Analyse Your Own Data - University of Cambridge Statistics Course

This file contains my notes after having attended the course from the 15th to the 19th of April 2024 - Jack Coutts. These notes are my interpretation of the course material and do not necessarily represent the opinions of the course instructors.

Contents

1. The Research Question

The first step in using statistics to address a research question is defining the question itself. The research question is key in narrowing the scope of the research and forces a researcher to focus on something specific. After the research question has been defined, a hypothesis can be formulated and the constraints on the data collection process will become clearer. The research question, hypothesis, and understanding of the proposed dataset data allow a researcher to determine the statistical tests that will be needed to analyse the data (as shown in figure 1).

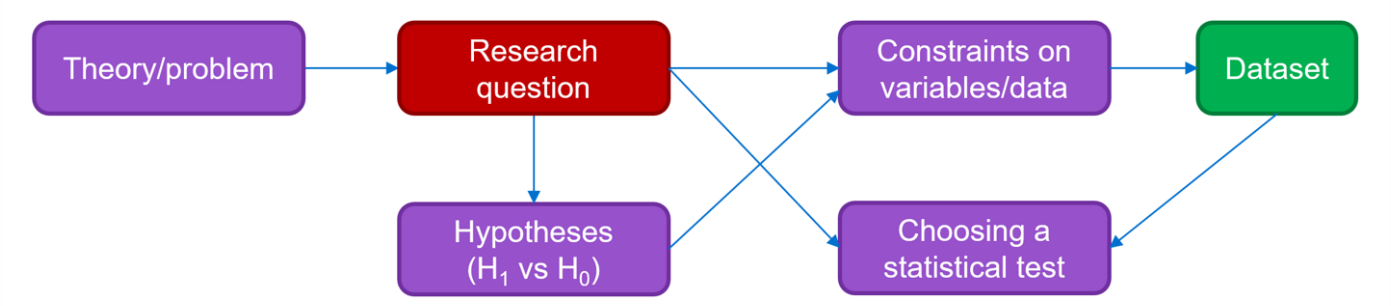

Figure 1. A flowchart depicting the sequence of events in the framework of statistical inference.

What makes a good research question?

A good research question will focus on a single topic. It should not encompass other topics or be two questions veiled as one. A focussed and specific research question will allow you to get much more from your statistics. The question must have measurable independent and dependent variables. A research question should also be useful/relevant, achievable, based on objective metrics, and free from any assumptions.

What happens if you don't have a research question?

What happens when you have data, but you don't have a specific research question? In this situation, several techniques are commonly applied to draw out initial insights which can then be used to generate a research question. The research question generated from this exploratory analysis can then be tested (using statistics) to assess whether the results are likely to have occurred by chance. These exploratory techniques include things like PCA, data mining, and clustering methods.

2. Statistical Inference

Statistical inference is the process of drawing conclusions about a population based on a smaller subset of that population known as a sample. A population refers to the complete set of 'things' that you want to know something about. For example, if you are looking at human height, then the population would be all humans. However, if you are interested in cancer incidence in UK adults over the age of 40, then the population would be all adults in the UK over the age of 40. It is typically not feasible to work with the entire population of interest so a smaller sample of the population is used. Researchers aim for the sample to be representative of the population as a whole but this can be challenging and difficult to assess. As a result, a simplifying framework is used. This framework is known as a **probability distribution**.

A probability distribution describes the likelihood of an individual from the population having a given value for a variable/feature of interest, if that individual was selected at random. The most famous probability distribution is the Gaussian distribution, more commonly called a Normal distribution or a Bell Curve, which is shown in figure 2. For most continuous variables, you can abstract your population to a normal distribution. In simple terms, a distribution is supposed to show how things would look if you had your whole population. From your distribution, you can estimate the population parameters like the mean, median, and variance. You can use distributions to generate hypotheses; for example, the mean global height in humans is 169 cm. There are many types of distributions (see figure 3), and the key thing for researchers to to be aware of is the type of distribution that their samples are drawn from as this allows them to make informed decisions about which statistical tests to implement.

Parameters and **Statistics** are key terms which can be easy to confuse when working with samples and distributions. Parameters refer to values you obtain from a distribution, and statistics refer to values you get from a sample (data which is a subset of a population). The mean and median statistics from a sample are typically a good estimate for their counterpart parameters (the mean and median of the population - which is represented by a distribution). This is not the case for the variance, the variance statistic is very unlikely to be a good estimate of the variance parameter. This is because a sample of a population is highly unlikely to contain the most extreme

values; for example, when considering global human height you are unlikely to sample the shortest and tallest person. Consequently, a different statistic called the sample variance is used instead of variance when estimating the population variance. The sample variance (S^2) applies a correction to make it a better estimate of the population variance.

Ultimately, statistical inference involves two things:

1. Estimating the population parameters.
2. Testing hypotheses regarding the population parameters, where a hypothesis refers to an assumption about the shape of a distribution and/or an assumption about a parameter value.

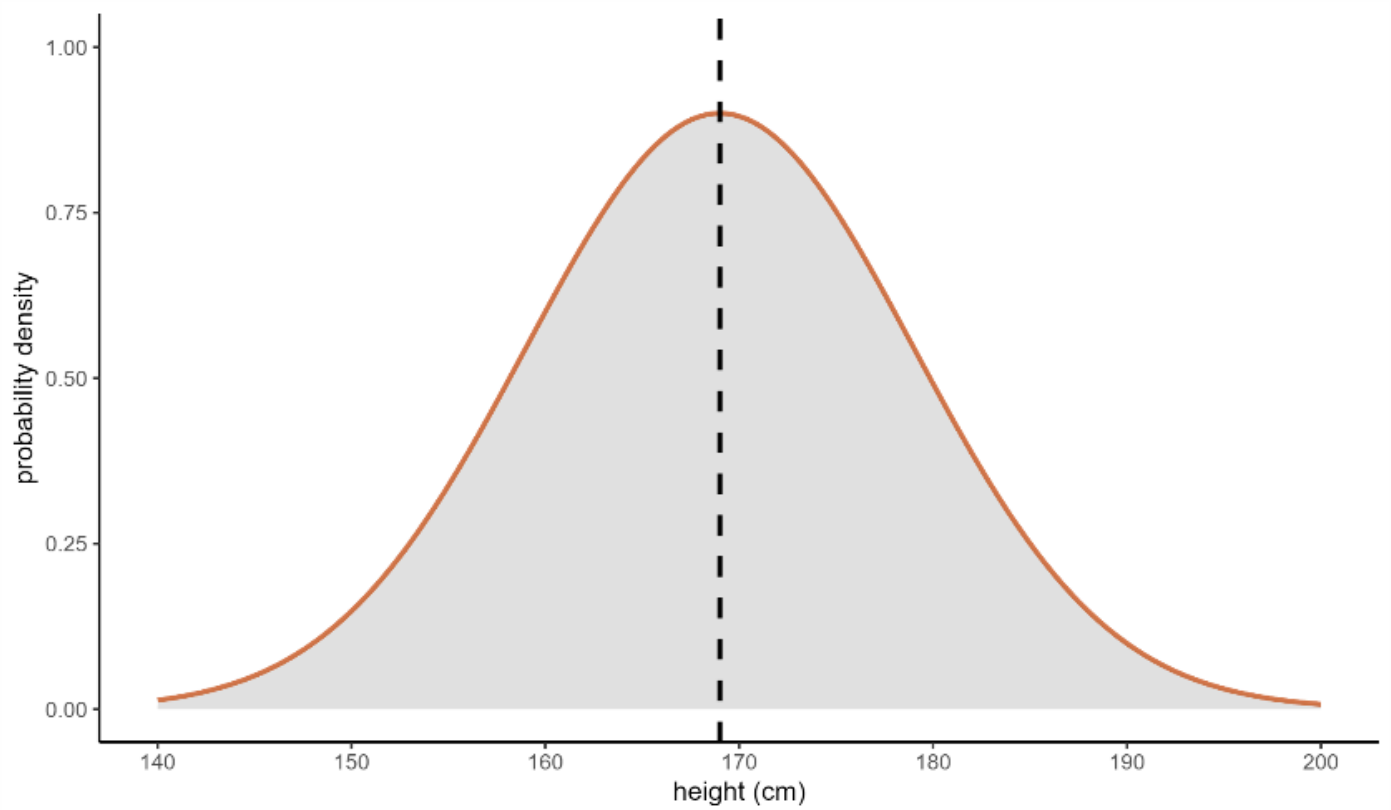

Figure 2. A Normal distribution representing global heights.

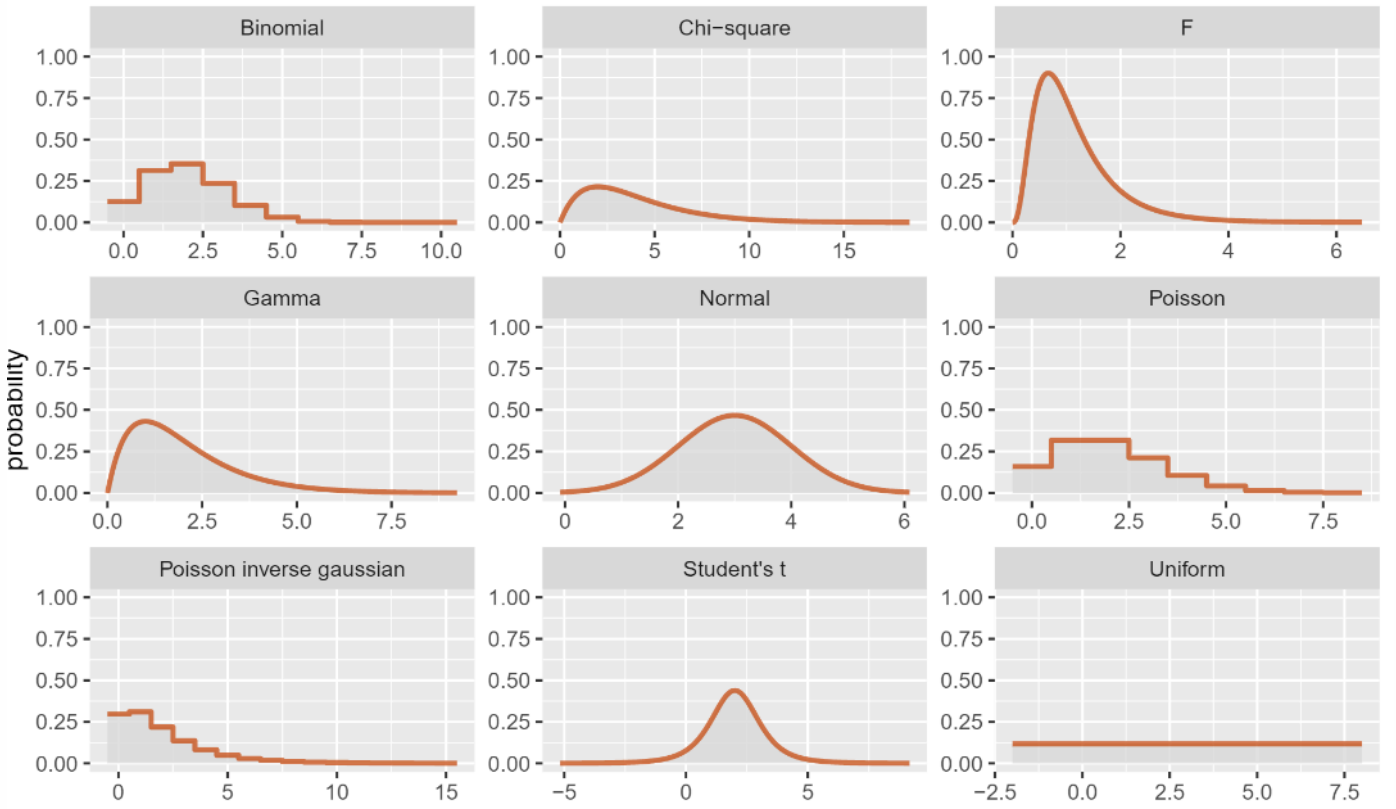

Figure 3. A grid showing distributions other than just a normal distribution.

3. Hypothesis Testing

Hypothesis testing involves asking testable questions that your data can answer.

