at = c(1, 2), labels = c("Control", "Pygmalion"))

# Add lines connecting observations from the same company
for(block in unique(pyg_data$Company)){
  temp <- pyg_data[pyg_data$Company == block, ]
  temp <- temp[order(temp$Treat_numeric), ]  # Order by treatment for correct line dra
  lines(temp$Treat_numeric, temp$Score, col = company_colors[block], lwd = 2)
}
```



If the assumption of additivity is met, we would expect relatively few lines to cross, i.e. we would expect mostly parallel lines. In the plot above, it seems that for most blocks, a low value in the control treatment is associated with a high value in the Pygmalion treatment. There are three lines (i.e. companies) that don't conform to this pattern, where it seems that Pygmalion treatment did not alter the scores or maybe even caused a reduction. It's important to remember that **sampling variability** prevents us from observing perfectly parallel lines in practice. The observed treatment means are always subject to **random variation**, which can introduce some deviations from the expected pattern.

What happens if this assumption is wrong, i.e. the blocking and treatment factors do interact? That is the treatment effect depends on which block it is in. Consider the following plots depicting an experiment with three treatments and three blocks. The first panel shows an example where treatments and blocks are additive – the lines connecting the same treatment in all blocks are parallel. Due to variability, we would of course never actually observe such parallel lines. In reality, the observed treatment means would be subject to random deviations from the true population means, and with lots of variability, the lines could cross and look more like the second panel, which is showing an example where the additivity assumption is violated.



With only one experimental unit per treatment in each block, as in a typical randomised complete block design, it is difficult to know which situation we have: the additivity assumption is violated, or there is simply a lot of random error. The interaction effect between treatment and block is confounded with the random error term. That is, $e_{ij}$ in the model equation is actually the sum of the interaction effect and the random error. So if the additivity assumption is violated, $e_{ij}$ is inflated and it will be harder to find differences between treatments.

With some replication of treatments within blocks, as in generalised randomised complete block designs (which we don't cover here), we can separately estimate the interaction effects. This is similar to what we will see when we talk about Factorial Experiment sin the next section.

# Chapter 10

# Linear model & ANOVA

## Linear model

We wish to compare $a$ treatments and have $N$ experimental units arranged in $b$ blocks each containing $a$ homogeneous experimental units: $N = ab$. The $a$ treatments are assigned to the units in the $j^{th}$ block at random. The design (blocking and treatment factors and the randomisation) determine the structural part of the model.

A linear model for the RBD is:

$$Y_{ij} = \mu + A_i + B_j + e_{ij}$$

where,

$$\begin{aligned}
Y_{ij} &= \text{observation on treatment } i \text{ in block } j \\
i &= 1, \dots, a \text{ and } j = 1, \dots, b \\
\mu &= \text{general/overall mean} \\
A_i &= \text{effect of the } i^{th} \text{ treatment} \\
B_j &= \text{effect of the } j^{th} \text{ block} \\
e_{ij} &= \text{random error with } e_{ij} \sim N(0, \sigma^2) \\
\sum_{i=1}^{a} A_i &= \sum_{j=1}^{b} B_j = 0
\end{aligned}$$

This model says that each observation is made up of an overall mean, a treatment effect, a block effect, and an error part. The block effect is interpreted in the same way as the treatment effect, it is the difference between block mean $j$ and the overall mean $\mu$.

It also says that these effects are additive. Additivity means that the effect of the $i^{th}$ treatment on the response $(A_i)$ is the same regardless of the block in which the treatment is used. Similarly, the effect of the $j^{th}$ block is the same $(B_j)$ regardless of the treatment. The additional constraint of $\sum_{j=1}^{b} \beta_j = 0$ follows the same logic as explained before.

Let's fit this model to the Pygmalion data. For the Pygmalion experiment, the researchers compared control to Pygmalion treatment so we have $a = 2$ treatments. The number of blocks, $b$, was 10. In R, on the right-hand-side of the formula, we have the treatment factor + blocking factor. The code looks exactly the same as before, except we **add** the Company (blocking) variable.

```
pyg_model <- aov(Score ~ Treat + Company, data = pyg_data)
```

We can again extract the model estimates with `model.table`:

```
model.tables(pyg_model, type = "means", se = TRUE)
```

```
Tables of means
Grand mean

73.31

 Treat
Treat
  Control Pygmalion
    67.92     78.70

 Company
Company
   C1   C10    C2    C3    C4    C5    C6    C7    C8    C9
71.60 73.70 73.50 72.20 68.00 80.85 84.35 68.35 69.65 70.90

Standard errors for differences of means
        Treat Company
        3.126   6.990
replic.    10       2
```

First we see the grand mean of 73.31 followed by the treatment means. Then, we have ten block means, these are the mean scores within each block. Lastly, we see the standard errors for the differences of means.

## Sums of squares and Analysis of variance

Now we have three sources of variability: differences between treatments, differences between blocks and experimental error. The total sum of squares can be split into three sums of squares: for treatments, blocks, and error respectively.

$$SS_{total} = SS_A + SS_B + SSE$$

with degrees of freedom

$$(ab - 1) = (a - 1) + (b - 1) + (a - 1)(b - 1)$$

The advantage of blocking becomes apparent here. If we had not blocked, i.e. used a completely randomised design, for example, the $SS_A$ (sums of squares for treatment) would be the same. However, in the completely randomised design, we would not be able to separate $SS_B$ from $SSE$ and the combined $SSE$ would therefore be larger.

When using a RBD, part of the unexplained variation is now explained and can be captured in the block sum of squares, $SS_B$. A small $SSE$ has the advantage of smaller standard errors, i.e. more precise estimates (for treatment effects and treatment means) and thus it is easier to detect differences between treatments.

The sums of squares are summarised in an ANOVA table.

You can think of the SSE as the variability among experimental units that cannot be accounted for by blocks or treatments.

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Treatments A | $SS_A = b\sum_i(\bar{Y}_i - \bar{Y}_{..})^2$ | $(a-1)$ | $\frac{SS_A}{(a-1)}$ | $\frac{MS_A}{MSE}$ |
| Blocks B | $SS_B = a\sum_j(\bar{Y}_j - \bar{Y}_{..})^2$ | $(b-1)$ | $\frac{SS_B}{(b-1)}$ | |
| Error | $SSE = \sum_{ij}(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$ | $(a-1)(b-1)$ | $\frac{SSE}{(a-1)(b-1)}$ | |
| Total | $SS_{total} = \sum(Y_{ij} - \bar{Y}_{..})^2$ | $ab-1$ | | |

Much of the table remains the same as in a one-way ANOVA, but now it includes an additional row for the blocking variable. The sum of squares for the blocking factor is calculated similarly to that of the treatment factor—by summing the squared deviations of observations within each block from the block's mean response. The residual sum of squares is also slightly different, but you don't need to worry too much about that[1]. Since the total SS is simply the sum of the treatment, block, and residual SS, you can always compute SSE by subtraction.

From this ANOVA table, we can test the hypothesis of no differences between the treatment means as before.

$$H_0 : \mu_! = \mu_2 = ... = \mu_a$$

Which is equivalent to testing:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$$

using the F-test which compares the mean square for treatments with the mean square for error:

$$F = \frac{MS_A}{MSE} \sim F_{a-1,(a-1)(b-1)}$$

Notice the degrees of freedom!

What does the ANOVA table look like for the Pygmalion data? Again, we use the `summary` function on the model object to obtain the table.

```r
summary(pyg_model)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)   
Treat        1  581.0   581.0   11.89 0.00729 **
Company      9  510.2    56.7    1.16 0.41433   
Residuals    9  439.8    48.9                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's first look at the degrees of freedom:

- We had $a \times b = 2 \times 10 = 20$ experimental units, so we should have $a \times b - 1$ total degrees of freedom.
- The treatment degrees of freedom are $a - 1 = 2 - 1 = 1$ and similarity, there are $b - 1 = 10 - 1 = 9$ degrees of freedom of the block effect.
- That leaves $(a - 1)(b - 1) = 1 \times 9$ degrees of freedom for the error term.

It is always a good idea to check that the df's match with what you expected them to be. One serious error that happens easily is one of the factor is fitted as continuous covariate because the levels were labelled using numbers. Hence why we converted the categorical variables (Treat and Company) to factors.

[1] You can easily get there by rearranging the model equation so that $e_{ij}$ is on the right-hand-side, replacing the effects with the difference in terms of means and simplifying.

The next column lists the sums of squares for the three components. The mean squares are calculated as $\frac{SS}{df}$, e.g. $\frac{581}{1} = 581$ for the treatment. The $F$-value for the treatment variable is the ratio of $MS_{treat}$ to the $MSE$. This is the same as in the one-way ANOVA. R then looks up the corresponding p-value, which is 0.00729.

So we have an very small p-value which means that we have strong evidence against our null hypothesis of equal treatment means. We cannot conclude that the effects are equal to zero. There is evidence to suggest that at least one treatment resulted in a different mean score. Here, because we have two treatments, the results indicate that the data are not compatible with a null hypothesis of equal means. We make the following conclusion:

"There is evidence to suggest that the two treatment means are different ($F = 11.89$, $p = 0.0073$)."

If we had more than two treatment means as is usually the case, we would conclude:

"There is evidence to suggest that at least one treatment resulted in a different mean response, there is evidence for a treatment effect ($F = 11.89$, $p = 0.0073$)."

There are many different ways to say that there is a difference somewhere. For example:

- One or more treatments had a mean response that differed from the others.
- Not all treatment means are the same; at least one is significantly different.
- The results indicate that not all treatment means are equal.

