STA2020 ANOVA Notes

Ané Cloete

2024-12-17

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 β, 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:

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

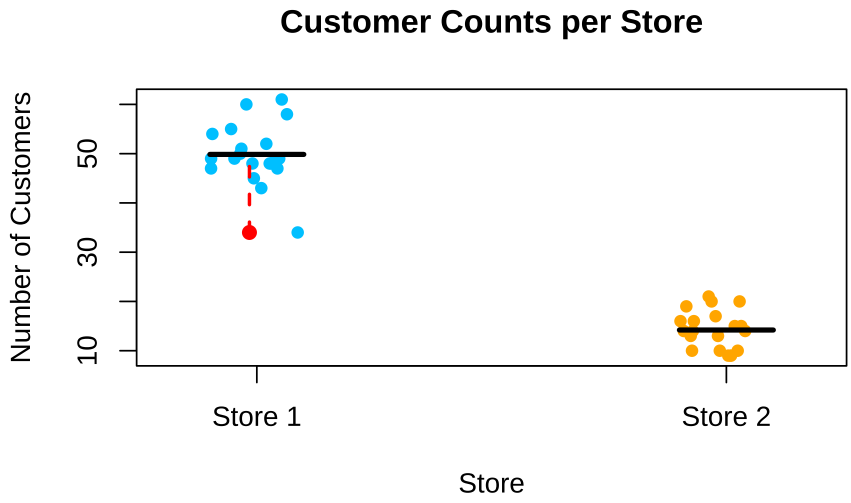
where:

- Y_{ij} is the number of customers observed on day j at store 1,
- μ_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"])
```

```r
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**:

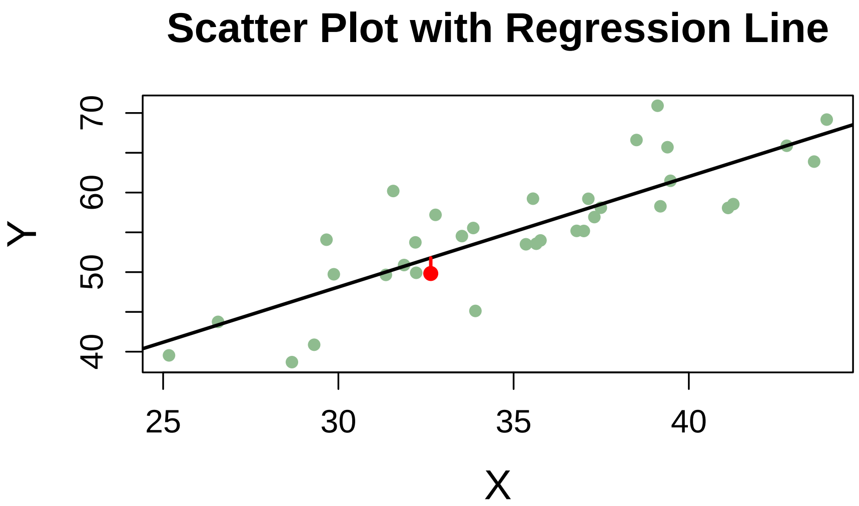
$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

where:

- Y_i is the observed response,
- β_0 and β_1 are unknown parameters representing the intercept and slope,
- X_i is the predictor variable,
- e_i is the random error term.

```r
# 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 = 2
```



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 α. If we obtain a P-value that is smaller than α, we say that we have obtained a "statistically significant result" or that "H_0 is rejected". If our P-value is larger than α, 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 α 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 α. 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 $\text{P} > 0.05$ or $\text{P} < 0.05$ 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 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 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.

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.

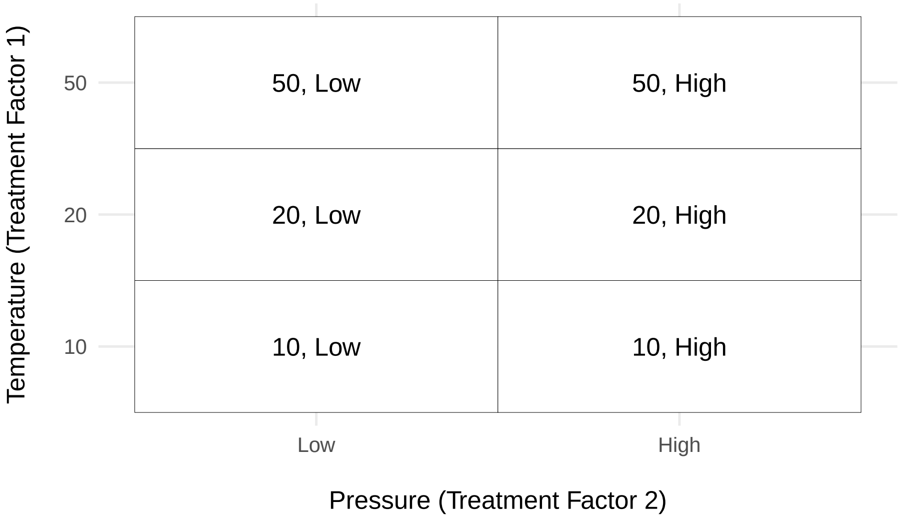

Temperature (Treatment Factor 1)
50
20
10
50, Low
50, High
20, Low
20, High
10, Low
10, High
Low
High
Pressure (Treatment Factor 2)