Classical Hypothesis Testing

1. Target a Population

Generally, researchers will already have a specific field of research and this will determine the population or populations of interest. This population of interest is referred to as the parent distribution and it is from this population/distribution that a sample will be taken to conduct the research. For example, the **parent distribution** for a researcher looking at lung cancer in non-smoking individuals will be all humans with lung cancer that are non-smokers. If a researcher was comparing individuals with cancer to individuals without cancer, then there are two parent distributions, all humans with cancer and all humans without cancer.

2. Formulate a Null Hypothesis (H_0)

After identifying the parent distribution, the next step in classical hypothesis testing is to formulate a **Null Hypothesis (H_0)** and an **Alternative Hypothesis (H_1)**. The null hypothesis and the alternative hypothesis are two opposing assumptions about the parent distribution(s) and its parameters. The null hypothesis is typically thought of as the default assumption in which there is no effect or no difference between populations. This is slightly different when you have a single sample of data, here the null hypothesis is generally that the mean (or another parameter) is a specific value and the alternative hypothesis is that the mean is not that specific value. The alternative hypothesis is generally what a researcher suspects might be true instead of the null hypothesis or is the alternative to the null hypothesis being true.

Example 1 - One-Sample Data:

Researchers are measuring the body length of fish in river X. The parent population is all of the fish in that river and the researchers have taken a sample of 100 fish. The body length is a continuous variable and there is only one sample of fish (a comparison is not being made with fish from another river).

H_0: The mean body length of fish from river X is equal to 20mm.

H_1: The mean body length of fish from river X is not equal to 20mm.

Example 2 - Two-Sample Data:

Researchers are testing a new drug and its efficacy in reducing blood pressure in healthy adults. The parent distribution will be all healthy adult humans. The researchers are comparing the efficacy of the drug against a placebo meaning there will be two samples. There will be a sample of individuals who have been given the placebo and another sample who have been given the drug.

H_0: There is no difference in blood pressure levels between the placebo and drug group.

H_1: There is a difference in blood pressure levels between the placebo and drug group.

3. Select a Significance Level

A significance level, also known as a significance threshold/cutoff, is the maximum tolerable risk of incorrectly concluding that a difference exists when the difference does not actually exist. In other words, it the maximum acceptable probability of getting 'false positive results' from statistical tests. These kind of incorrect conclusion is known as Type I error. In science, a significance level/threshold of $\alpha = 0.05$ is generally used. This threshold of $\alpha = 0.05$ indicates that there is a 5% risk of concluding a difference exists when there is actually no difference. If we continue with the blood pressure drug example, then a significance level of $\alpha = 0.05$ would mean there would be a 5% risk of rejecting the null hypothesis and accepting that there is a difference in blood pressure levels between the placebo and drug groups, when in realty the blood pressures of the two groups are the same.

4. Choose a Statistical Test

Statistical tests are based on a **test statistic**. A test statistic is a value that is calculated to serve as a bridge between the data collected in the study and decision made based on that data. The test statistic is a single number that summarises the data that has been collected. The primary purpose of the test statistic is to measure how far the observed effect in a sample(s) deviates from the effect predicted by the null hypothesis. The value of the test statistic reflects the degree of agreement between what is observed in the collected data and the null hypothesis. For any given type of parent distribution, the test statistic will follow its own theoretical distribution (see figure 4), and statistical tests will use this theoretical distribution to determine the probability of observing the calculated

test statistic in the null hypothesis were true. The probability of observing the calculated test statistic if the null hypothesis were true is known as the **p-value**. This is how statistical tests calculate p-values.

In order to choose a statistical test and a test statistic researchers need to know several things:

- What type of distribution the parent distribution follows. If this is not known, a different type of test will need to be selected and this test will come from a category of tests known as non-parametric statistical tests. Different types of parent distributions will require different statistical tests. Non-parametric tests and other different types of statistical test will be discussed later.
- The type of data that will be collected. Researchers will need to know how many samples/groups will be compared, how many variables will be measured, and what form these variables will take e.g. categorical or continuous etc.

NOTE: It is important not to confuse the term sample(s) with how it is used in many fields (including metabolomics) where it refers to the actual measurements themselves or each data point. Here a sample refers to a subset of the parent population or all of the data points for a specific condition or group.

5. Collect the Data

Measurements from a subset of the parent population need to be collected. One of the biggest challenges in data collection is trying to get a sample that is as representative as possible of the parent population.

6. Interpret the Results

Once the data has been collected, the selected statistical tests can be applied where the tests statistic and p-values can be calculated. Based on the test statistic and p-value, the null hypothesis can be accepted or rejected.

- Reject H_0 if p-value $< \alpha$.
- Fail to reject H_0 if p-value $\geq \alpha$.

Where α is the significance threshold selected by the researcher. This should be done before the data has been collected and the statistical tests applied.

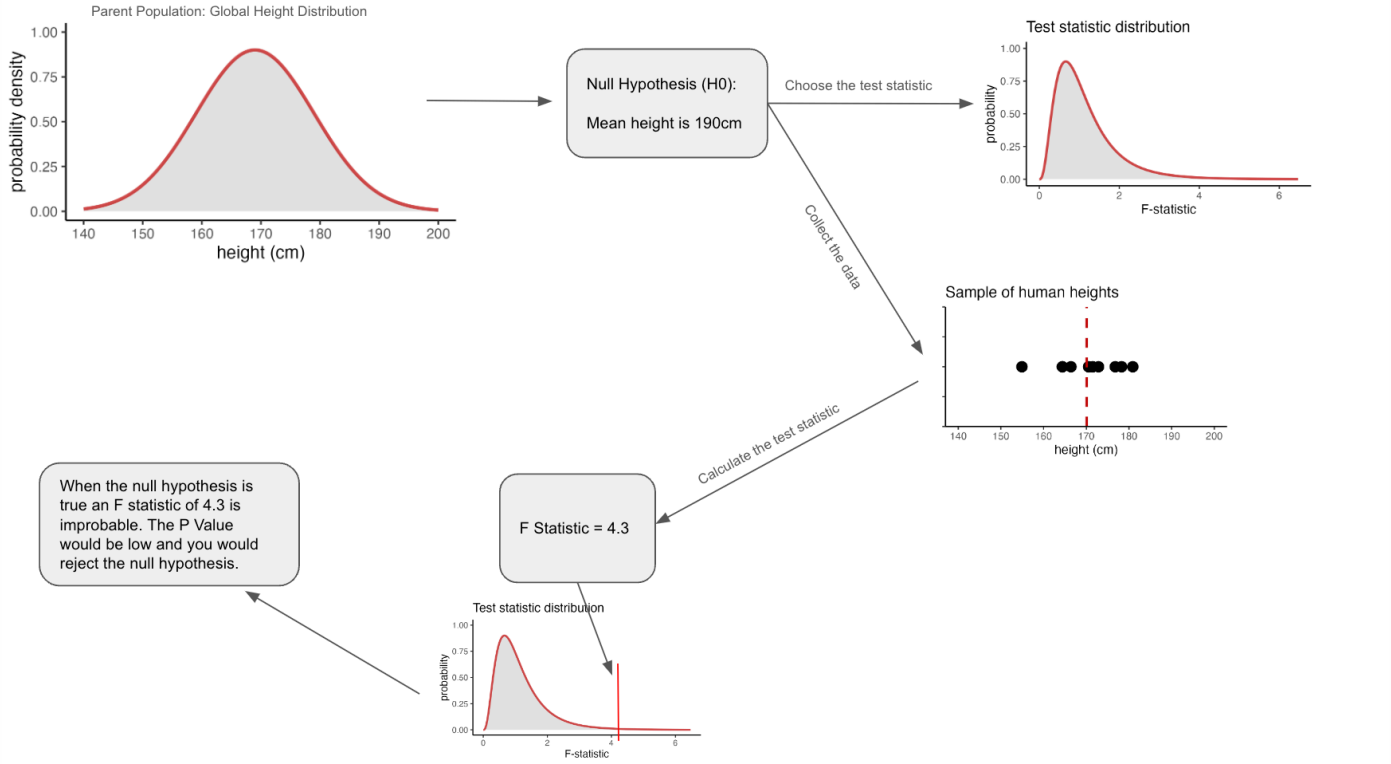

Figure 4. Process of classical hypothesis testing - the numbers/graphs in the figure are made up and NOT accurate.

Hypothesis Testing in Practice

In practice, the process is generally a bit different to what you see in classical hypothesis testing. Typically, researchers will choose a statistical test based on a dataset, and the research question they have. It is relatively common for researchers to start with a dataset and to work from there. In this situation, researchers will often generate their research question from the sample and then, select a statistical test that fits the data and research question. Also unlike classical statistical testing, the statistical test will be executed computationally. This makes the process very simple where researchers simply need to select the test and they will receive the relevant outputs (including the test statistic and p-value). Here, the statistical testing itself is easy, it is the selection of an appropriate statistical test that is challenging.

4. Choosing a Statistical Test

When selecting a statistical test there are generally three things that need to be considered; what is the question you want to answer, what type of data do you have, and what are the assumptions of the test you plan to use.

The Question & The Data

Unsurprisingly, the statistical test a researcher chooses will need depend on what the statistical test is trying to answer. The type of question you are trying to answer will also depend on the type of data you have. There are broadly two types of statistical tests: tests that predict and tests that identify differences. Both of these types of tests are forms of linear models, which fit lines of best fit to datasets to summarise linear relationships. There are non-linear models as well but they will not really be discussed here.

Predictive Statistical Tests

Predictive tests are typically used when all/both of the variables are continuous. An example would be looking at the effect of height on BMI. These types of statistical test build a **model** to describe the dataset. The term model and be used interchangeably with the term **line of best fit**. The function of the model/line of best fit is to capture the signal or overall pattern/trend in the data, and to discard the noise (see figure 5). If a good line of best fit has been produced, it could be used to predict a y value for a given x value.

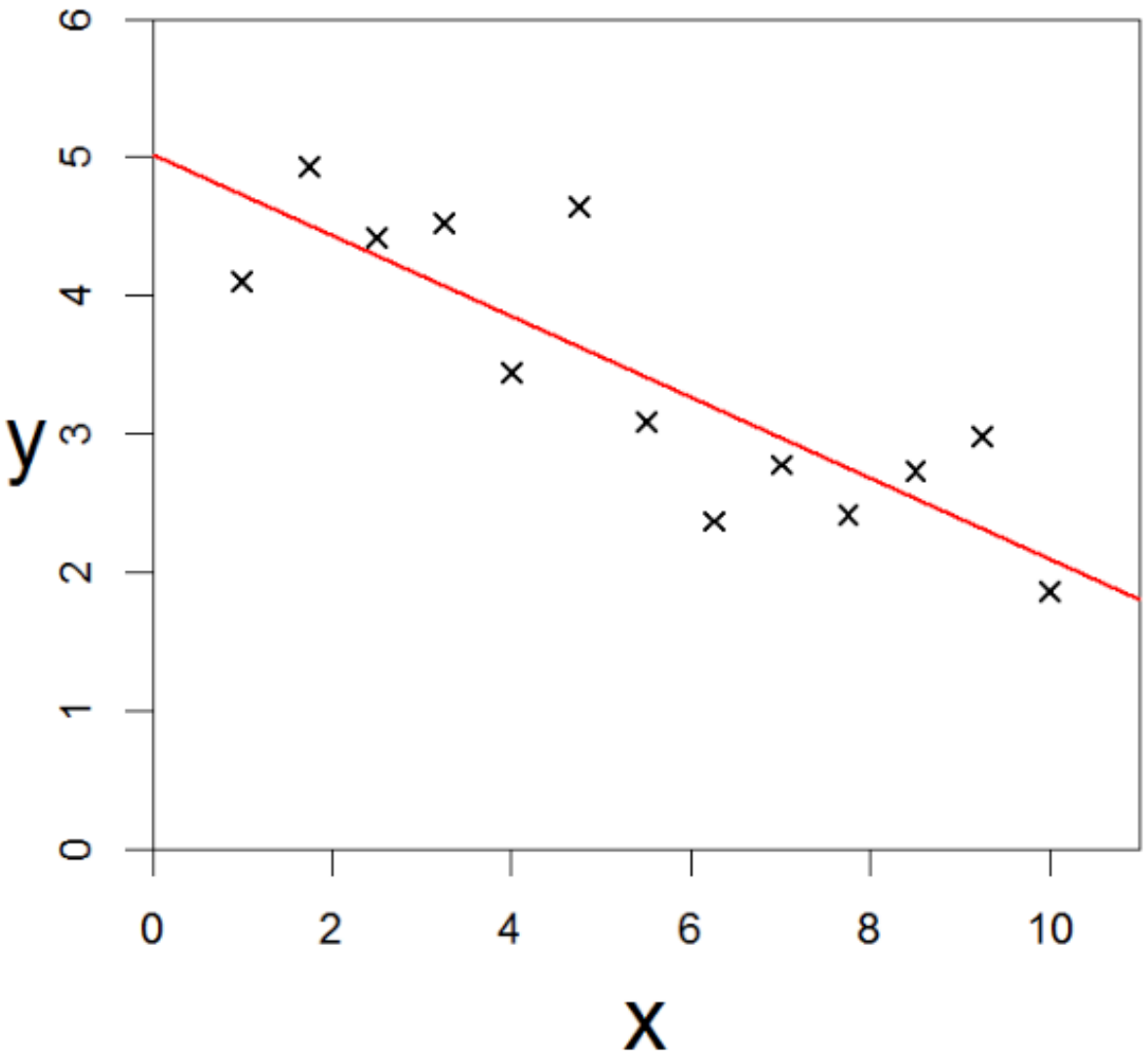

Figure 5. An example line of best fit where x is a continuous predictor variable. The crosses are individual data points and the red line is the line of best fit/the model.