You get the idea! As long as it is clear that a 'significant' result indicates that there is a difference somewhere, we don't know where, but there is evidence for a treatment effect.

## What about the F-test for the blocking variable?

We see that the blocks accounted for a similar fraction of the sums of squares as the other two components (just over a third). If we did not block, this variation would be part of the SSE but then the error degrees of freedom would also be larger (the 9 degrees of freedom would be part of the error degrees of freedom). In fact, here, the blocking did not significantly reduce the unexplained variability, since the F-value is close to one. The variability explained by the blocks is close to what would be expected due to random noise.

Remember, we aren't particularity interested in formal inference about block effects (we knew or suspected that they were different) and we should always be careful about interpreting the F-test for the blocking variable (as blocks typically cannot be randomised to experimental units - see the previous chapter). We might, however, be interested in whether blocking increased the efficiency of the design by reducing the unexplained variation (SSE). There exists a more thorough method of assessing the relative efficiency of blocking - that is, relative to if a simpler design (i.e. CRD) was used instead[2]. Here, however, we focus on a simple and quick check of block efficiency using the F-ratio.

We would like the block factor to explain a lot of variation. If the mean square of the blocking variable is larger than the error mean square we conclude that blocking was effective (compared to a CRD).

- If $F > 1$ then blocking did reduce unexplained error variance.
- If $F \approx 1$ then the blocks did not improve the power of the experiment and you would have been equally well off with a CRD.
- If $F < 1$ which happens rarely, it means that blocking did not account for much of the variability because experimental units within blocks are more heterogeneous than between blocks (or there are strong interactions between blocks and treatments). Blocks actually reduced the power of the experiment but this should really not happen if you choose your blocks sensibly.

If blocking was not efficient, we would still leave the block factor in the model (**design dictates analysis**), but we might decide not to use blocking in a similar experiment in the future because it didn't assist in reducing experimental error variance and only cost us degrees of freedom.

# Estimation

To obtain estimates for the treatment and block effects, we minimize the error sum of squares (method of least squares).

$$SSE = \sum_i \sum_j (Y_{ij} - \mu - A_i - B_j)^2$$

$Y_{ij} - \mu - \alpha_i - \beta_j$ is the observed value minus the expected value (the structural part of the model). This difference is just the error $e_{ij}$. If we minimise the error sum of squares we obtain the following estimates:

$$\hat{\mu} = \bar{Y}_{..}$$

$$\hat{A}_i = \bar{Y}_{i.} - \bar{Y}_{..}, \quad i = 1 \dots a$$

$$\hat{B}_j = \bar{Y}_{.j} - \bar{Y}_{..}, \quad j = 1 \dots b$$

[2] Kuehl (2000) wrote a great textbook (freely available) that explains the relative efficiency check in detail.

To estimate the $i^{th}$ treatment effect we take the observed treatment mean minus the overall mean, similarly to obtain block effect estimates.

# Chapter 11

# Contrasts

After finding that there is evidence to suggest that at least two of the treatments differed from each other (another way to say there is a treatment effect!), we want to find out which ones differed and estimate these differences. In the Pygmalion experiment, there were only two treatments so we know the difference is between them. In fact, we could have used a paired t-test to analyse the data and we would get the same results. Generally though, there will be more than two treatments and then after concluding that there are differences, we want to know where the differences lie.

To do that, we use the coefficients from fitting a linear regression model to estimate the difference between the two treatments (as we did for CRD experiments as well). The null hypothesis is again:

$$H_0 : \mu_C - \mu_P = 0$$

where C stands for Control and P for Pygmalion. We use the `lm` function as in regression.

When we have blocks in RCBD, the observations are paired and data from two treatments can be analysed using a paired t-test. When we do not have blocks, the observations are not and data from two treatments can be analysed using a standard t-test.

```
pyg_model_reg <- lm(Score ~ Treat + Company, data = pyg_data)
summary(pyg_model_reg)
```

```
Call:
lm(formula = Score ~ Treat + Company, data = pyg_data)

Residuals:
   Min     1Q Median     3Q    Max 
-9.390 -3.217  0.000  3.217  9.390 

Coefficients:
```

We use the `lm` function when we are primarily interested in the coefficient estimates and difference. We use `aov()` when we want a breakdown of how much each factor can explain of the overall variation in the response, and when we want a general test for ‘are there *any* difference between the treatments’.

```
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      66.210      5.184  12.771 4.52e-07 ***
TreatPygmalion   10.780      3.126   3.448  0.00729 ** 
CompanyC10        2.100      6.990   0.300  0.77069    
CompanyC2         1.900      6.990   0.272  0.79191    
CompanyC3         0.600      6.990   0.086  0.93348    
CompanyC4        -3.600      6.990  -0.515  0.61897    
CompanyC5         9.250      6.990   1.323  0.21839    
CompanyC6        12.750      6.990   1.824  0.10147    
CompanyC7        -3.250      6.990  -0.465  0.65303    
CompanyC8        -1.950      6.990  -0.279  0.78659    
CompanyC9        -0.700      6.990  -0.100  0.92243    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.99 on 9 degrees of freedom
Multiple R-squared:  0.7127,    Adjusted R-squared:  0.3936 
F-statistic: 2.233 on 10 and 9 DF,  p-value: 0.1211
```

A few things to note here.

1. We are only interested in the second line starting with `TreatPygmalion`. R pasted the name of the factor and the name of the treatment level together.

2. The Control treatment was taken as the baseline (it comes first in the alphabet).

3. The remaining lines are not of interest to us. It estimates the differences between each block and the first and tests whether these effects are different from zero. R doesn't know we aren't interested in these, so it computes the effects and hypothesis test as if we are. We ignore this part.

4. If we didn't know that this code was for analysing a RCBD we would probably think that it is linear regression with two categorical variables. Think back to the regression module, what does this intercept represent? It represents the mean score for some treatment level and company. Since 'C' comes before 'P' in the alphabet and C1 is before everything else, the intercept is the average score for the control group in the first block. This is because R uses the *treatment contrast* parameterisation by default for all the factors. We can change this by letting R know that the block effects sum to zero.

```
pyg_model_reg2 <- lm(Score ~ Treat + C(as.factor(Company), contr.sum), data = pyg_data
summary(pyg_model_reg2)
```

```
Call:
lm(formula = Score ~ Treat + C(as.factor(Company), contr.sum), 
    data = pyg_data)
```

```
Residuals:
   Min     1Q Median     3Q    Max 
-9.390 -3.217  0.000  3.217  9.390 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         67.920      2.211  30.725 2.01e-10 ***
TreatPygmalion                      10.780      3.126   3.448  0.00729 ** 
C(as.factor(Company), contr.sum)1   -1.710      4.689  -0.365  0.72379    
C(as.factor(Company), contr.sum)2    0.390      4.689   0.083  0.93554    
C(as.factor(Company), contr.sum)3    0.190      4.689   0.041  0.96857    
C(as.factor(Company), contr.sum)4   -1.110      4.689  -0.237  0.81818    
C(as.factor(Company), contr.sum)5   -5.310      4.689  -1.132  0.28675    
C(as.factor(Company), contr.sum)6    7.540      4.689   1.608  0.14232    
C(as.factor(Company), contr.sum)7   11.040      4.689   2.354  0.04300 *  
C(as.factor(Company), contr.sum)8   -4.960      4.689  -1.058  0.31775    
C(as.factor(Company), contr.sum)9   -3.660      4.689  -0.780  0.45514    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.99 on 9 degrees of freedom
Multiple R-squared:  0.7127,    Adjusted R-squared:  0.3936 
F-statistic: 2.233 on 10 and 9 DF,  p-value: 0.1211
```

Now, the intercept is what we expect, the mean score for the control treatment across all blocks.

```r
mean(pyg_data$Score[pyg_data$Treat == "Control"])
```

```
[1] 67.92
```

Let's interpret the hypothesis test for the difference between the treatment means. The estimated difference is 10.78 with a standard error of 3.126. The test statistic is 3.448 which has a p-value of 0.00729. Look familiar? It's the exact same p-value we found in the ANOVA table! That is because the ANOVA is an extension of the t-test to more than two groups and when we only have two treatments, they are equivalent. In fact, the test statistics have the following relationship:

$$t^2 = F$$

Test the result to confirm that it holds. Now, let's return to the interpretation. The test shows that the difference between the control and Pygmalion treatment is statistically significant, as indicated by the extremely small p-value. This provides strong evidence against the null hypothesis of equal means.

To recall the experiment's design: The researchers aimed to test the Pygmalion effect while eliminating interpersonal contrasts by assigning treatments to entire groups. Specifically, platoons within companies were used as treatment units, and since there were 10 companies, each with 2 platoons, companies served as blocks. The response variable, theoretical specialty knowledge, was measured through test scores.

The results of a two-way ANOVA provide evidence of a treatment effect ($F = 11.89$, $p = 0.0073$). More precisely, the estimated difference between the control and Pygmalion treatment was 10.78 (s.e. $= 3.13$, $t = 3.45$, $p = 0.0073$). This suggests that the Pygmalion effect was successful, as soldiers in the Pygmalion group scored higher on average than those in the control group.

Nice! We're done. Before we move on, I'll summarise the results of the actual study. The researcher had four different responses:

In an actual analysis, we would not report both the ANOVA and t-test since they are equivalent when we have two treatments.