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`.

```r
# 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] "D" "D" "C" "B" "D" "C" "A" "B" "A" "A" "B" "B" "B" "C" "D" "C" "A" "A" "A"
[20] "C" "B" "D" "C" "D" "C" "C" "A" "B" "B" "A" "D" "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 **??** 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 (**?**), 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 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.

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:

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

```
[1] 120
```

```r
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
```

```
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
```

```
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:

```
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 (**?**) 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. (**?**) 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 (**?**) 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
```

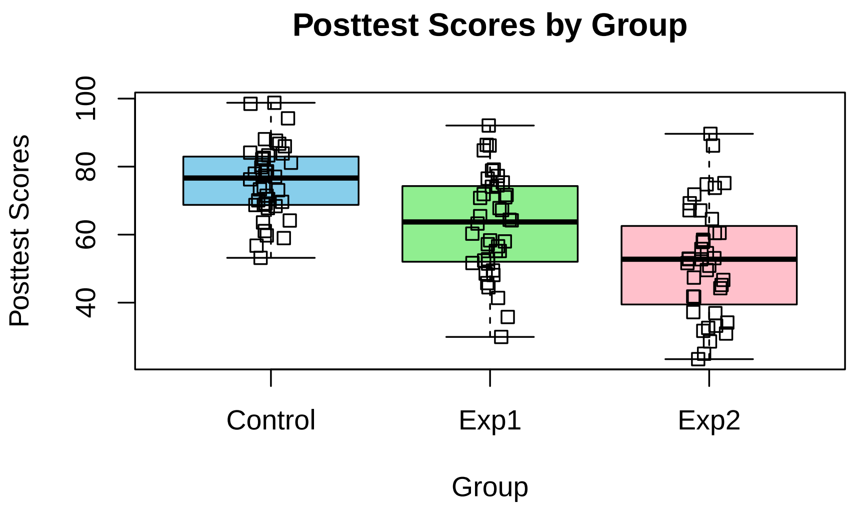


Figure 4.1: Box-plots of Post treatment scores by group.

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:

```
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:

```
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 **??**, 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")
```

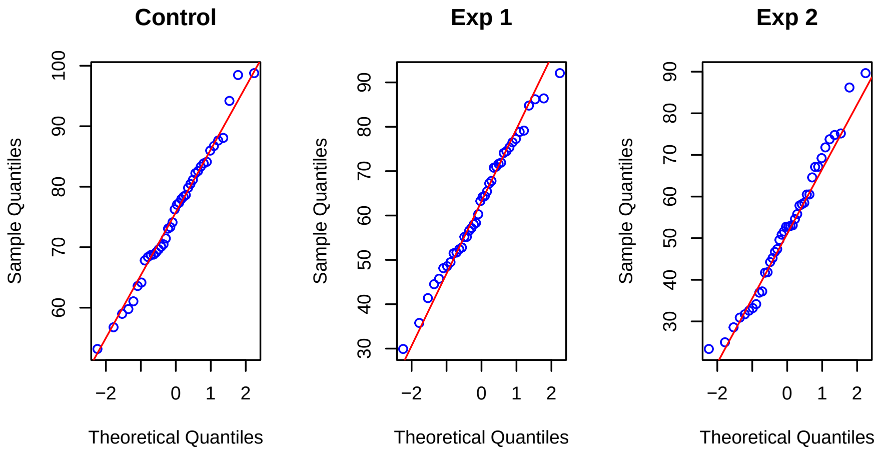


Figure 4.2: Q-Q plots of response per treatment group.

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 (?) 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.

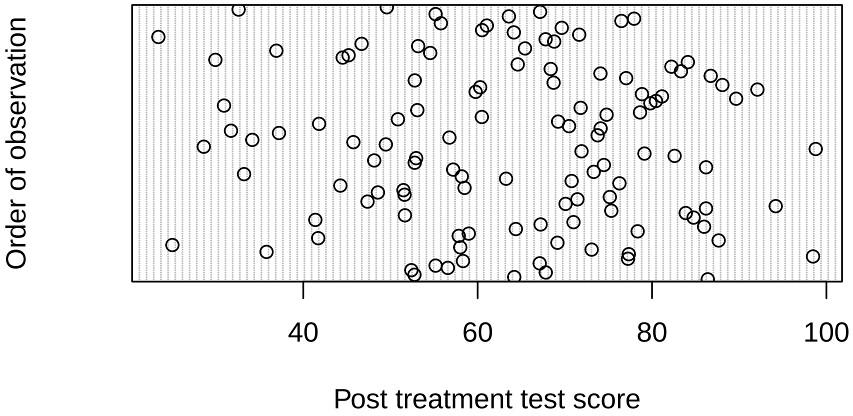


Figure 4.3: Cleveland dot chart of response values in the order in which they appear in the data set.

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 (α) for the test you will conduct. This significance level is the Type 1 error rate (probability of falsely rejecting H_0). A common α 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 μ (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 σ_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):

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

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 μ 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 μ_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

$$\begin{aligned} i &= 1, \dots, a \quad (a = \text{number of treatments}) \\ j &= 1, \dots, r \quad (r = \text{number of replicates}) \\ Y_{ij} &= \text{observation of the } j^{th} \text{ unit receiving treatment } i \\ \mu &= \text{overall or general mean} \\ A_i &= \text{effect of the } i^{th} \text{ level of treatment factor A} \\ e_{ij} &= \text{random error with } e_{ij} \sim N(0, \sigma^2) \end{aligned} \tag{5.1}$$

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 μ and A_i instead of β_0 and β_i which have different meanings.

Regression	ANOVA
β_0 is the mean of the baseline category	μ is the overall mean

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