The line of best fit is constructed by looking at the **residuals**. The residuals are the difference between the real value and the model value on the y-axis. This can be thought of as the distance of the cross/datapoint in figure 5 from the red line in the y direction. This is visualised in figure 6 where a line has been drawn from each data point to the line of best fit, the length of each line is the residual for that data point. The term **error** is often used interchangeably with residual and this does make sense intuitively. The distance of a datapoint from the line of best fit could be thought of as the error the line of best fit has made when predicting that datapoint.. So why are the residuals important?

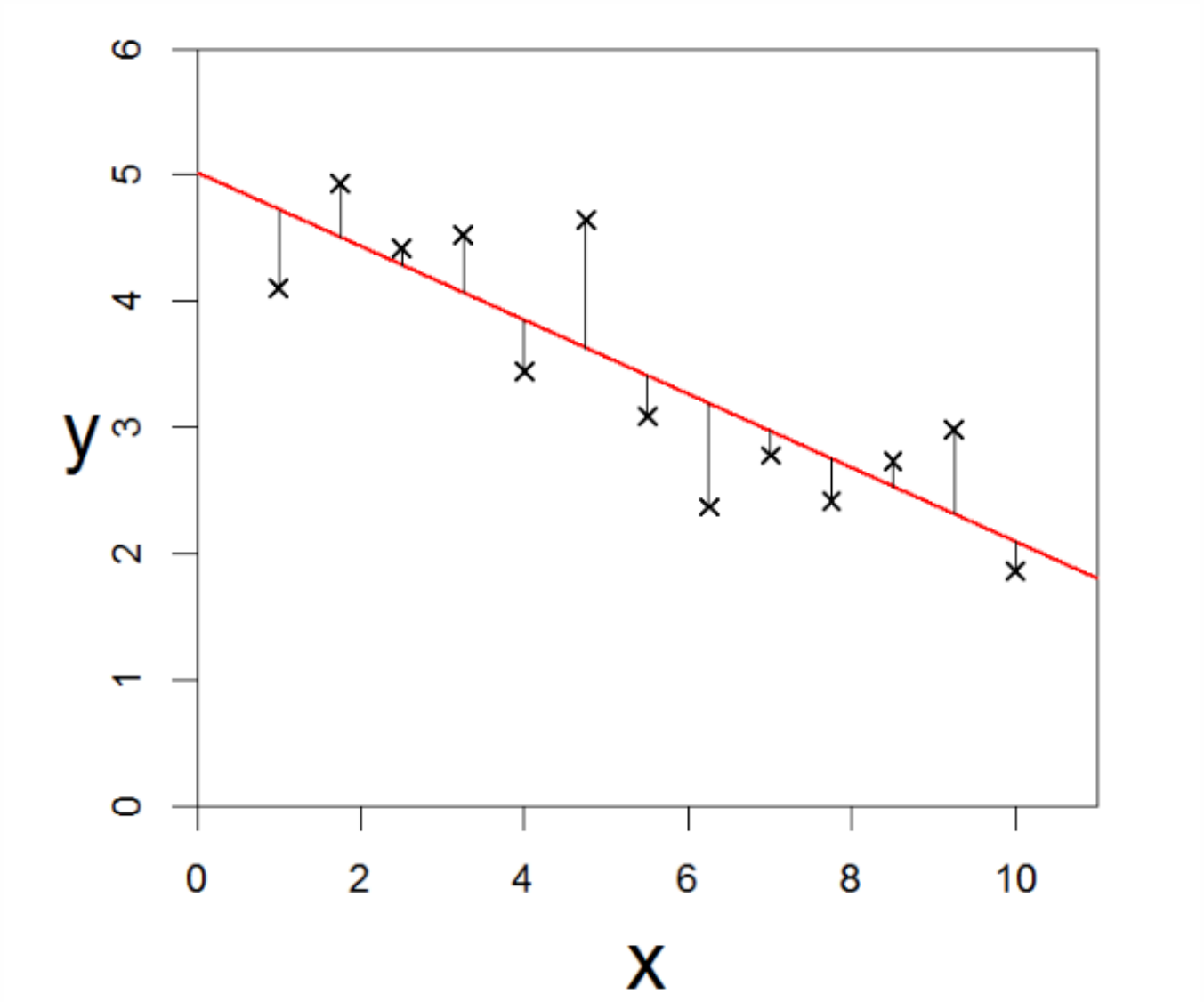

Figure 6. The residual for each data point is the distance of the data point from the model/line of best fit (red line) in the y direction.

The residuals are used to determine what line best fits the data. The residuals are used as a success metric for the line for best fit and allow multiple possible lines to be compared. Conversely, the best model can also be mathematically calculated using the errors as well. For the errors to be used as a metric that can be optimised they need to be condensed into a single value. To account for the fact that the errors/residuals can be positive or negative, they are squared and then all of the errors are summed together, providing a single number that represents the success of a model/line of best fit. This number is called the **sum of squares** or **residual sum of squares**. The best line of best fit will have the smallest residual sum of squares. Computationally it is very easy to just test out hundreds or thousands of lines and just optimise for the residual sum of squares. Now that the best possible line of best fit has been created, what do you do with this?

The next part of a statistical test like this is a **regression analysis**. A regression analysis is conducted to determine the significance of the line of best fit. Figure 7 shows two lines of best fit that are the same, they have the same gradient and intercept, but it should be clear from looking at them that the model on the right is a better description of the data. In this example, the line of best fit in the graph on the right would be more significant. It is this level of significance that a regression analysis is calculating. How does it do this?

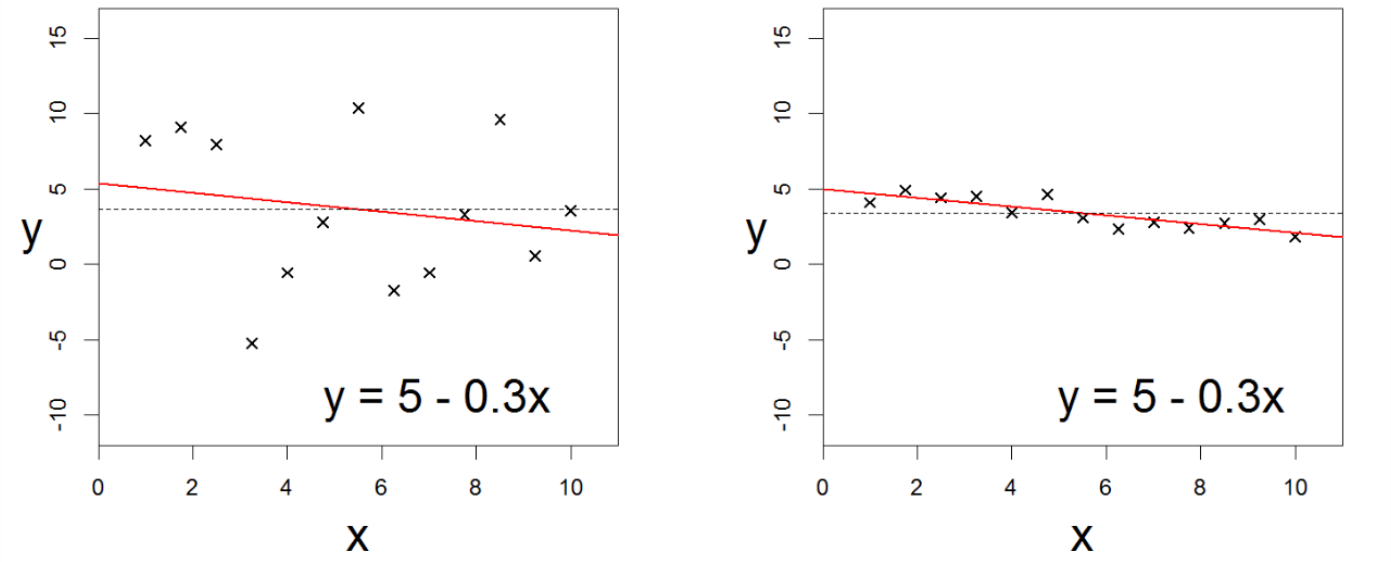

Figure 7. Two lines of best fit that are the same, but have differing levels of significance. The p value for the model on the left is 0.56 and the p value for the model on the right is 0.000263.

The way a regression analysis works is by adding a line to the graph with no gradient and a y value equal to the mean value of y in the data. The regression analysis checks that the line of best fit outperforms this flat line where y is the mean value of y. The extent to which the line of best fit outperforms the flat line determines the p value or how significant your predictions were.

The actual calculation is done by calculating the error/residual for each of the predicted data points, and then the **variance** of these residuals. Here variance refers to the spread of the data around the mean value. The variance of the predicted data points is then also calculated, this variance calculation can be thought of as a comparison of the predictions to the flat line with the mean y intercept (see graph B in figure 8). This flat line (with a y value equal to the mean value of y) can be thought of as the null hypothesis, and the line of best fit as the alternative hypothesis. If the line of best fit has a large gradient, then the variance will be high, and this indicates a strong **signal** in the data. If there is a large variance in error, a lot of data points sitting far away from their predicted values, then this can be thought of as a lot of **noise** in the data (see graph A in figure 8 for a representation of the noise in the data). Since the p value is calculated by dividing the variance in residuals by the variance in predicted values, the p value can be thought of as a signal to noise ratio. This is true for most statistical tests. Graph C in figure 8 shows the variance in predicted values on the right and the variance in error on the left, the ratio of these two values gives us the p value and in this example we can see that the p value should be low due to the minimal noise in relation to the large signal.