- Theoretical specialty knowledge (taught by platoon leaders)
- Practical specialty skills (taught by platoon leaders)
- Physical fitness (assessed independently)
- Target shooting (assessed independently)

Significant treatment effects were found for theoretical and practical specialty scores ($F = 13.74$, $p < 0.01$ and $F = 6.37$, $p < 0.05$, respectively). No significant difference was found for physical fitness or target shooting, confirming that the Pygmalion effect was specific to areas influenced by leader expectations! This suggests that high expectations from others can enhance performance, particularly in areas where they have direct influence. With this in mind, I want you to know that **I believe in your potential to excel in this course and expect nothing less. ;)** On to the next section!

# Part IV

# Factorial Experiments

# Chapter 12

# Introduction

So far, we have explored experiments with a **single treatment factor**. However, in many cases, analyzing factors one at a time does not fully explain the behavior of the response variable. This is particularly true when factors interact, meaning that the effect of one factor depends on the level or setting of another factor.

A factorial experiment involves more than one treatment factor, allowing us to study how factors interact. In a complete factorial experiment, every possible combination of factor levels is tested. The total number of treatments is the product of the number of levels for each factor. In other words, each treatment is a combination of one level from each factor.

## 12.1 Factorial Structure vs. Experimental Design

It is important to note that a **factorial experiment is not a design by itself**—it is a treatment structure. The underlying design can be:

- A Completely Randomized Design (CRD)
- A Randomized Complete Block Design (RCBD)

In the social media multitasking example, suppose the researchers wanted to know whether the effect of social media multitasking on academic performance is mitigated by lecture format? We would ask:

> *Does the effect of social media multitasking on academic performance depend on lecture format?*

The experiment would still follow a Completely Randomized Design (CRD) but now with two treatment factors instead of one.

Similarly, if we extended the Pygmalion experiment to include an additional factor, we would have an RCBD with two treatment factors.

## 12.2 Notation and Structure of Factorial Experiments

In general, if an experiment has two treatment factors with $a$ and $c$ levels, then there are $a \times c$ treatments. This is called an $a \times c$ factorial treatment structure.

To clarify the terminology:

- A **treatment factor** has **different levels** (e.g., social media multitasking: *none, texting, Facebook*).
- **Treatments** are the **combinations** of factor levels (e.g., *no multitasking + lecture format A*, *texting + lecture format B*).

In factorial experiments, the treatment factors are said to be *crossed*, meaning that all levels of one factor appear at all levels of the other factor.

## 12.3 Randomisation in Factorial Experiments

Randomization in factorial experiments depends on the chosen design and is carried out similarly to single-factor experiments. In R, it is helpful to number or name the treatments systematically.

Suppose we have two factors:

- Marketing Strategy (2 levels: $m_0, m_1$)
- Product Promotion (2 levels: $p_0, p_1$)

This creates **four treatments**:

$m_0p_0$, $m_0p_1$, $m_1p_0$, $m_1p_1$

If we have 12 experimental units and no need for blocking, we conduct a Completely Randomized Design (CRD) as follows:

```
treats <- c("m0p0", "m0p1", "m1p0", "m1p1")
treats <- rep(treats, each = 3)  # Repeat each treatment 3 times
treats
```

```
 [1] "m0p0" "m0p0" "m0p0" "m0p1" "m0p1" "m0p1" "m1p0" "m1p0" "m1p0" "m1p1"
[11] "m1p1" "m1p1"
```

```
r1 <- sample(treats)  # Randomly assign treatments

cbind(1:12, r1)  # Display the assignments
```

```
           r1
 [1,] "1"  "m0p1"
 [2,] "2"  "m1p0"
 [3,] "3"  "m0p0"
 [4,] "4"  "m0p0"
 [5,] "5"  "m0p0"
 [6,] "6"  "m1p1"
 [7,] "7"  "m0p1"
 [8,] "8"  "m1p1"
 [9,] "9"  "m1p1"
[10,] "10" "m1p0"
[11,] "11" "m1p0"
[12,] "12" "m0p1"
```

This code assigns treatments randomly and prints the experimental unit number alongside its assigned treatment. If we had blocking, we would repeat the randomization separately for each block.

## 12.4 Is comprehension affect by playback speed and lecture modality?

In keeping with the theme of students, learning and teaching. Have you ever wondered whether playback speed affects your comprehension of a lecture? Or whether your comprehension is better with audio-only lectures such as podcast versus recorded lectures with visuals? What about if you listen to a podcast at double speed versus a recorded lecture at double speed, is there difference in comprehension? to answer this question, researchers from the University of California conducted a 2 × 2 factorial experiment.

> Lecture modality and playback speed
>
> Chen et al. (2024) conducted an experiment to find out whether visual information improves comprehension when lectures are played at faster speeds. Specifically, they wanted to investigate the effect of *lecture modality* (audio-only or audio-visual) and *playback speed* (1x or 2x) on comprehension of students and whether these factors interact. We can summarise the research questions as follows:
>
> 1. Does lecture modality have an effect on comprehension?
> 2. Does playback speed have an effect on comprehension?
> 3. Is there an interaction effect of modality and playback speed on comprehension?
>
> A total of 200 undergraduate students were randomly assigned to one of four groups:
>
> 1. Audio-only at normal speed (1x)
> 2. Audio-visual (with slides) at normal speed (1x)
> 3. Audio-only at double speed (2x)

4. Audio-visual (with slides) at double speed (2x)

The researchers chose two lectures: one about about real estate appraisals and another bout the history of the Roman Empire. The lectures were either presented as audio-visual clips which consisted of presentation slides and instructor images, and no subtitles or captions were provided. All the graphics (maps, figures) in the slides were static. For lectures presented as audio-only clips, only the instructor's audio was made available.

Each student was presented both lectures in the modality and speed they were assigned. Afterwards, they completed a comprehension test consisting of 25 multiple choice questions on each topic. The average of the scores was taken as the final measure of comprehension.

Right! Let's begin with identifying the design. It should be clear that we have two treatment factors: *lecture modality* and *playback speed* each with the treatment levels. this means that we have a total of $2 \times 2 = 4$ treatments which are the combinations of the treatment levels. They investigated the effect of these factors on the comprehension of students - that means, comprehension is the response.

We will be using the actual data collected but we will only be using a subset of the information they recorded. The authors conducted a different analysis which incorporates this extra information. We will not be doing this as the method they used is outside th scope of this course.

- **Response Variable:** Comprehension
- **Treatment Factors:** Lecture modality and playback speed
- **Treatment Levels:** Lecture modality: Audio-only or Audio-visual; Playback speed: 1x or 2x
- **Treatments:** Audio-only at normal speed (1x); Audio-visual at normal speed (1x); Audio-only at double speed (2x); Audio-visual at double speed (2x)

Each student was assigned to one the treatments indicating that students were the experimental units. The response was also measured on each student, they are the observational units as well. Therefore, since we had 200 students and 4 treatment groups, there was 50 students per group, the experiment had 50 replicates.

- **Experimental Unit:** Student (200)
- **Observational Unit:** Student (200)
- **Replicates:** 50 students per group

Lastly, we need to determine how randomisation was conducted. There is no indication of any blocking and treatments were randomised to the whole group of experimental units. So this is a Completely Randomised Design, specifically it is a $2 \times 2$ factorial CRD.

- **Design Type:** $2 \times 2$ factorial Completely Randomized Design (CRD)

Before we do any further analysis, we need to talk a bit about effects!

# Chapter 13

# Interactions

Interactions between treatment factors are an important reason for conducting factorial experiments. If the effect of a factor would always be the same, no matter which other factors are present, and at what levels, it would be enough to investigate this factor on its own in a single-factor experiment. However, many factors interact with other factors, which means that the effects change, depending on the levels of the other factors.

Up until now, we have spoken rather loosely about 'effects'. But at this point, we need to define more clearly what we mean by the effect of a treatment or the effect of an explanatory variable. By the effect of a treatment, we mean the change in response relative to either a reference or baseline treatment, or often in experiments, to an overall mean response.

In regression, the effect of a continuous explanatory variable is measured by the slope, which is the change in response for a one-unit increase in the explanatory variable, i.e., relative to one unit less. The effect of categorical or factor variables in regression is the change in response relative to a reference category.

In experiments, when using an ANOVA model, the effect of a treatment is mostly measured as the change in response relative to an overall mean response.

There are different kinds of effects: **main effects, interaction effects, and random effects**.

The **main effect** of a treatment measures the average change in response, averaged over all levels of the other factors, relative to the overall mean. When there is only a single factor in an experiment, we only have main effects.

If the effect of a factor depends on the level of another factor that is present, then the two factors interact. The **interaction effect** represents the change in response relative to the main effects with a particular treatment.

If there are multiple factors in an experiment, and the effects of one factor depend on the level of the other factor, i.e., the two factors interact, the (average) effects might not give a good idea of changes in the response, or of how the factors affect the response. In such cases, we need to study the individual treatments more closely. We look at the combinations of factor levels with large interaction effects.

The figure below illustrates a number of possible interaction situations in a $2 \times 2$ factorial experiment, with treatment factor A having levels a1 and a2, and factor B having levels b1 and b2. To determine whether main effects of A are present, we must ask whether the average response changes when moving from a1 to a2, and similarly for main effects of B.

To determine whether **interaction effects** are present, we must ask whether the change in response when moving from a1 to a2 depends on the level of B. Main effects and interaction effects can both be present simultaneously.

Before we do anything, orient yourself. What is represented by each axis, how is the graph differentiating between treatment factors? In the figure below, the response is on the y-axis and we have the levels of treatment factor A on the x-axis, the levels of B are denoted by the colour of the line.

Let's start with (a). The main effect of A is the average change in response - averaged over all levels of the other factors. Essentially, we need to determine what happens to the response when we ignore the levels of B. To do this, we have to calculate the average of the points at a1 and separately, at a2. When points in a line are eveny distributed, the average is the mid-point. If you do this for both levels of A and connect the dots, you will have drawn a flat horizontal line. Now, going from a1 to a2, what happens to the response? In other words, does the average response change? No, there is n change which means there is no main effect of factor A.