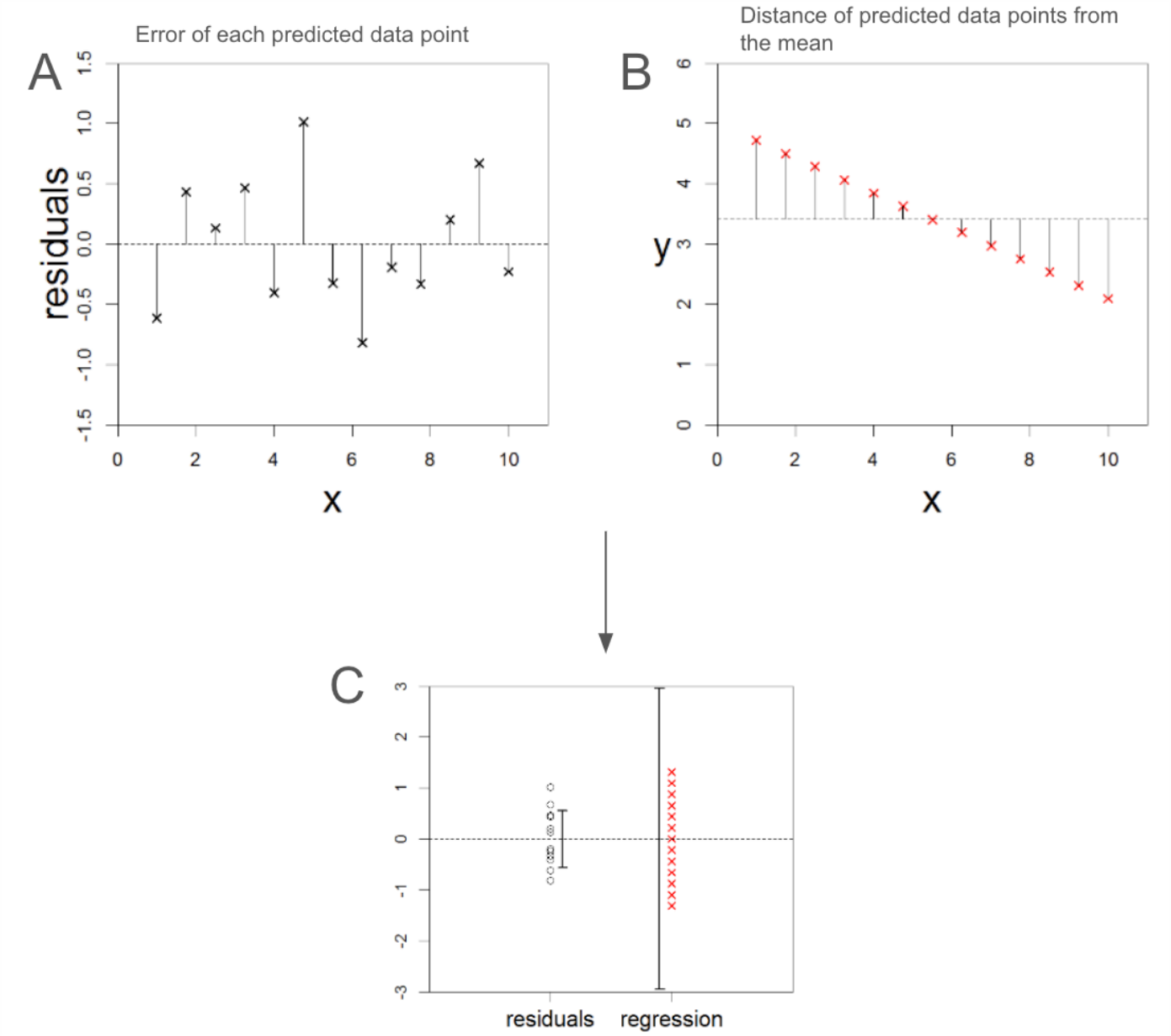

Figure 8. Graph A shows the error of the predicted data points. This graph can be considered as a representation of the noise in the data, or at least what the model considers to be noise. Graph B shows the distance of each predicted data point from the mean value of y, where the mean value of y acts as the null hypothesis, and gives a visual representation of the gradient of the line of best fit. The greater the distance of the predicted data points from the line, the greater the signal in the data. Graph C shows the variance in the residuals and the variance in the predicted values, the ratio of these values is used to derive a p value.

If your p value is within the significance threshold, the result could be reported with the following statement: the slope of the line of best fit is significantly different from zero. This would mean that a change in x would result in a significant change in y. An additional note, is that since the variance accounts for the number of residuals/data points, the p value that is produced is not naive to the number of data points in the sample.

When you have implemented a linear regression as described above, you can also look at the **correlation** between the variables. The term correlation describes the strength of association between variables and always takes a value between -1 and 1. The two extreme values, -1 or 1, describe either a perfect positive or negative association. When considering the correlation between variables, we are uninterested in the gradient of the line of best fit and are instead interested in the degree of scattering of points around the line. To e certain degree, it can be thought of as the amount of noise in the relationship between two variables where 0 is 100% noise and -1 or 1 would be 0%

noise. The correlation can only be calculated between a maximum of two variables. A commonly used measure of correlation is the **Pearson product moment correlation coefficient** and this is denoted by **r**. This is an estimation (statistic) of the true population correlation and will vary between -1 and 1. The value that is more useful and more frequently reported is **r squared (r^2)**, which is the amount of variance in the data explained by linear model that we have fitted. An r^2 value of 0 implies that none of the variance has been explained and an r^2 value of 1 implies that all of the variance has been explained by the model. For example, and r^2 value of 0.72 suggest that the model explains 72% of the variance in the data.

In summary, a simple linear regression involves three components:

1. Fit a linear model to determine how much one variable affects the other.
2. Conduct a regression analysis to determine whether there is sufficient evidence to conclude that one variable is actually affected by another.
3. Calculated the r^2 to determine how much of the variation int eh data is explained by your model.

Categorical Statistical Tests

When researchers want to look at two or more groups of samples with continuous response data, **analysis of variance tests (ANOVA)** are typically used. An example data set here would be the heights of 5 samples of 50 men with each samples coming from men born in a different country. Another example would be the abundance of a specific metabolite, metabolite x, in two samples of 30 people where one sample is made up of individuals with lung cancer and the other sample contains individuals without lung cancer. For this kind of data, the statistical tests would be used to determine whether the samples came from parent distributions with the same mean. When there are only two samples/groups an ANOVA can be used but a t-test is typically used instead. An ANOVA can be thought of as a generalisation of the two-sample Students's t-test. The purpose of these tests is to compare multiple groups and ask whether the means or medians of these groups are different, here comparing means will be the focus.

ANOVA considers the spread of the means of the groups and the spread of the data within the groups (variance). Figure 9 shows how the within group variance is relevant in deciding whether the groups are different. The larger the amount of within group variance, the harder it is to separate groups.

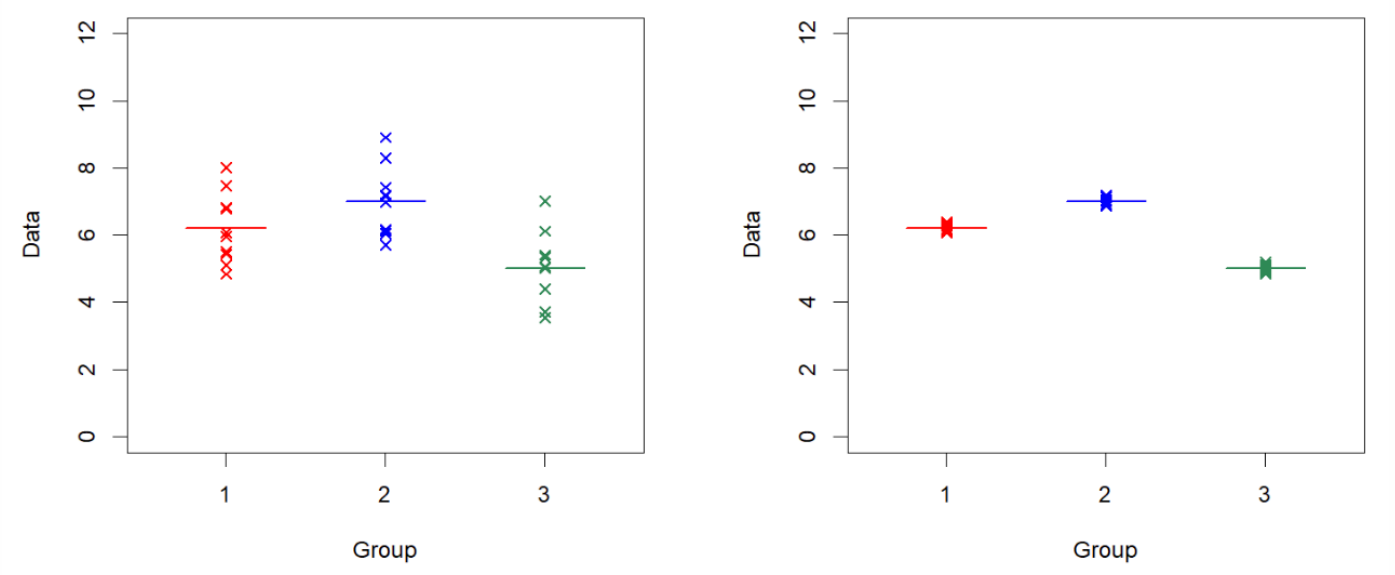

Figure 9. The graph on the left shows three groups that have a similar means and a 'medium' amount of within group variance. It is hard to say whether these groups are significantly different from one-another. The graph on the right has the same three groups with the same means and the same number of data points, the only difference here is that the variance within each group is extremely small. In the graph on the right it more likely that the groups are significantly different from one-another.

Similarly, the spread of means is also important when considering whether groups are significantly different from each other, and this is shown in figure 10. The smaller the spread of the group means, the harder it is to separate groups.

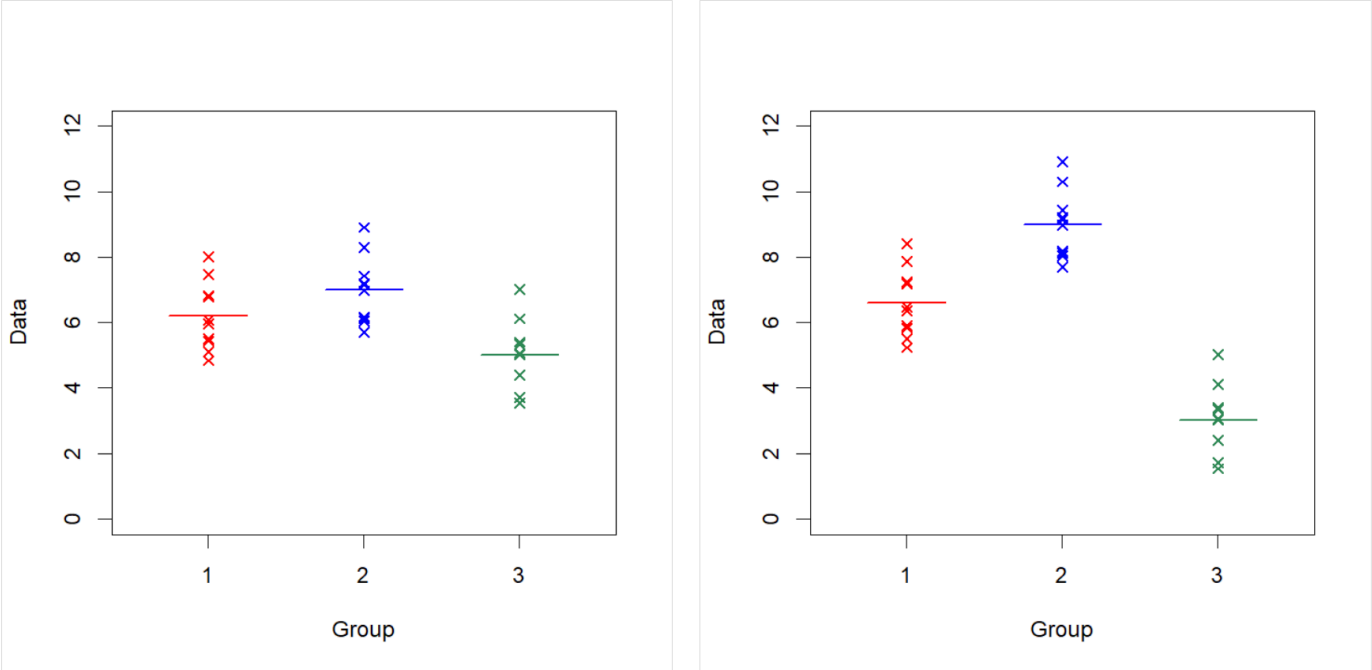

Figure 10. The graph on the left shows three groups that have a similar means and a 'medium' amount of within group variance. It is hard to say whether these groups are significantly different from one-another. The graph on the right has the same three groups with the same within group variance and the same number of data points, the only difference here is that the means of the groups are much further apart. In the graph on the right it more likely that the groups are significantly different from one-another.

It should now be clear why the group means (between group variance) and within group variance are important in deciding whether groups are different from each other, but how does ANOVA (and other tests) use these values? Two numbers need to be calculated, one to represent the spread of the means and one to represent the within group variance (the spread of the data within a group). In an ANOVA test, the between group variance is measured by comparing the combined mean for all groups with the means of each group (simplification but essentially what is happening). To measure the within group variance for each group, you compare each datapoint with the mean of that group. The between group variance can be presented on a graph like in figure 11, with the groups means being shown on the right hand side in a more simplified manner.

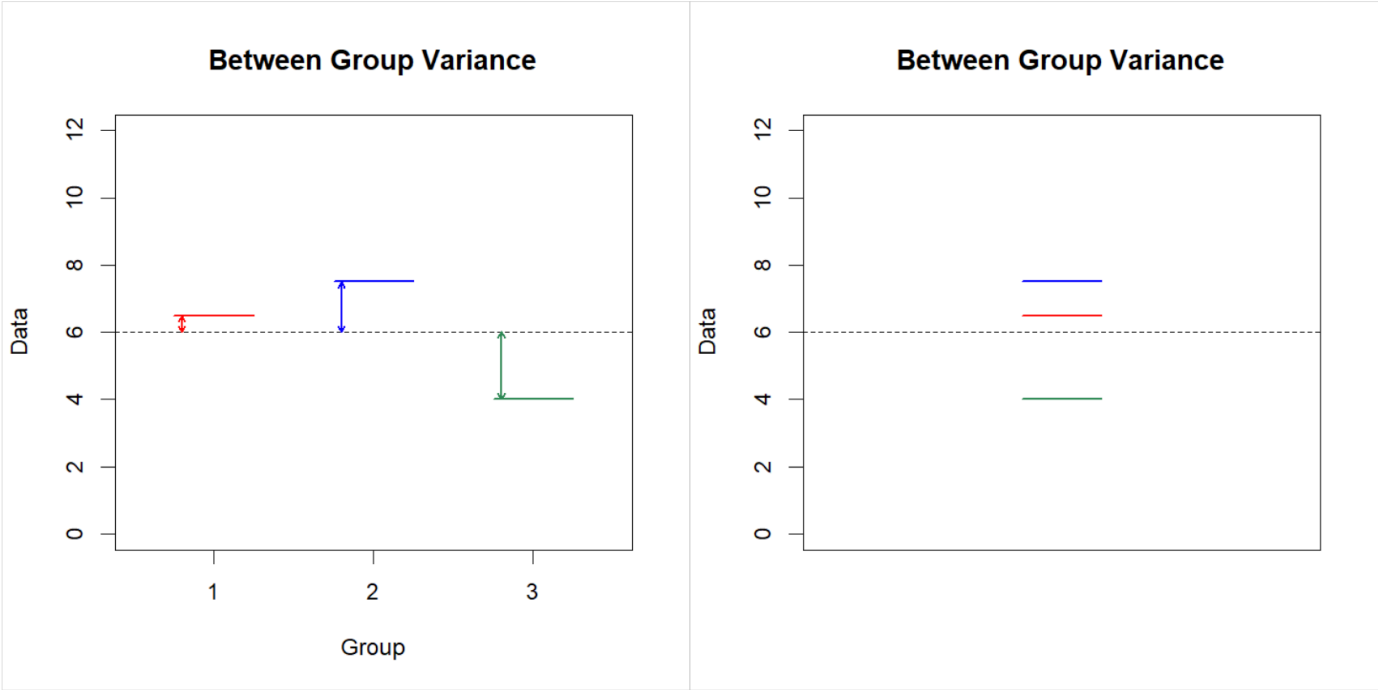

Figure 11. The spread of group means, also known as the between group variance. The dotted line is the overall mean of the groups. The between group variance can be visualised in a more simplified manner as shown in the graph on the right where the groups are purely denoted by colour rather than being in separate columns as well.

To measure the within group variance, for each of the groups is calculated and then these three variance values are added together. To easily visualise and easily compare the amount of variance in each group you can plot the spread of data around each group mean where the mean for that specific group is set to zero and this can be seen in figure 12.

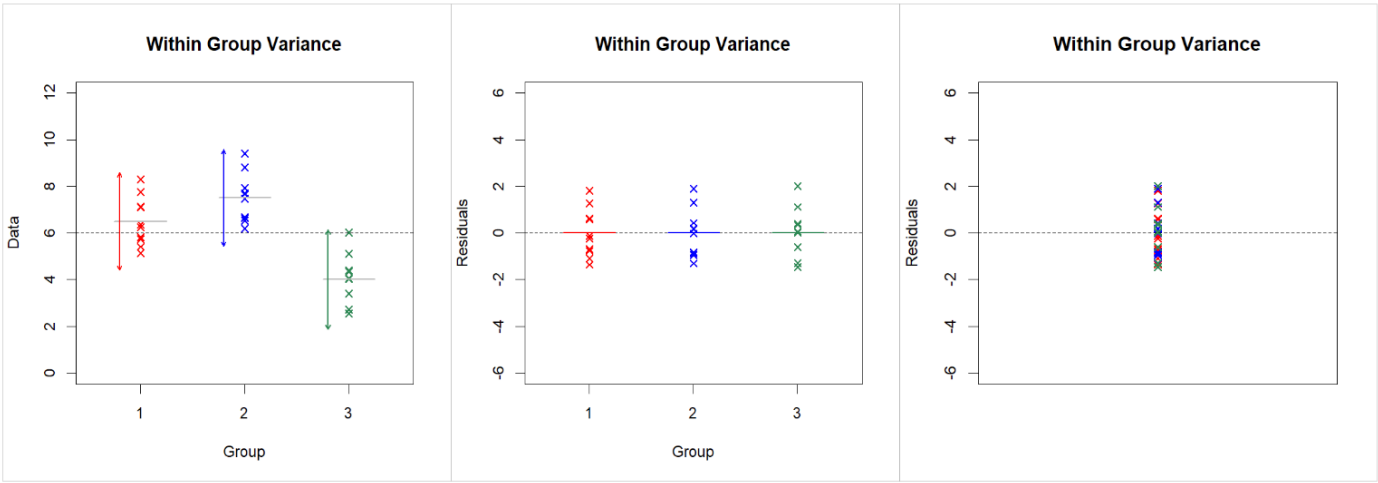

Figure 12. Graphs to represent the within group variation of three different sample groups. Moving from left to right, each group is a simplification of the former to all ow for easier comparison of the variance in each group.

Now we have two numbers, the between group variance and the within group variance, what do we do with them? We cannot compare them yet as the between group variance figure does not account for the fact that there are multiple samples in each group. A correction needs to be applied to address this and in this case, the correction is simply multiplying the between group variance by the number of samples in each group (this is actually embedded within the variance calculation and the multiplication done on a per group basis). This correction is represented in figure 13.

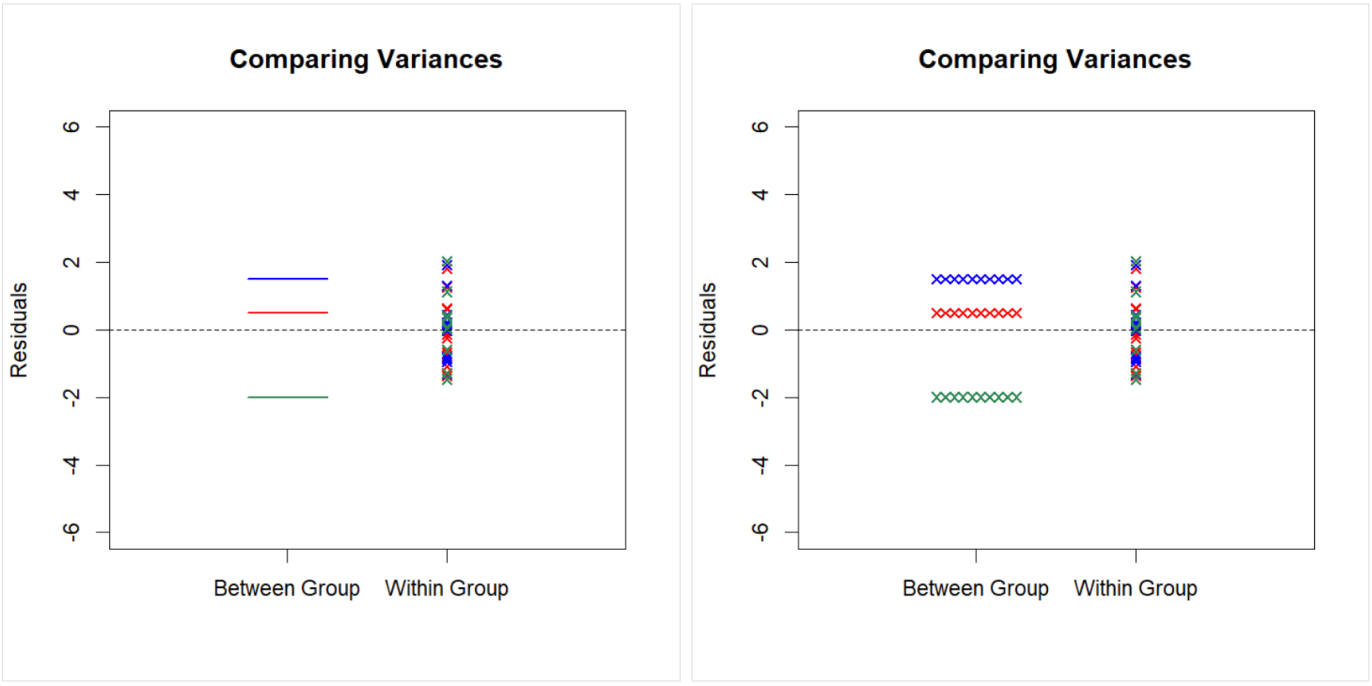

Figure 13. Visual comparisons of the between group and within group variances. The figure on the right represents the correction that is applied to account for the sample size of each group when looking at the between group variance.

Now that the between group variance has been corrected, this can be compared with the within group variance to generate the F statistic. The F statistic is calculated by dividing the between group variance by the within group variance. This F statistic can be thought of as a signal to noise ratio, and if this ratio is big there is a significant effect. How do we know whether this ratio is big or not? The F statistic is an absolute value which does not scale in a way that allows for comparison between tests, so we need a way to know whether the F statistic if big enough for us to conclude there is a significant difference between groups. Again, the p value is used here and it is generated by looking at an F statistic probability distribution to determine the likelihood of seeing the observed F statistic under the null hypothesis. Figure 14 shows the F statistic for the example we have been looking at.

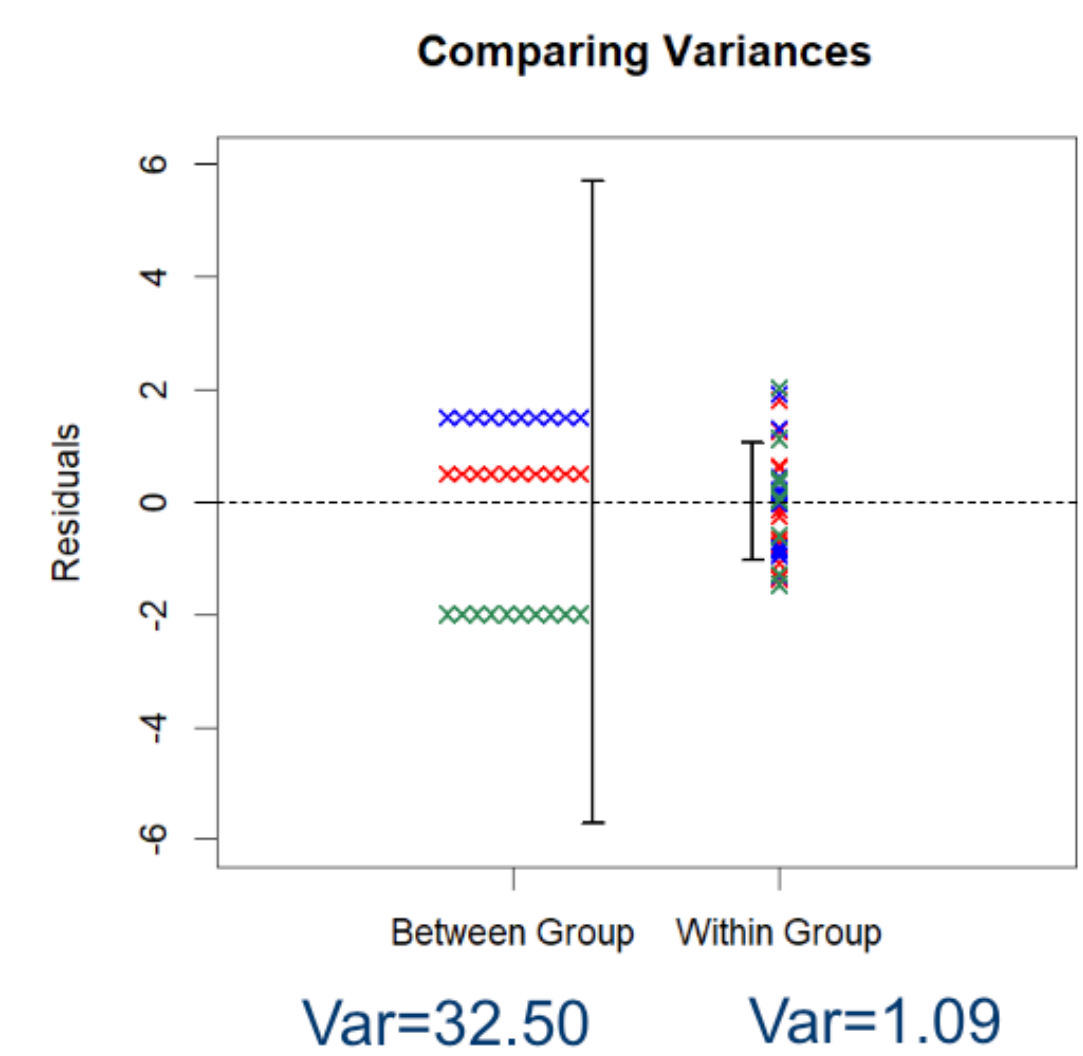

Figure 14. The F statistics for this data is 32.5/1.09 = 29.7. This number alone does not tell us much, but the ANOVA test will generate a p value which in this case is 0.000000153.