Going through the same motions for factor B, reveals that going from b1 to b2 increases the mean response. There is a main effect of factor B. What about an interaction effect? We ask: does the effect of A on the response change when we consider the levels of the other factor? now we do not avearge over the other factor, we take it into account. Looking at the plot, if we focus on the red line and go from a1 to a2, nothing happens to the response. If we focus on the black line, the same (i.e. nothing) happens to the response as well. If we reverse this and focus on the points at a1, going from b1 (red) to b2 (black) increases the response. At a2, the response increases as well by the same amount. So nothing changed when we considered the other factors, there is no interaction.

Now, consider plot (e).

- Averaging the response at a1 and a2 results in a horizontal line again. No main effect of factor A.
- Averaging the response at b1 and b2, leads to the same conclusion as before. Going from b1 to b2, increases the response. There is a main effect of factor B.

- If we focus on the red line (b1), going from a1 to a2 increases the response. However, at b2, going from a1 to a2 decreases the response! The effect of treatment factor A depends on the level of B, they interact with each other.

So does this mean that A and B have no effect on the response? No, they both affect the response, their effects, however, depend on the level of the other factor present.

Plot (e) demonstrates clearly why sometimes main effects cannot be understood or interpreted when interactions are present. In such a case, the interaction plot is very helpful to illustrate the effects. Try deciding for the other plots whether there are main effects for factor A and B and whether A and B interact. The answers are given in the figure caption.

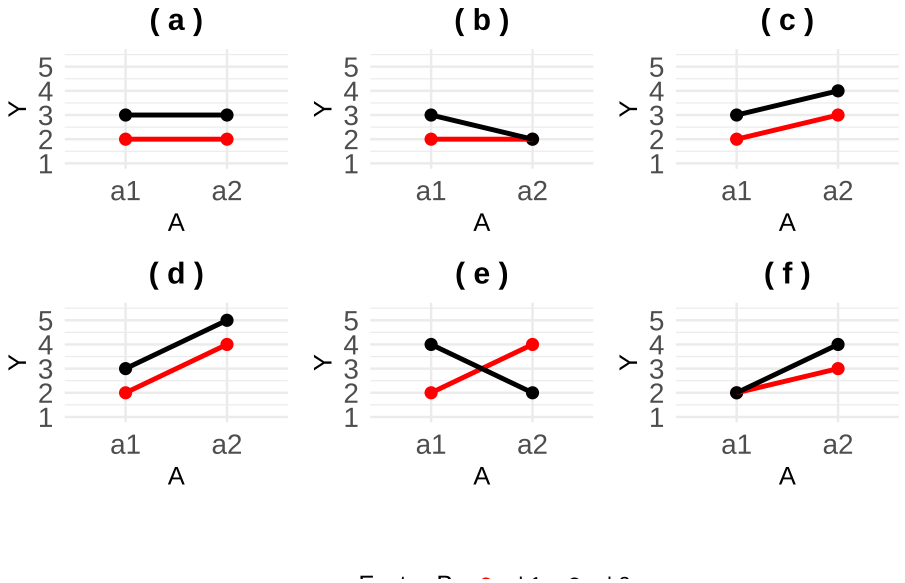


Figure 13.1: Interaction plots for six hypothetical 2 × 2 factorial experiments. (a) only main effect of B, (b) no main effects and no interaction effects, (c) only main effect of A, (d) main effect of A and main effect of B, (e) interaction between A and B, but both main effects are 0, (f) main effect of B, small main effect of A, A and B interact.

Interaction effects are calculated as the difference between the treatment mean and the sum of the main effects. To express this more precisely, it is useful to write down the model.

## 13.1 Can going for a brief walk help with memory performance?

We often hear about the benefits of exercise for physical health, but what about its impact on learning and memory? Would taking a brisk 10-minute walk before studying help us to remember more?

Researchers set out to explore this question by testing whether a short bout of exercise before learning could enhance memory performance. They had students either walk or sit before studying a list of words and then predict how well they would remember them. Later, the students took a recall test to see how much they actually retained.

Before studying, some students took a 10-minute brisk walk, while others remained seated and inactive. After this, everyone studied a list of words and rated how well they thought they would remember them (Judgements of Learning, or JOLs). Later, they took a free recall test to see how many words they actually remembered. The researchers wanted to find out if walking before studying could boost memory and whether students were aware of any benefits.

Warning

Salas, Minakata, and Kelemen (2011) conducted a study to examine whether a brief bout of aerobic exercise influences memory performance and judgements of learning (JOLs).
A total of 80 college students participated in a 2 × 2 factorial between-subjects design where they were randomly assigned to one of four conditions:
Walking-Walking: Participants walked before both encoding and retrieval. Walking-Sitting: Participants walked before encoding but sat before retrieval. Sitting-Walking: Participants sat before encoding but walked before retrieval. Sitting-Sitting (Control): Participants sat before both encoding and retrieval.
After the activity, all students studied 30 English nouns, provided immediate JOLs, and then took a free recall test

# Chapter 14

# Model for Factorial Experiments

Assuming we have a continuous response variable for which we assume a normal distribution, no blocking factors and a factorial experiment with two treatment factors, the following model is plausible:

$$Y_{ijk} = \mu + A_i + C_j + (AC)_{ij} + e_{ijk}$$

where $Y_{ijk}$ is the $k^{th}$ observation on the $(ij)^{th}$ treatment combination and

$$\begin{aligned}
Y_{ij} &= \text{observation on treatment } i \text{ in block } j \\
\mu &= \text{general/overall mean} \\
A_i &= \text{main effect of the } i^{th} \text{ level of A} \\
A_i &= \text{main effect of the } j^{th} \text{ level of C} \\
(AC)_{ij} &= \text{interaction between the } i^{th} \text{ level of A and the } j^{th} \text{ level of C.} \\
e_{ijk} &= \text{random error with } e_{ijk} \sim N(0, \sigma^2)
\end{aligned}$$

**Note that (AC) is a single symbol and does not mean the interpaction is the product of the two main effects.**

$\mu + A_i + C_j + (AC)_{ij}$ is the structural part of the model which describes the mean or expected response with treatment $ij$, i.e. at the $i^{th}$ level of factor A and $j^{th}$ level of factor C. Depending on the estimates for the main effects, each treatment will have a different estimated mean response. For every level of A there is a main effect, the $A_i$'s. For every level of factor B there is a main effect, $B_i$. For every combination of A and B levels there is an interaction effect,

$(AC)_{ij}$. So the model implies that each treatment mean is made up of an overall mean, two main effects and and interaction term.

## 14.1 Replication

Replication is crucial in any experiment! Without replication, we cannot estimate the experimental error variance ($\sigma^2$), which is essential for assessing variability and conducting hypothesis tests.

If we only have one observation per treatment, that observation becomes the treatment mean. Since we cannot compute deviations from the treatment mean, there is no estimate of error variance. This means that while we can technically estimate the model parameters, the model itself is practically useless—we cannot perform hypothesis tests without an estimate of error variance. And if we can't test anything, what's the point?

In factorial experiments, the situation gets even worse when we don't replicate treatments. Specifically, we can't calculate the interaction effect. In general, the interaction effect is calculated as the difference between the treatment mean and the sum of the main effects and the overall mean:

$$(AB)_{ij} = \bar{Y}_{ij} - (\mu + A_i + B_j)$$

Now consider the first plot in the series below. There is only one observation in the hypothetical treat $i = 1$ and $j = 3$. That means that the treatment mean $\bar{Y}_{ij}$ is just the mean of this single observation. We can't calculate any deviations from this mean with only one observation as we usually would for the error variance. We always need to calculate an error term and this is always calculated as the deviation of the observation to the next closest mean. With only one observation per treatment, the next closest mean to that observation is the sum of the main effects: $\mu + A_i + B_j$. But wait, that means the error term is also the interaction term since the treatment mean and the observation are the same? Jup! Now you see the problem. With no replication, the error term and the interaction effect are confounded.

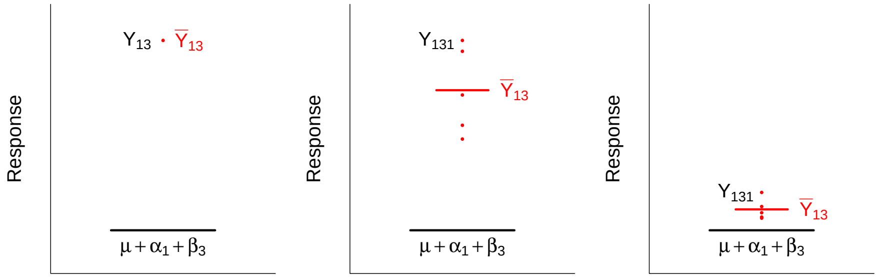


Figure 14.1: How is the interaction effect calculated? (a) only one observation, interaction effect confounded with error term, cannot estimate interaction effect; (b) large interaction present; (c) interaction statistically not discernable (very small).

Have a look at the second plot. Now we have five observations within the treatment. We can calculate the 5 error terms:

$$r_{13k} = Y_{13k} - \bar{Y}_{13}$$

and we can calculate the interaction effect:

$$(\hat{AB})_{13} = \bar{Y}_{13} - (\hat{\mu} + \hat{A}_1 + \hat{B}_3)$$

They are no longer the same thing, they are separable! The interaction effect for this treatment is quite big if you look at the difference visually. For the last plot, there are also five observations, but now the deviation of the treatment mean from the sum f the main effects is almost zero; it's just due to random variation. The interaction effect is too small to detect statistically.

## 14.2 Parameter estimation

The maximum likelihood/least squares estimates are found by minimizing the error or residual sum of squares:

$$S = \sum_{ijk}(Y_{ijk} - \mu - \alpha_i - \beta_j - (\alpha\beta)_{ij})^2$$

The solutions to these equations are the least squares estimates:

$$\hat{\mu} = \bar{Y}_{\dots}$$

$$\hat{A}_i = \bar{Y}_{i..} - \bar{Y}_{...}, \quad i = 1, \ldots, a$$

$$\hat{C}_j = \bar{Y}_{.j.} - \bar{Y}_{...}, \quad j = 1, \ldots, c$$

$$(\hat{AC})_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}, \quad i = 1, \ldots, a, \quad j = 1, \ldots, c$$

The main effects are as before, except that now they refer to differences between row or column means ([Figure 6.3]) and the overall mean. The interaction effects are estimated as the differences between treatment means and the sum of the main effects.

## 14.3 Back to the example

Let's fit this model to the data of the playback and lecture modality experiment. This time, we have access to the actual data collected! Let's explore the data and check our assumptions. The assumptions are the same as in a one-way ANOVA. That is normality of errors, equal population variances, independent errors and no outliers.

As always we start by reading in our data, checking that it has been read in correctly and looking at some descriptive statistics.

```
data <- read.csv("Datasets/Exp2DataPlayback.csv")

head(data)
```

```
  Participant.ID       Condition Speed Content.Type Accuracy
1     945445adf5 1x Audio-Visual     2 Audio-Visual       42
2     23afb88ef3        1x Audio     1   Audio-Only       56
3     1bc24e0480        1x Audio     1   Audio-Only       62
4     4fbdbd41a5        1x Audio     1   Audio-Only       44
5     442adf227a        1x Audio     1   Audio-Only       56
6     3ca9d09e2e 1x Audio-Visual     2 Audio-Visual       48
```

The data set contains 5 columns:

1. `Participant.ID` – This column contains a unique identification code for each participant in the study.

2. `Condition` – Indicates the experimental condition or treatment, which includes both playback speed (1x or 2x) and content type (Audio-Only or Audio-Visual).

3. `Speed` – A numeric column that explicitly represents the playback speed, with `1` for normal speed (1x) and `2` for double speed (2x).

4. `Content.Type` – Specifies whether the participant received Audio-Only or Audio-Visual content.

5. `Accuracy`– The participant's performance score, representing comprehension accuracy.

```r
summary(data)
```

```
 Participant.ID      Condition             Speed      Content.Type
 Length:200         Length:200         Min.   :1.0   Length:200
 Class :character   Class :character   1st Qu.:1.0   Class :character
 Mode  :character   Mode  :character   Median :1.5   Mode  :character
                                       Mean   :1.5
                                       3rd Qu.:2.0
                                       Max.   :2.0
    Accuracy
 Min.   :14.00
 1st Qu.:42.00
 Median :54.00
 Mean   :52.78
 3rd Qu.:64.00
 Max.   :90.00
```

From the summary, you should notice a few things:

- All the columns are read in as character values except `Speed` and `Accuracy`. We need the relevant columns to factors if we want to use them in our analysis.

- `Speed` and `Accuracy` seem to be read in as numeric values. This makes sense for `Accuracy` but not `Speed`! `Speed` is a categorical variable with levels 1x and 2x, we need to correct this.

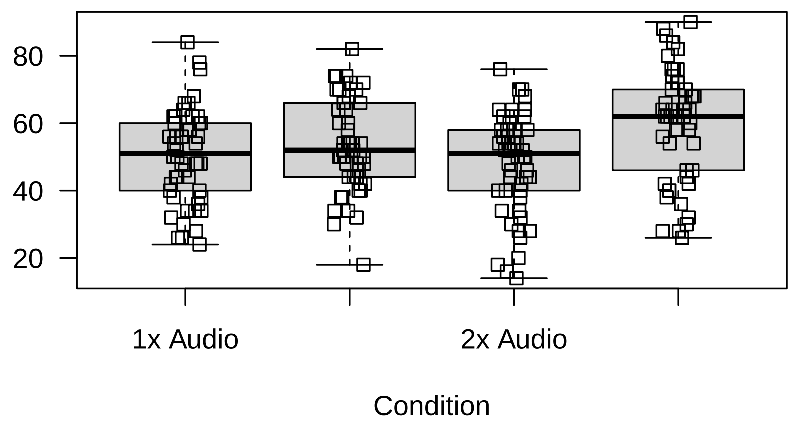
```r
data$Condition <- factor(data$Condition)
data$Content.Type <- factor(data$Content.Type)
data$Speed <- factor(data$Speed)

summary(data)
```

```
 Participant.ID               Condition  Speed         Content.Type
 Length:200         1x Audio        :50  1:100   Audio-Only  :100
 Class :character   1x Audio-Visual:50  2:100   Audio-Visual:100
 Mode  :character   2x Audio        :50
                    2x Audio-Visual:50
```

```
    Accuracy
 Min.   :14.00
 1st Qu.:42.00
 Median :54.00
 Mean   :52.78
 3rd Qu.:64.00
 Max.   :90.00
```

Great, now we can see that each treatment was replicated 50 times as we expected. To check our assumptions we start as always by plotting the response against treatments.

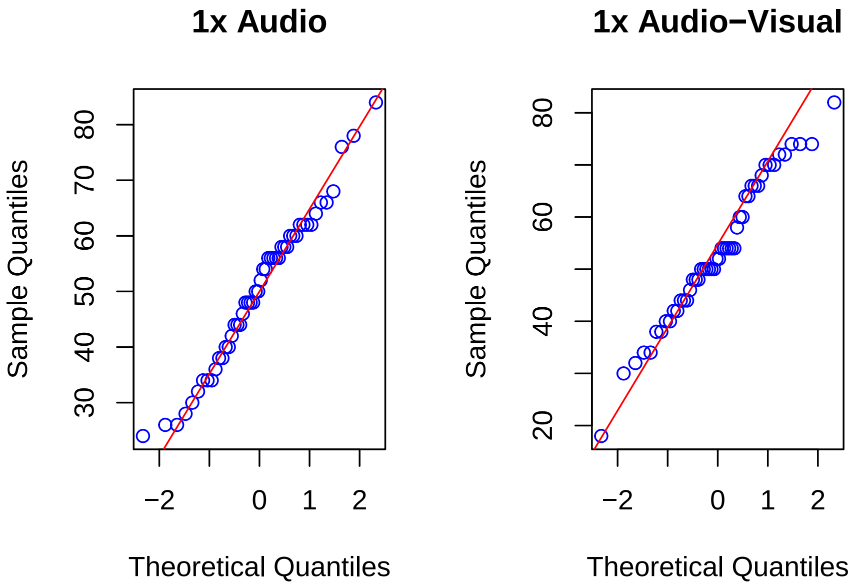
```
boxplot(Accuracy ~ Condition, data = data,
        ylab = "", main = "", las = 1)

# we could also have specified the first argument as Accuracy ~ Content.Type * Speed

stripchart(Accuracy ~ Condition, data = data, vertical = TRUE, add = TRUE, method = "j
```



There are no clear signs of deviation from normality, the box-plots look relatively symmetric. We could plot Q-Q plots for each treatment as well. Let's do that for two of the treatments.

```
par(mfrow=c(1,2)) # splitting the plotting window into 1 row with 2 columns

qqnorm(data$Accuracy[data$Condition == "1x Audio"], pty = 4, col ="blue", main = "1x Au
qqline(data$Accuracy[data$Condition == "1x Audio"], col = "red")
```

```
qqnorm(data$Accuracy[data$Condition == "1x Audio-Visual"], pty = 4, col ="blue", main = "1x Audio
qqline(data$Accuracy[data$Condition == "1x Audio-Visual"], col = "red")
```



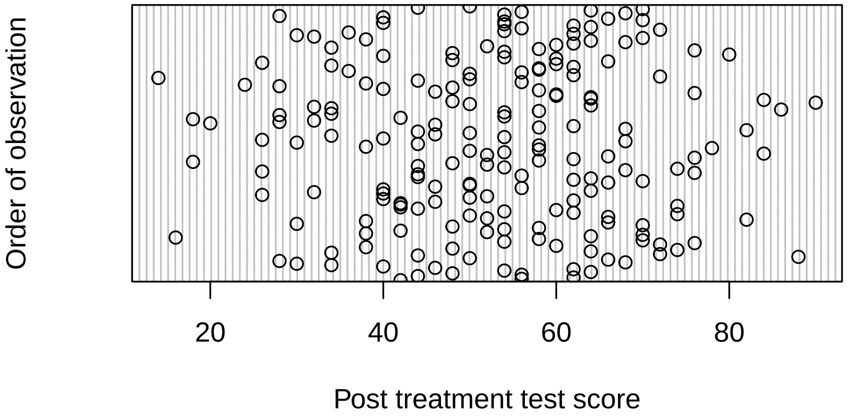
No worrisome patterns! Next, the box-plots also suggest that there are no outliers and there are no clear indications that the assumption of equal population variance is not reasonable. Let’s have a look a the standard deviations.

```
sort(tapply(data$Accuracy, data$Condition, sd))
```

```
1x Audio-Visual        1x Audio        2x Audio 2x Audio-Visual
       13.71690        14.01084        14.93064        16.76386
```

The ratio of the smallest to largest is roughly 1.22 which is much smaller than five. Lastly, we need to check the assumption of independence. We start by assuming that the order in which the data are in the data set is the order in which the measurements were taken and we construct a Cleveland dot chart.

```
dotchart(data$Accuracy, ylab = "Order of observation", xlab ="Post treatment test score")
```



No clear patterns in the measurements, so no reason to suspect any dependence between successive measurements. The students were randomly assigned to each group and there are no other reasons to believe that independence was violated based on the description of the experiment.