The p value for the data in figure 14 is 0.000000153 which means we can reject the null hypothesis that the is no difference between the three groups. It is important to know that this means you have only confirmed that **at least one** of the groups have been drawn from a population with a different mean to the mean of the populations the other groups have been drawn from. So, if you have 15 groups and 14 are from the same population and 1 is from a different population, you are very unlikely to be able to detect that that one group is different (unless that group is massively different). Consequently, it is not recommended to use ANOVA with too many groups.

Assumptions of Statistical Tests

Statistical tests are specific tools designed to be used for specific types of data. A danger of working with statistical tests is that when they are used incorrectly, they will still output results. Consequently, it is extremely important to understand the limitations of each statistical test you might use and the assumptions the test make about the datasets. A lot of the criticism that statistical tests receive is due to the misuse of the tests due to a lack of understanding about the assumptions they make.

All statistical tests make assumptions about the data. If these assumptions are not met, you cannot have any confidence in the results of the test!

There are some common assumptions:

- The population from which a sample is drawn follows a normal distribution and the samples themselves follow a normal distribution.
- The parent distributions have the same variance.
- Each data point is independent from all off the others. E.g. if each data point represents a person, no relatives could be included.

Parametric Tests

The assumption of normality is one of the most common and important assumptions of statistical tests. Tests that assume normality are called **parametric tests**. When working with a parametric test, the sample data doesn't always necessarily need to follow a normal distribution. In practice, at least one of the following requirements need to be met:

- The assumption of normality is met.
- If the data in the samples does not follow a normal distribution, a transformation can be applied to make the data normal (e.g. a log transformation). With most tests it is actually the residuals that need to be normally distributed rather than the data points themselves, so a transformation would need to be applied so that they become normally distributed.
- The sample(s) is sufficiently large enough. This is an annoying and ambiguous caveat. What does this actually mean? As a general rule, if there are more than 30 observations in the sample then you don't need to worry about the distribution. I have no references to validate this though. The less normal the data, the more observations you would need. If the data clearly follows another distribution, then often it could be better to use a non-parametric test.
- For all linear statistical tests (all that have been discussed so far), the relationships between the variables must be linear.

Generally, if you can use a parametric test, then you should. If your data does not meet any of the above requirements, then a **non-parametric test** can be used. Non-parametric tests do not assume a normal distribution (or even any distribution) but they do have their own assumptions. These types of test can be particularly useful when you have ordinal or ranked data which cannot be normally distributed, or you have a small non-normal sample which cannot be transformed. **Outliers are important and should not be removed when trying to fit a distribution.** Non-parametric tests tend to be more conservative and have less power to detect differences. Commonly used parametric tests will normally have an equivalent non-parametric test that can be used when the parametric test is not suitable.

How to check whether your data meets the assumptions of a test?

The results of a statistical test should **never** be reported without evidence that the assumptions of the model have been met! Unless the reviewers have the code that was used AND the exact dataset used, it is not possible for them to assess (review) the use of a statistical test. **A statistical test cannot, and should not, be trusted if** diagnostic plots (or some other validation technique) have not been provided. Where possible plots of the model should also be included.

Bartlett's test is the most common way of testing the equality of variances between samples. It is important to note when using Bartlett's test that you are looking for a p-value greater than 0.05. This is because the null hypothesis of the test is that the population variances are equal.

A common way of checking whether data is normal or not is to use the Shapiro-Wilk test. It is probably best to not use this test. The problem with the Shapiro-Wilk test is that it is not very reliable. Generally, if you have a small sample the test will conclude it is normal and if you have a large sample it will conclude the sample is not normal. Instead of using the Shapiro-Wilk test, there are four graphs that can be used. These graphs are collectively known as **diagnostic plots** and are typically shown as a panel of all four plots (as shown in figure 15).

The first plot is known as the **residuals plot** and is used to assess that there is no pattern in the residuals. The residuals should be distributed randoms on the graph with an equal number above and below the red line. The red line itself should appear flat and is a representation of the trend in the data. If there is a curve in the data or in the red line, then you are seeing a sign of a non-linear relationship.

The second plot is the **Q-Q plot** which shows how closely the residuals follow a normal distribution. In this plot, the blue line shows a perfect normal distribution and you hope to see that the data points are close to the line. It is important to remember that the residuals do not need to follow perfect normal distribution and some 'peeling' at the ends is expected.

The third plot is the **location-scale plot** which shows the cloud of data points. Here you are looking to see that the variance is staying the same across the samples or range of predicted values. Like the residuals plot, you are looking for the red line to be mostly flat.

The fourth plot is the **Cook's D plot** which is used to check for influential points. You do not want a linear model fitted as part of a statistical test to be massively skewed by a single data point. Ideally the influence of each data point on the model should be similar. The Cook's D plot shows how influential each data observation is on the model (and thus the statistical test) and a general rule is that any observation that has a Cook's D value of 0.5 or higher needs to be investigated further. These plots work for both categorical and continuous data because even when working with categorical data (like when using an ANOVA), a linear model is still fitted to the data. These plots are the 'gold standard' in checking that datasets meet the assumptions of a model and should be used without exception.

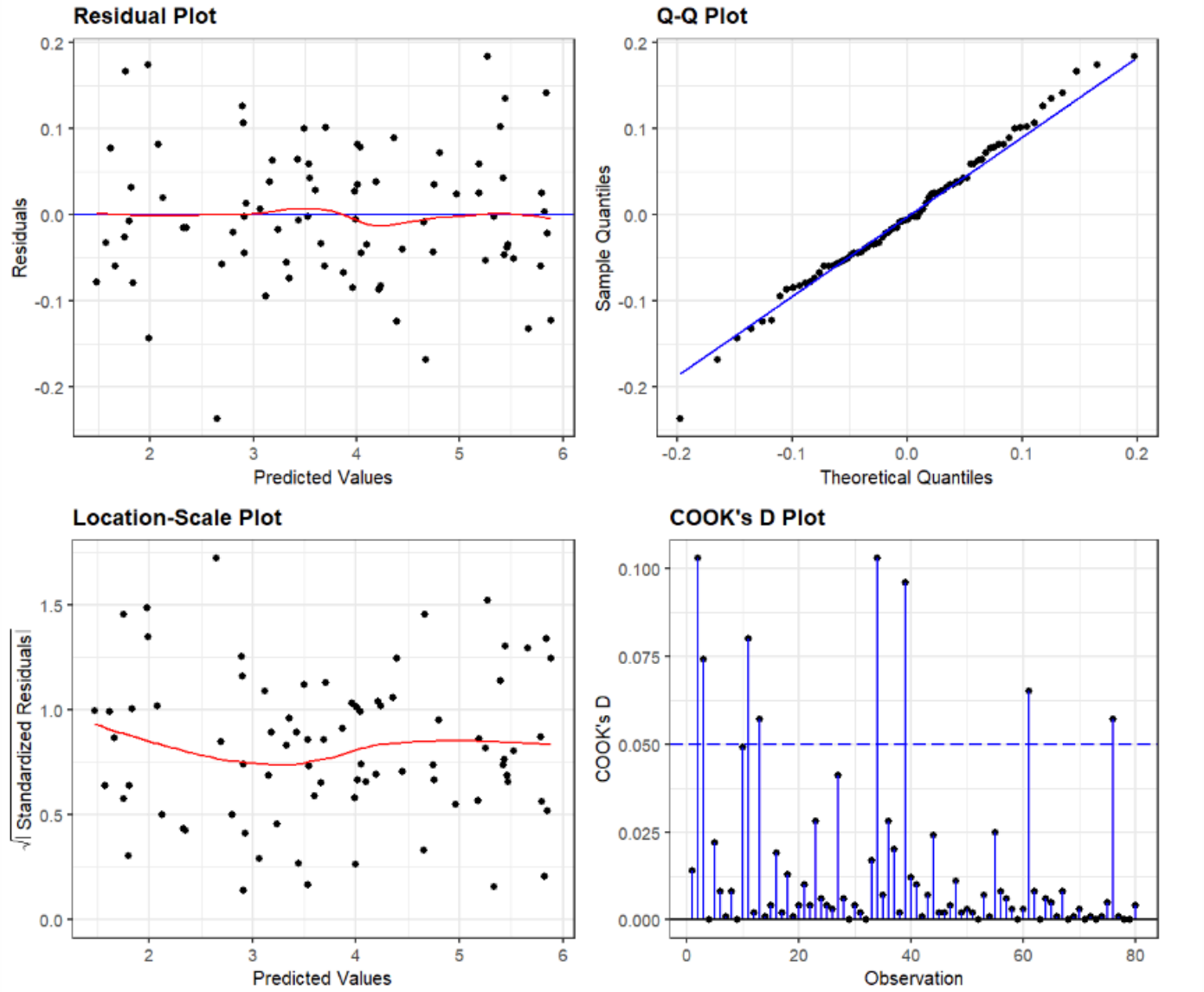

Figure 15. A panel of the four diagnostic plots typically used to assess the assumptions of statistical models are met. The panel above is what the graphs should like like for continuous data (like when using a t-test) but it should look similar for catagegorical dat you might find when using an ANOVA. The difference for categorical data would be that the Residuals plot and the Location-Scale plot would have the data points clustered along the x-axis by group. You would still expect the same red line shape.

The key takeaway from this section is that researcher **ALWAYS** need to check that the assumptions of a statistical test are met **AND** these checks need to be reported/published with the results.

5. Probability & P Values

All statistics can be wrong! Statistical tests are probabilistic in a nature and this means there will always be a chance that they are wrong, even when all of the assumptions of a test have been met. This needs to be kept in mind when using statistical tests. A key part of doing this is ensuring that you correctly interpret p values.

P values are a weird concept and can be hard to understand. It is common for people to think of a p value as the probability of the null hypothesis being true. This is incorrect. What a p value actually represents is the probability of seeing your data if the null hypothesis is true. This is a subtle but important distinction. The statistical tests that generate p values are based on the assumption of an underlying probability distribution, often a normal distribution (see figure 16). It is important to understand how these distributions are used to generate a p value.

Example:

Assume we know that the height of Uk men follows a normal distribution with a mean height of 169cm. We have taken a random sample of 50 Greek men and we want to know whether their height is different to the height of British men. The null hypothesis is that the two groups of men come from the same distribution, so our null hypothesis is that the mean height of greek men is 169cm. The alternative hypothesis is that the mean height of greek men is not 169cm. Here, the p value represents the probability of seeing the 50 heights we've sampled if we assume we've taken these observations from the distribution shown in figure 16.

As we can see by looking at the graph, our p value will be much lower if we observe a high proportion of the Greek men with heights over 195cm and very few below 185cm. We would see a low p value in this scenario because it is very unlikely we would randomly select 50 men with this range of heights from the distribution in figure 16. We would be far more likely to randomly sample a large number of men with heights between 165cm and 175cm. If this was what we had observed in the Greek men, we would have seen a large p value. Given this low probability of observing the heights we observed (lots between 185-195cm) from the distribution shown in figure 16 (a distribution describing the null hypothesis), we would likely reject the null hypothesis and conclude that the Greek men have a different underlying population distribution.

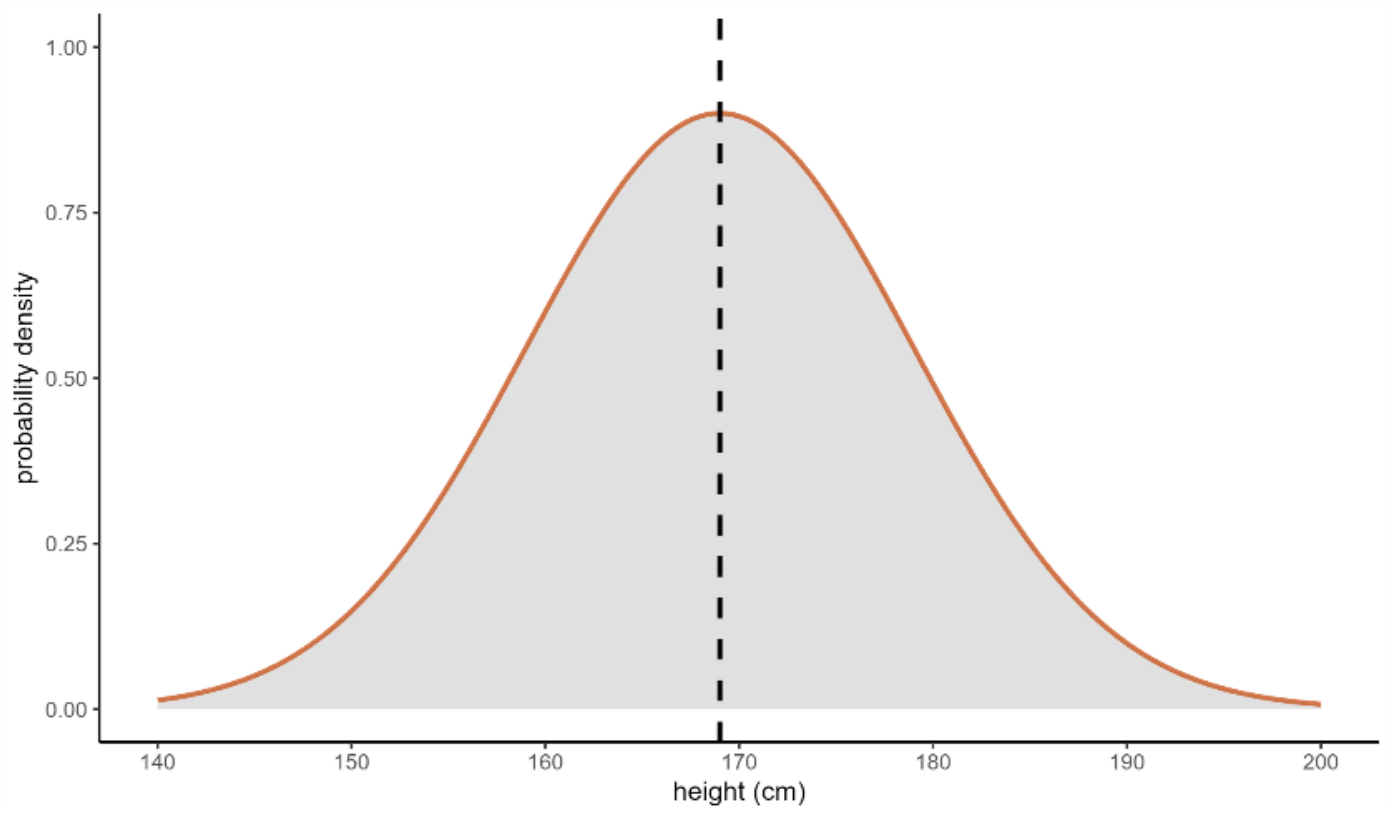

Figure 16. A normal distribution representing the height of British men (this is not a real.)

When reporting p values, researchers sometimes find themselves in the difficult position of having set a significance threshold (usually $p < 0.05$) but observing interesting findings that slightly exceed this threshold. Some argue that you should not report any results that do not fall with this threshold. There is logic to the idea that a threshold is a threshold, and making any kind of exception calls into question the whole idea of significance testing or setting a threshold in the first place.

However, since a p value is a probability, it is reasonable to report interesting findings that do not meet the significance threshold but still have a relatively low p value. This might be particularly true when working with pilot studies that might have small sample size. Insignificant findings might become significant if the work is repeated with a larger sample. In this sense, the p value cutoff or significance threshold is somewhat arbitrary as the context of the work can often change what kinds of probabilities are might be 'interesting' or even replicable.

As a result, in my opinion, reporting findings that are not significant is fine. This is only true when the p values of these findings are reported honestly and they are not presented in a misleading manner.

6. Experimental Design

All hypothesis testing involves making a decision: is a result significant or not? The decision can be wrong in two different ways:

1. You reject the null hypothesis when the null hypothesis it is actually true. This is known as a **False Positive** or **Type I Error**. Here you have likely concluded something interesting is going on when the reality is boring.
2. You do not reject the null hypothesis (accept the null hypothesis) when the alternative hypothesis is actually true. This is known as a **False Negative** or **Type II Error**. Here you might have accepted that the realty is boring when something interesting is actually going on.

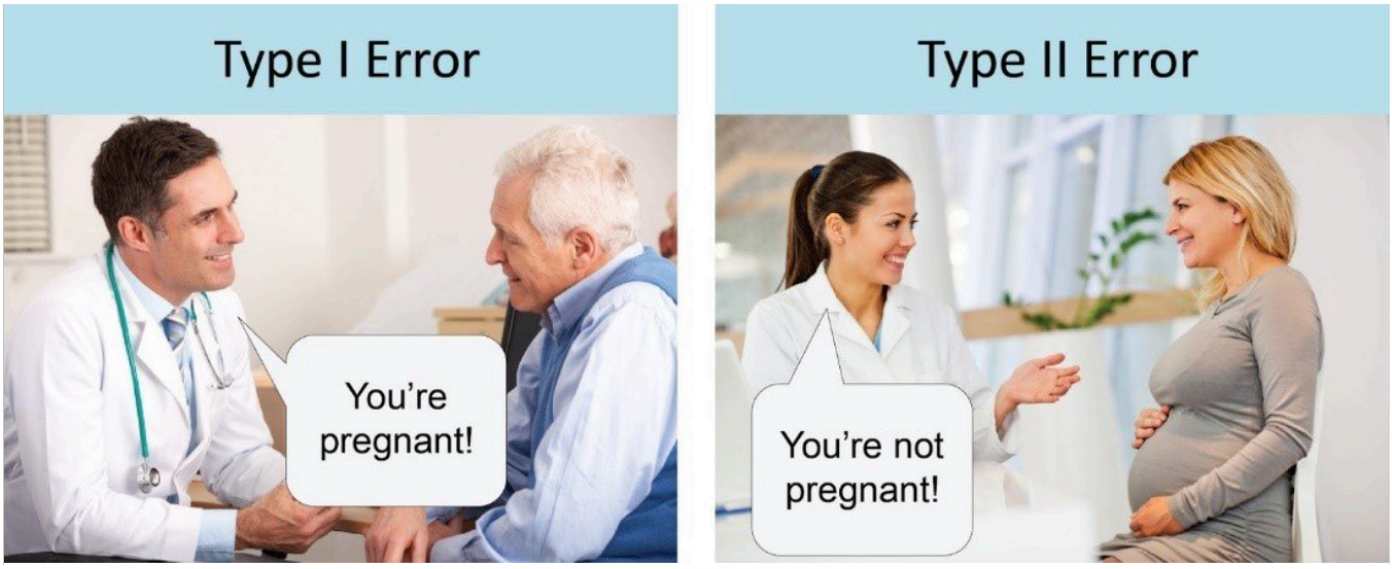

Figure 17. Error types.

We know that the p value is the likelihood of obtaining the specific dataset in question given the null hypothesis is true. So, the probability of encountering type I error is essentially equal to the p value. Consequently, it seems simple to reduce the amount of type II error, simply increase the p value threshold. Obviously researchers cannot do that, as there is no point (normally) to reduce the amount of type I error at the expense of type II error.

Power is essentially the probability of avoiding type II error or the probability of detecting an effect when it does exist. How can we increase the power?

1. Increase the size of the effect you are measuring.
2. Increase the significance threshold at the cost of increasing type I error.
3. Add more data points.

Increasing the size of the effect you are measuring is not really possible if you are trying to measure a specific effect. Normally the size of the effect will be static and something you are trying to capture. We have discussed why you don't want to increase the significance threshold. Why does adding data points increase the power? Adding data points or increasing the size of the sample(s) decreases the variance of the distribution(s). The decrease in variance of the distribution occurs because as you add more data points/observations, the higher the proportion of observations that are close to the mean. For each observation you have added in the tail of the distribution, you might add 10 observations close to the mean. Decreasing the variance in the distribution of data increases the statistical power because, for groups with differing means but an overlap in the tails of the distribution, the proportion of data points/observations that overlap between the two groups is reduced (this is visualised in figure 18).

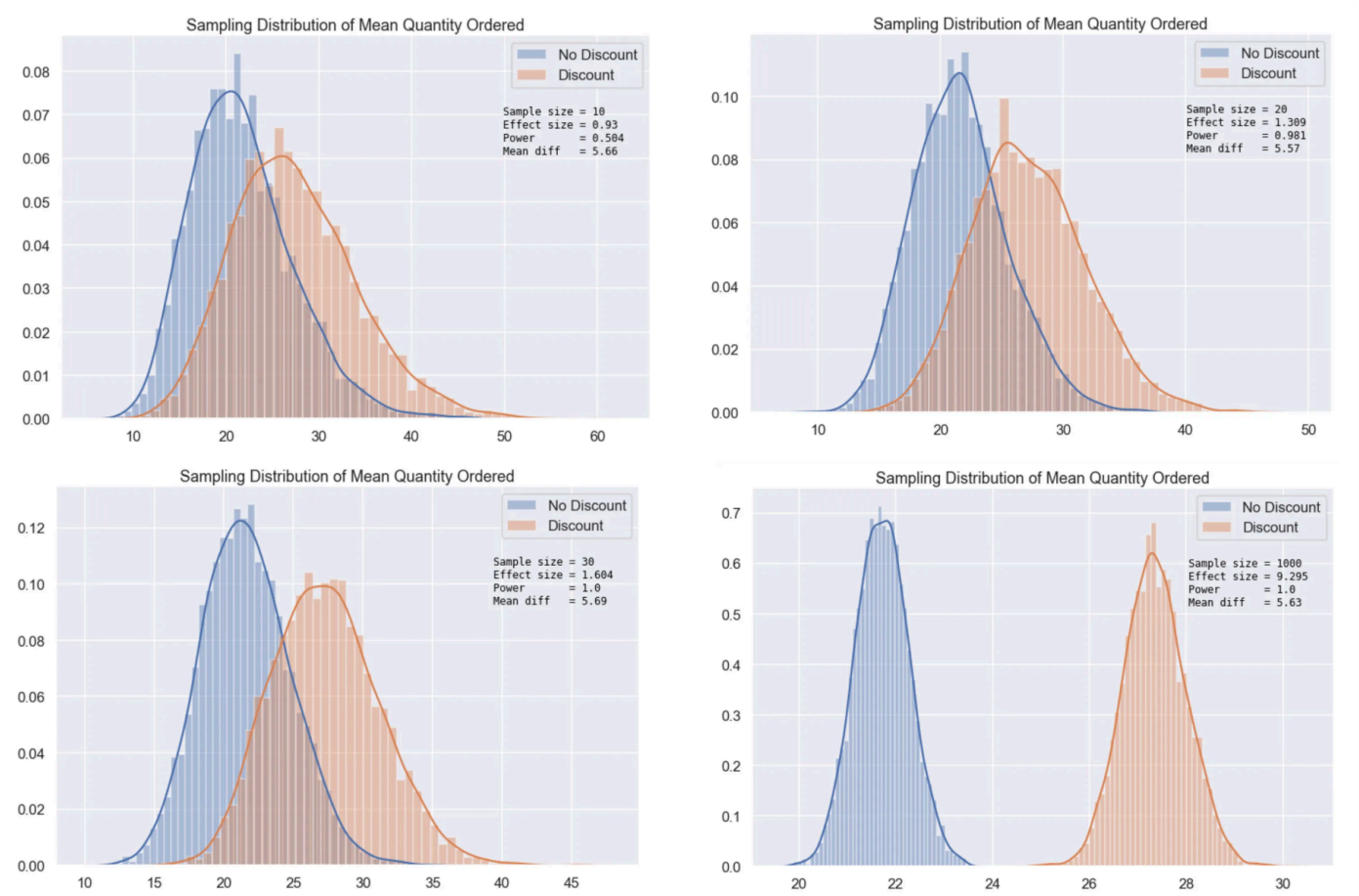

Figure 18. The four graphs in the panel show the probability distributions for the amount of a specific item ordered from a shop when there is a discount and without the discount. Each graph has a different sample size and the panel shows visually how increasing the sample size allows for better distinguishing between group means.