With all the assumptions checked, we can move onto fitting the linear model to our data and inspecting the model estimates. Here is the model equation:

$$\text{Accuracy}_{ijk} = \mu + \text{Speed}_i + \text{Content.Type}_j + (\text{Speed:Content.Type})_{ij} + e_{ijk}$$

where,

$$\begin{aligned} i &= 1, 2 \text{ and } j = 1, 2 \\ e_{ijk} &= \text{random error with } e_{ijk} \sim N(0, \sigma^2) \end{aligned}$$

In R, we fit the model like this:

```r
model <- aov(Accuracy ~ Speed + Content.Type + Speed:Content.Type, data = data)
model.tables(model, type = "means", se = TRUE)
```

```
Tables of means
Grand mean

52.78

 Speed
```

```
Speed
    1     2
54.94 50.62

 Content.Type
Content.Type
  Audio-Only Audio-Visual
       49.10        56.46

 Speed:Content.Type
     Content.Type
Speed Audio-Only Audio-Visual
    1 50.32      59.56
    2 47.88      53.36

Standard errors for differences of means
        Speed Content.Type Speed:Content.Type
        2.108        2.108              2.981
replic.   100          100                 50
```

R allows a convenient shorthand for this type of model. Instead of typing out all three terms, you can shorten the right hand side of the formula to `Speed*Content.Type`. The `*` operator indicates to R that we want main effects and interaction effects. Try it yourself to see that you get the same result.

We extract the treatment means as before. The grand mean is shown first. Now with a factorial treatment structure, we get the mean values for each level of the treatment factors included and the treatment means. In the output below, we see the means of the 1x and 2x speed followed by the means for the levels of content type. Lastly, the treatment means are presented in a 2 by 2 matrix format. The treatment “1x Audio Only” had a mean accuracy of 50.32, “2x Audio-Only” mean response is 47.8, and so on. And then lastly, the standard errors for differences between means.

We can also extract the estimated effects as before.

```
model.tables(model, type = "effects", se = TRUE)
```

```
Tables of effects

 Speed
Speed
    1     2
 2.16 -2.16

 Content.Type
Content.Type
  Audio-Only Audio-Visual
```

```
       -3.68         3.68

 Speed:Content.Type
     Content.Type
Speed Audio-Only Audio-Visual
    1 -0.94       0.94
    2  0.94      -0.94

Standard errors of effects
        Speed Content.Type Speed:Content.Type
        1.490        1.490              2.108
replic.   100          100                 50
```

First we get the main effects for `Speed` and `Content.Type`. Then we get the interaction effects and standard errors. Let's check that we understand how these interaction effects are calculated. Remember:

$$(AB)_{ij} = \bar{Y}_{ij} - (\mu + A_i + B_j)$$

So for treatment $i = 1$ and $j = 1$, the equation becomes:

$$(\hat{AB})_{11} = \bar{Y}_{11} - (\hat{\mu} + \hat{A}_1 + \hat{B}_1)$$

We go by the dimensions of the matrix returned by R, so then treatment $i = 1$ and $j = 1$ is “1x Audio-Only”. Substituting the estimated values:

$$\begin{aligned}(AB)_{11} &= 50.32 - (52.78 + 2.16 - 3.68)\\ &= -0.94\end{aligned}$$

Which is what R outputs as well. Now, we want to ask is there evidence for an interaction effect? To do this we need to construct the ANOVA table.

# Chapter 15

# ANOVA

The model for a factorial experiment with two treatment factors was:

$$Y_{ijk} = \mu + A_i + C_j + (AC)_{ij} + e_{ijk}$$

If we move $\mu$ to the left-hand side of the equation, we get:

$$Y_{ijk} - \mu = A_i + C_j + (AC)_{ij} + e_{ijk}$$

Now, each of the terms on the RHS is a deviation from a mean.

- The main effects come from the overall mean $\mu$,
- The interaction effects from $\mu + A_i + C_j$,
- The error terms from the treatment means.

We can square and sum the corresponding observed deviations and obtain sums of squares. For a **balanced** factorial experiment, the total sum of squares on the LHS can be split into four parts, corresponding to:

1. Main effects of factor **A**,
2. Main effects of factor **C**,
3. Interaction between **A** and **C** effects,
4. Error.

$$SS_{total} = SS_A + SS_C + SS_{AC} + SS_E$$

The degrees of freedom for these sums of squares are:

$$abn-1=(a-1)+(b-1)+(a-1)(b-1)+ab(n-1)$$

where $n$ is the number of replicates per treatment. The degrees of freedom on the right-hand side add up to the total degrees of freedom. Once again, we summarise all this in a table.

## ANOVA Table

The following table summarizes the partitioning of variation:

| Source | SS | df | MS | F |
|---|---|---|---|---|
| A Main Effects | $SS_A = nb\sum_i(\bar{Y}_{i..}-\bar{Y}_{...})^2$ | $(a-1)$ | $MS_A$ | $\frac{MS_A}{MS_E}$ |
| C Main Effects | $SS_C = na\sum_j(\bar{Y}_{.j.}-\bar{Y}_{...})^2$ | $(b-1)$ | $MS_C$ | $\frac{MS_C}{MS_E}$ |
| AC Interactions | $SS_{AC} = n\sum_{ij}(\bar{Y}_{ij.}-\bar{Y}_{i..}-\bar{Y}_{.j.}+\bar{Y}_{...})^2$ | $(a-1)(b-1)$ | $MS_{AC}$ | $\frac{MS_{AC}}{MS_E}$ |
| Error | $SS_E = \sum_{ijk}(Y_{ijk}-\bar{Y}_{ij.})^2$ | $ab(n-1)$ | $MSE$ | - |
| Total | $SS_{total} = \sum_{ijk}(Y_{ijk}-\bar{Y}_{...})^2$ | $abn-1$ | - | - |

There are three F-tests in this ANOVA table.

1. $H_{AB} : (\alpha\beta)_{ij} = 0$ for all $i$ and $j$ (Factors A and B do not interact)
2. $H_A : \alpha_i = 0 \quad i = 1, \dots, a$ (Factor A has no main effects)
3. $H_B : \beta_j = 0 \quad j = 1, \dots, b$ (Factor B has no main effects)

The alternative hypothesis is, in each case, that at least one of the parameters considered is non-zero.

While discussing interactions, we saw that sometimes, with strong interaction effects, the main effects of a factor may disappear (be close to zero). But this does not mean that the factor has no effect. On the contrary, it has an effect on the response; the effects just differ over the levels of the other factor and may average out.

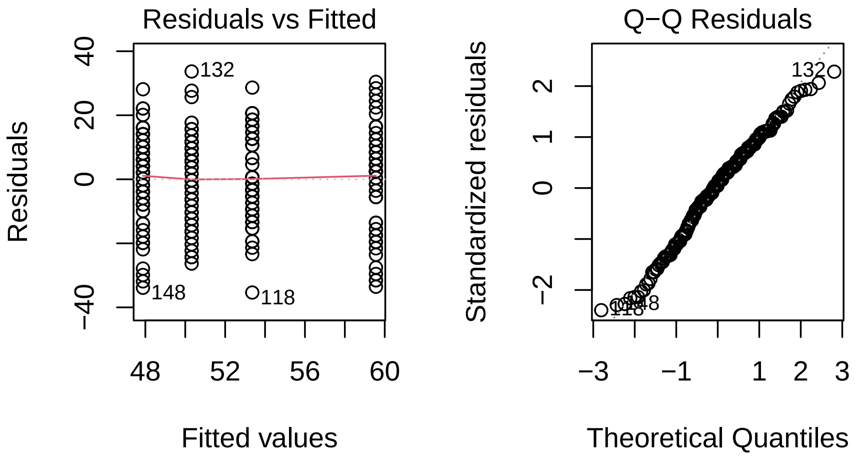
Therefore, we usually start by testing the interaction effects. If there is evidence for the presence of interactions, we have to examine the main effects with this in mind, i.e., be careful with the interpretation of the main effects. Some people say that it becomes meaningless to test for main effects if there is evidence of interactions. However, this depends on what we want to know. The main effects still tell us whether or not the average response changes with changing levels of the factor.

The **F-ratio** always has the mean square for error in the denominator. As before, it is a ratio of two variance estimates, and in each case, it can be seen as a **signal-to-noise ratio**: how large are the effects relative to the experimental error variance?

## 15.1 Back to the example

Before we inspect the ANOVA table for the working example, we need to check the assumptions about the errors after model fitting. We do this by inspecting the residuals.

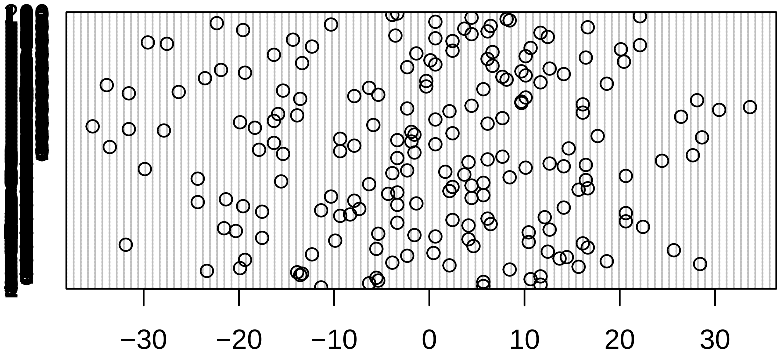
```
par(mfrow = c(1,2))
plot(model, which = 1)
plot(model, which = 2)
```



There are no clear violations, in the first plot, the residuals appear to be centered around zero and the spread is reasonably equal across groups. The second plot is a Q-Q plot of the residuals which shows nothing worrisome. Remember we can also plot a histogram of the residuals to check normality.

For the independence assumption, we construct the dot chart once again but with the residuals.

```
dotchart(model$residuals) # note the different way of extracting residuals!
```



The y-axis is messy but we can ignore that, it shows the index of each observation and there are 200 hence why it overlaps so much. The residuals look uniform, there are no systematic patterns or trends in the plot.

Let's see what the ANOVA table looks like for our working example.

```
summary(model)
```

```
                    Df Sum Sq Mean Sq F value   Pr(>F)    
Speed                1    933   933.1   4.201 0.041724 *  
Content.Type         1   2708  2708.5  12.195 0.000592 ***
Speed:Content.Type   1    177   176.7   0.796 0.373485    
Residuals          196  43532   222.1                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Verify that the degrees of freedom are what you expected! First, we look at the interaction. The F-value is quite small which leads to a large p-value of 0.37. This means that we really have no evidence against the null hypothesis that the factors interact. There is some evidence for a main effect of Speed but there is much stronger evidence as indicated by the small p-value for a main effect of lecture modality.

# Chapter 16

# Contrasts

There are two approaches to analysing data from experiments. The first is to construct a set of a-priori contrasts, test these, and perhaps afterwards use unplanned comparisons to see if there are any other interesting treatment effects or differences that we might want to follow up with in a future experiment.

The second approach is an analysis of variance (ANOVA). This usually tests much more general hypotheses about the presence of main and interaction effects. The two approaches are not mutually exclusive, but if the questions we are interested in are not answered by an analysis of variance, we should concentrate on the contrasts. The two approaches may also give what seem to be different answers.

For example, from the ANOVA F-test, we may see no evidence for interactions, but if we look at specific contrasts for interactions, there is evidence. This can happen; it is not a mistake in the methods, it is just a difference in the hypotheses that are being tested.

Often, an ANOVA is expected in journal publications and research reports, even if it does not answer the specific research questions. The more specific questions are answered by constructing confidence intervals or tests for contrasts.

Let's revisit the specific research questions for the working example:

1. Does lecture modality have an effect on comprehension?
2. Does playback speed have an effect on comprehension?
3. Is there an interaction effect of modality and playback speed on comprehension?

With these question, conducting an ANOVA is enough. We simply want to know if there are any main effects or interaction effects. We have answered that with the ANOVA above. But what if the questions were a bit more specific:

1. Does audio-visual content increase comprehension?

2. Does increased playback speed decrease comprehension?
3. Is the effect of playback speed improved by audio-visual content?

So far we have only contrasted two treatments. Sometimes we want to compare groups of treatments to one another. More generally, a contrast is defined as a linear combination of the parameters where the coefficients add up to zero:

$$L = \sum_{1}^{a} h_i A_i$$

such that $\sum_1^a h_i = 0$. This ensures a fair comparison. For example, in a comparison of two group means we have:

$$L = \mu_1 - \mu_2 = 1 \times \mu + (-1) \times \mu_2$$

Here, the coefficients are $h_1 = +1$ and $h_2 = -1$ which sum to zero. This simple difference is the simplest form of a contrast. Effectively, $\sum_1^a h_i = 0$ represents the null hypothesis, that the difference equals 0.

Let's start with the first question. Remember the treatments were:

1. 1x Audio-Only (1AO)
2. 2x Audio-Only (2AO)
3. 1x Audio-Visual (1AV)
4. 2x Audio-Visual (2AV)

To answer the first question, our contrast should compare Audio-Visual vs. Audio-Only and we do this by averaging over the levels of playback speed.

First we compute the average response for the two levels of content type, AV and AO.

$$\frac{(\mu_{1AV} + \mu_{2AV})}{2}$$

$$\frac{(\mu_{1AO} + \mu_{2AO})}{2}$$

Now we are comparing groups of means. The first group contains the means for all treatments that included Audio-Visual level and the second contains the Audio-Only level. We are asking whether the AV level increased comprehension. So we are testing:

We could specify the difference either way, that is AO - AV. Then we would be doing a one-sided lower tailed test.

$$H_0 : \frac{(\mu_{1AV} + \mu_{2AV})}{2} = \frac{(\mu_{1AO} + \mu_{2AO})}{2}$$

$$H_1 : \frac{(\mu_{1AV} + \mu_{2AV})}{2} > \frac{(\mu_{1AO} + \mu_{2AO})}{2} <=> \frac{(\mu_{1AV} + \mu_{2AV})}{2} - \frac{(\mu_{1AO} + \mu_{2AO})}{2} > 0$$

The coefficients of the contrast sum to zero:

$$\begin{aligned}
&\frac{(\mu_{AV1} + \mu_{AV2}) - (\mu_{AO1} + \mu_{AO2})}{2} \\
&\frac{(1)\mu_{AV1} + (1)\mu_{AV2} + (-1)\mu_{AO1} + (-1)\mu_{AO2}}{2} \\
&(0.5)\mu_{AV1} + (0.5)\mu_{AV2} + (-0.5)\mu_{AO1} + (-0.5)\mu_{AO2} \\
&0.5 + .0.5 - 0.5 - 0.5 = 0
\end{aligned}$$

This is a linear combination of the model parameters. What does the contrast and coefficients look like for the second question? To test whether playback speed decreases comprehension, we need to compare treatments at 1x speed vs. 2x speed:

$$\frac{(\mu_{1AV} + \mu_{1AO})}{2} - \frac{(\mu_{2AO} + \mu_{2AV})}{2}$$

The coefficients sum to zero as before. This might be confusing but we are simply grouping treatments together and comparing them. To compute these contrasts in R, we first fit the model using lm() and extract the treatment means using `emmeans` from the `emmeans` package.

```r
model_reg <- lm(Accuracy ~ Content.Type * Speed, data = data)

means <- emmeans(model_reg, ~Content.Type * Speed)

means
```

```
 Content.Type Speed emmean   SE  df lower.CL upper.CL
 Audio-Only       1   50.3 2.11 196     46.2     54.5
 Audio-Visual     1   59.6 2.11 196     55.4     63.7
 Audio-Only       2   47.9 2.11 196     43.7     52.0
 Audio-Visual     2   53.4 2.11 196     49.2     57.5

Confidence level used: 0.95
```

The `emmeans` function returns the treatment means, the standard error, degrees of freedom and the bounds of 95% confidence interval. Now we want to perform the two contrasts using the means saved in the object we created, `means`. First, note the order in which `emmeans` outputs the treatments:

AO1, AV1, A02, AV2.

We are going to use this order and the coefficients were determined earlier to perform the ocntrasts with the function `contrast()` also from the package `emmeans`:

```
contrast(means,
         list(
           c1 = c(-1,1,-1,1)/2, # AV - AO
           c2 = c(1,1,-1,-1)/2 # 2x - 1x
         ),
         by = NULL, side = ">")
```

```
 contrast estimate   SE  df t.ratio p.value
 c1           7.36 2.11 196   3.492  0.0003
 c2           4.32 2.11 196   2.050  0.0209
```

```
P values are right-tailed
```

We supply the emmeans object `means` and then a list of contrasts we call `c1` and `c2` corresponding to the first and second question. Each contrast consists of the coefficients in the order in which the means appear in the `means` object and the scaling by 2. Then we need to specify `by = NULL` because we have manually coded the contrasts and don't need to specify by which factor the contrasts should made. Lastly, we specify the type of test we want, that is, is it one sided or two sided. If it is one-sided, in which direction? We have specifically constructed the contrasts so that both are "one-sided greater than".

The output shows the estimate of each contrast, the standard error of the difference in means, t-value and associated p-value. For the first contrast we see the difference in comprehension scores between the Audio-Visual and Audio-Only groups was 7.36, this means that the average response in the Audio-Visual group was higher than the average response in the Audio-Only group. We see that the p-value to test this contrast is 0.0003 which is extremely small, so it is unlikely that the difference in mean response is due to chance. There is strong evidence to indicate that the audio-visual type increased the mean response, the estimate of this the difference between groups is 7.36% ($t = 3.492$, $df = 196$, $p = 0.0003$).

For the second contrast, the p-value still provides sufficient evidence against the null hypothesis that the difference is zero but it is not as strong as for the first contrast. However, we are still satisfied with the evidence against $H_0$. The 2x speed decreased the average accuracy (averaged over the levels of content type) by 4.32% ($t = -2.050$, $df = 196$, $p = 0.021$).

When we have factors with two levels (as we do here) and we conduct two sided contrasts, then the contrast is equivalent to testing for the presence of main effects which what the ANOVA table does! Remember we said that the ANOVA is an extension of the t-test and with two levels. Let's go through this step-by-step:

- We conducted one-sided tests.

- If we conducted two-sided tests, the results would be the same as in ANOVA table.
- This is because when we have two levels per treatment factor, the contrasts are equivalent to testing whether there are main effects of Speed and Content.Type.

Since we conducted one-sided tests, the p-value is has been split between the tails. To get to the value of the p-value for a two-sided tests, we multiply the one-sided p-value by 2.

```
# For AV - AO = 0

0.0003*2
```

```
[1] 6e-04
```

```
# For 1 - 2 = 0

0.0209 * 2
```

```
[1] 0.0418
```

Check that these are the same as in the ANOVA table. The test statistics are also related in this case, $t^2 = F$.

Let's answer the third question. Since we have two levels per factor, this question is asking about the interaction. The contrast for the interaction should compare the *difference between audio-visual and audio-only in the two levels of playback speed*:

At 1x playback speed, the effect of content type is given by:

$$(\mu_{AV1} - \mu_{AO1})$$

At 2x playback speed, the effect of content type is given by:

$$(\mu_{AV2} - \mu_{AO2})$$

Now to examine whether the effect of content type is consistent across playback speeds, we compute:

$$\begin{gathered}(\mu_{AV1} - \mu_{AO1}) - (\mu_{AV2} - \mu_{AO2}) \\ = \mu_{AV1} - \mu_{AO1} - \mu_{AV2} + \mu_{AO2}\end{gathered}$$

This contrast assesses whether the difference between Audio-Visual and Audio-Only is the same at 1x and 2x speeds.

We are not dividing by two because we are not averaging across conditions, we are computing the difference of two differences.

```
contrast(means,
         list(
           c3 = c(-1, 1, 1,-1) # interaction
         ),
         by = NULL)
```

```
 contrast estimate   SE  df t.ratio p.value
 c3           3.76 4.22 196   0.892  0.3735
```

We get the same p-value as in the ANOVA table which indicates a lack of evidence against the null hypothesis, there is no evidence to suggest that the two factors interact ($t = 0.892$, $df = 196$, $p = 0.374$).

In practice, we would test the interaction first and then interpret the main effects if there is evidence to support their presence. Here we have done it this way around purely for educational purposes.

We can also visualise the interaction (especially useful for understanding the interaction if there is evidence for one!). There is a built-in function in R that can do this for us, but it will be useful to construct the plot from scratch to ensure you understand what it visualises.

Have a look at the data set again.

```
head(data)
```

```
  Participant.ID       Condition Speed Content.Type Accuracy
1     945445adf5 1x Audio-Visual     2 Audio-Visual       42
2     23afb88ef3        1x Audio     1   Audio-Only       56
3     1bc24e0480        1x Audio     1   Audio-Only       62
4     4fbdbd41a5        1x Audio     1   Audio-Only       44
5     442adf227a        1x Audio     1   Audio-Only       56
6     3ca9d09e2e 1x Audio-Visual     2 Audio-Visual       48
```

We want to visualise the response per treatment for each combination of Speed and Content.Type (which is already combined in the column Condition). We did this with the `emmeans` function and stored the treatment means in the object `means`! To use it to visualise the treatment means we need to convert to a data frame, currently it is something called a "emmGrid"

```
means
```

```
 Content.Type Speed emmean   SE  df lower.CL upper.CL
 Audio-Only       1   50.3 2.11 196     46.2     54.5
 Audio-Visual     1   59.6 2.11 196     55.4     63.7
 Audio-Only       2   47.9 2.11 196     43.7     52.0
 Audio-Visual     2   53.4 2.11 196     49.2     57.5

Confidence level used: 0.95
```

```
class(means)
```

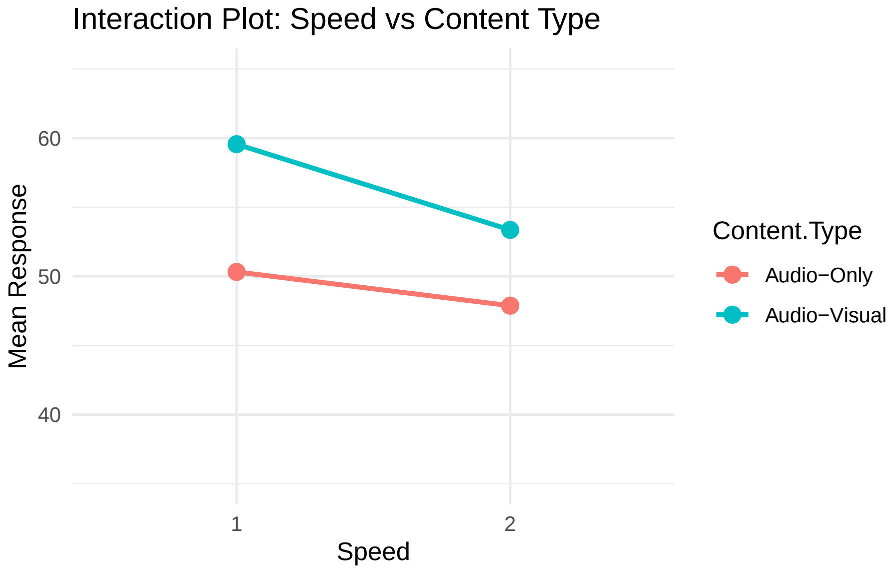
```
[1] "emmGrid"
attr(,"package")
[1] "emmeans"
```

It is easy to convert the object to a data frame:

```
means_data <- data.frame(means)
```

We need to decide which factor will be on the x-axis, let's do Speed. Below, I use a new package called `ggplot2` to visualise the data. It creates nicer looking plots and is more intuitive in my opinion. If you want to see how to use base R to plot this, see the code at the end of this section.

```
# install.packages("ggplot2")
library(ggplot2)

# Create the ggplot with interaction lines
ggplot(means_data, aes(x = factor(Speed), y = emmean, colour = Content.Type, group = Content.Type
  geom_point(size = 3) +     # Add points for each Content Type
  geom_line(linewidth = 1) +      # Connect points with lines
  labs(title = "Interaction Plot: Speed vs Content Type",
       x = "Speed",
       y = "Mean Response") +
  scale_y_continuous(limits =c(35,65)) + # to visualse the magnitude a bit better
  theme_minimal()
```



The `aes()` function maps Speed to the x-axis, Mean Response to the y-axis, and uses Content Type for color and grouping. `geom_point(size = 3)` adds individual data points, while `geom_line(size = 1)` connects them to show trends. The `labs()` function provides axis labels and a title, and `theme_minimal()` specifies the theme for the plot.

It is evident that increasing Speed has a negative effect on the response, and switching from Audio-Visual to Audio-Only content reduces the mean response. When moving from 1x to 2x Speed in the Audio-Visual condition, the response decreases. A similar decline is observed for the Audio-Only condition. Although the decrease appears slightly larger for Audio-Visual than for Audio-Only, the difference is not substantial enough to conclude a significant interaction effect between Speed and Content Type as evidenced by the ANOVA and contrasts we did before.

If you want to know more about how to visualise data with `ggplot2` have a look at this link. There are plenty of resources mentioned.

## 16.1 Conclusion

```r
# Set up an empty plot
plot(means_data$Speed[means_data$Content.Type == "Audio-Visual"],
     means_data$emmean[means_data$Content.Type == "Audio-Visual"],
     type = "o",
     col = "#F79256",
     pch = 16,
     ylim = range(means_data$emmean),
     xlab = "Speed",
     ylab = "Mean Response",
     main = "Interaction Plot: Speed vs Content Type",
     xaxt = "n")

# - plot(...) initializes the graph using Speed as the x-axis and Mean Response as the
# - The subset `means_data$Speed[means_data$ContentType == "Audio-Visual"]` extracts o
# - type = "o" specifies that both points and lines should be drawn.
# - ylim = range(means_data$emmean) ensures that the y-axis spans the full range of da
# - xaxt = "n" suppresses the default x-axis, allowing for manual customization in the
#
# Since the x-axis represents discrete categories (Speed levels), we manually specify

# Add x-axis labels manually
axis(1, at = unique(as.numeric(means_data$Speed)), labels = unique(means_data$Speed))

# Add Audio-Only group
points(means_data$Speed[means_data$Content.Type == "Audio-Only"],
       means_data$emmean[means_data$Content.Type == "Audio-Only"],
       col = "#5BC0EB",
       pch = 16,
       type = "o")
```

```
# Add legend
legend("topright", legend = c("Audio-Visual", "Audio-Only"), col = c("#F79256", "#5BC0EB"), pch =
```

```
# OR WITH BUILT IN

interaction.plot(x.factor = means_data$Speed, #x-axis variable
                 trace.factor = means_data$Content.Type, #variable for lines
                 response = means_data$emmean, #y-axis variable
                 fun = mean, #metric to plot
                 ylab = "Counts",
                 xlab = "Seasons",
                 col = c("red", "blue"),
                 lty = 1, #line type
                 lwd = 2, #line width
                 trace.label = "Species")
```

Aguinis, Herman, Ryan K Gottfredson, and Harry Joo. 2013. “Best-Practice Recommendations for Defining, Identifying, and Handling Outliers.” *Organizational Research Methods* 16 (2): 270–301.

Chen, Ashley, Suchita E Kumar, Rhea Varkhedi, and Dillon H Murphy. 2024. “The Effect of Playback Speed and Distractions on the Comprehension of Audio and Audio-Visual Materials.” *Educational Psychology Review* 36 (3): 79.

Demirbilek, Muhammet, and Tarik Talan. 2018. “The Effect of Social Media Multitasking on Classroom Performance.” *Active Learning in Higher Education* 19 (2): 117–29.

Kuehl, Robert O. 2000. “Design of Experiments: Statistical Principles of Research Design and Analysis.” *(No Title)*.

Salas, Carlos R, Katsumi Minakata, and William L Kelemen. 2011. “Walking Before Study Enhances Free Recall but Not Judgement-of-Learning Magnitude.” *Journal of Cognitive Psychology* 23 (4): 507–13.

Underwood, A. J. 1996. “Analysis of Variance.” In *Experiments in Ecology: Their Logical Design and Interpretation Using Analysis of Variance*, 140–97. Cambridge University Press.