Consequently, adding observations to your sample is typically the only way to add power to a study in the real world. While extremely rare in the real world, in theory having a sample size that is too large does run the risk of identifying noise as signal. With a huge sample size, tiny effect sizes can be resolved and these will frequently not be biologically meaningful.

How do you calculate the statistical power of a study?

The four concepts of **effect size**, **sample size**, **significance level**, and **power** are all linked with three of the four values being required to calculate the fourth.

A power analysis is typically an a priori analysis conducted before a study starts to determine how big the sample(s) need to be. Most studies set/require a power of 80% and a significance level of 0.05, which leaves the effect size. The effect size is often what makes power analyses difficult, especially in the case of novel research. The effect size itself can be different things, in the case of t-tests it refers to the difference between means (Cohen's D) but when using linear models it can also be the amount of variance explained (r^2). As a result, researchers first need to determine what their effect size metric will actually be for the research they'll be conducting. The effect size can be calculated individually for each predictor or be a total value for a of the predictors combined. Once the type of effect size is determined, you need to estimate the effect size of the phenomenon being studied. This estimation can come from a pilot study, the literature, or it could be the desired minimum biologically meaningful size.

It is important to note that the more complex a model becomes, either by adding additional variables/predictors or by extracting more parameters, the larger your sample needs to become. The phenomenon is known as the degrees of freedom, and is often used as a proxy for sample size when inputting the arguments for a power analysis. Typically, the degrees of freedom equals your sample size minus the number of parameters you need to calculate during an analysis.

Power analyses, as described above, can only really be conducted for linear models and the assumptions of these models must still be met. This kind of power analysis cannot really be done for non-linear models or mixed effects models (both discussed later). However, there is an alternative. Simulation.

The accessibility of computational power now means that brute force trial and error is often a good alternative to traditional statistical calculations. This is true for power analysis. There are several possible reasons someone might want to use simulation instead of traditional power analysis methods; a non-linear relationship/model is expected, mixed effects models will be used, or you have no idea what the effect size might be so you want to test the power with multiple combinations of effect sizes and sample sizes. The drawbacks of using simulation is that it uses much more computational power and that it's a relatively new idea so there are very few programs or packages that can make the coding simpler. However, implementing simulations to test the effect of combinations of sample sizes, effect sizes, and even model types on the power of the analysis has a LOT of potential. Simulation provides the researcher with a lot of flexibility and allows for the exploration of study designs that might not have been considered with more traditional techniques.

Multiple Comparisons: 'correcting' p values

When you perform a large number of statistical tests, some will have p values lower than 0.05 even if the null hypothesis is actually true. If you do 10000 tests then, with a significance threshold of 0.05, you will expect 500 of these false positives. A good example of when this might be relevant is when looking a 1000 metabolites in individuals with cancer and healthy control individuals. For each metabolite, a statistical test would be run to determine whether the levels of that specific metabolite were different between the two conditions. We would expect 50

metabolites to show as being significantly different when in reality they are the same between the two conditions.

There are two approaches to addressing this issue. The first is the idea of controlling the **family-wise error rate (FWER)** which refers to the aging of the pairwise significance level such that the probability of getting any else positive across all tests is less than 0.05. The compromise here is that this correction will reduce the power of each test. The most famous FWER correction is the Bonferroni correction which is equivalent to multiplying all of the p values by n (where n is the number of observations). By applying the Bonferroni correction, you will only see the extremely significant results and a lot if significant results will be lost.

The second approach to controlling the number of false positives in multiple comparisons is by controlling the **False Discovery Rate (FDR)**. This method is less stringent as it accepts false positives as an inevitable component of results. This method involves setting a maximum acceptable proportion of false discoveries, with the most famous technique being the Benjamini-Hochberg Procedure.

7. Specific Statistical Tests Discussed

All of the following statistical test use the same concepts as, or build on from what was explained in section 4 (choosing a statistical test). For each test, all assumptions must be met and not all of them will be discussed here so always ensure you look up the relevant assumptions for any test you might use. It is also important to keep in mind exceptions to some of the assumptions, like the decreased importance of normally distributed data for parametric tests when you have a sufficiently large sample size.

1. One-sample t-test

One sample t-tests are used when you have a single sample of continuous data. The test is used to determine whether the sample came from a parent distribution with a given mean. You are essentially just finding out whether the sample mean is "close enough" to the hypothesised parent population mean. If your data is normally distributed, then a one-sample t-test is an appropriate choice for a single sample of continuous data.

2. Wilcoxon signed rank-sum test

The Wilcoxon signed rank-sum test is also used on single samples of continuous data. Like the one-sample t-test, the Wilcoxon signed rank-sum test is used to determine whether the sample came from a parent distribution with a given median. If your data is not normally distributed and their distribution is symmetric then the Wilcoxon signed rank-sum test might be a better choice than the one-sample t-test. The Wilcoxon signed rank-sum test may also be a better choice when the sample size is very small.

3. Two-sample student's t-test

Two-sample student's t-test is used when you have two samples of continuous data where the aim is to find out whether the two samples came from the same parent distribution or not. This can essentially be reduced down to whether there is a difference in the means of the two samples. This is another parametric test and does assume a normal distribution.

4. Mann-Whitney U test

The Mann-Whitney U test is the non-parametric counterpart to the Two-sample student's t-test. It aims to find out whether the two samples have different median values in order to determine whether they come from the same parent distribution. It is non-parametric so it doesn't assume normal distributions, but it does assume that the shape of the two samples is the same.

5. Paired t-test

A paired t-test is used when we have paired data. Paired data refers to a dataset where each data point in the first sample can be linked (or is related to) a datapoint in the second sample.

For example, suppose we measure the cortisol levels in 20 adult females first thing in the morning and again in the evening. We want to test whether the cortisol levels differs between the two measurement times. This is paired data because in each sample, there is a measurement relating to the same subject.

This is a parametric test and the non-parametric equivalent which will compare the medians of the samples is the two-tailed Wilcoxon signed rank test.

6. ANOVA

Analysis of variance or ANOVA is a test than can be used when we have multiple samples of continuous data. Whilst it is possible to use ANOVA with only two samples, it is generally used when we have three or more groups and a t-test used when there are only two samples. It is used to find out if the samples came from parent distributions with the same mean. It is a parametric test. Generally, the more samples we use, the less effective the test will become. It is only telling you is at least one of the samples comes from a parent distribution with a different mean. With an ANOVA, the category that separates the different samples is treated as the predictor variable, making the test a one-way ANOVA.

7. Kruskal-Wallis

The Kruskal-Wallis one-way analysis of variance test is an analogue of ANOVA that can be used when the assumption of normality cannot be met. In this way it is an extension of the Mann-Whitney test for two groups.

8. Linear Regression

Linear regression is used to investigate the relationship between two continuous variables. Regression analysis not only tests for an association between two or more variables, but also allows you to investigate quantitatively the nature of any relationship which is present. This can help you determine if one variable may be used to predict values of another.

9. Two-way ANOVA

A two-way analysis of variance is used when we have two categorical predictor variables (or factors) and a single continuous response variable. For example, when we are looking at how body weight (continuous response variable in kilograms) is affected by sex (categorical variable, male or female) and exercise type (categorical variable, control or runner).

This aims to answer to questions:

1. Does either of the predictor variables have an effect on the response variable i.e. does sex affect body weight? Or does being a runner affect body weight?

2. Is there any interaction between the two predictor variables? An interaction would mean that the effect that exercise has on your weight depends on whether you are male or female rather than being independent of your sex. For example if being male means that runners weigh more than non-runners, but being female means that runners weight less than non-runners then we would say that there was an interaction.

If an interaction effect is found, you cannot then draw conclusions about each predictor individually. If no interaction effect is found, you can then look at the effect of each interaction individually. An interaction plot is the best way to visualise this.

10. Linear Regression with grouped data

A linear regression analysis with grouped data is used when we have one categorical predictor variable (or factor), and one continuous predictor variable. The response variable must still be continuous however.

When analysing these type of data we want to know:

1. Is there a difference between the groups? Does the continuous predictor variable affect the continuous response variable?
2. Is there any interaction between the two predictor variables? Here an interaction would display itself as a difference in the slopes of the regression lines for each group.

In this case, no interaction means that the regression lines will have the same slope. Essentially the analysis is identical to two-way ANOVA.

11. Multiple Linear Regression

The above can essentially be scaled to as many variables as you like. You can create a linear model (line of best fit) for any number of interactions. However, the more complex your model is, the more data you need. You can create these regression analyses for any number of predictor variables in a data set. You also, then have the ability to compare these models. You can manually compare these models but there is a better way.

There a several methods that can be used to compare different models in order to help identify "the best" model. More specifically, we can determine if a full model (which uses all available predictor variables and interactions) is necessary to appropriately describe the dependent variable, or whether we can throw away some of the terms (e.g. an interaction term) because they don't offer any useful predictive power.

One way of comparing the models is to use the Akaike Information Criterion (AIC). Here you would start with the full model, then you would define the next most simple model and remove a parameter. Once you have done this you obtain an AIC score for each model. If the difference between two AIC scores is greater than 2, then the model with the smallest AIC score is more supported than the model with the higher AIC score. If the difference between the two models' AIC scores is less than 2 then both models are equally well supported. In this situation, we use the AIC scores to decide whether our reduced model is at least as good as the full model. Here if the difference in AIC scores is less than 2, we can say that dropping the interaction term has left us with a model that is both simpler (fewer terms) and as least as good (AIC score) as the full model. Next, we see which of the remaining terms can be dropped.

This method of finding a minimal model by starting with a full model and removing variables is called backward stepwise elimination. Although regularly practised in data analysis, there is increasing criticism of this approach, with calls for it to be avoided entirely. Read into this further before utilising this method.

12. Covariates of No Interest - ANCOVA

Sometimes, there is no good way of mitigating a confound in your experimental design. All is not necessarily lost, however, because in some of those cases, you can include a confound as a covariate of no interest in your model.

There are two things that need to be true to be able to include a confound as a covariate of no interest:

1. It needs to be continuous.
2. The confound cannot interact with any of your predictors of interest.

So, how do you actually include a covariate of no interest in a model?

It's as simple as including the covariate of no interest in the model as if it were any other predictor (though, depending on the exact function you're using, you may need to ensure that you include it as the first term in the model). An ANCOVA is an example test which allows you to handle covariates of no interest.

12. Generalised Linear Models

The linear models we have talked about previously require the data to meet certain assumptions. For example, what if we have data where we can't describe the relationship between the predictor and response variables in a linear way?

One of the ways we can deal with this is by using a generalised linear model, also abbreviated as GLM. The GLM makes the linear model more flexible in two ways:

1. In a standard linear model the linear combination becomes the predicted outcome value. With a GLM a transformation is specified, which turns the linear combination into the predicted outcome value. This is called a link function.
2. A standard linear model assumes a continuous, normally distributed outcome, whereas with GLM the outcome can be both continuous or integer. Furthermore, the outcome does not have to be normally distributed. Indeed, the outcome can follow a different kind of distribution, such as binomial, Poisson, exponential etc.

When the residuals for a linear model do not fit a normal distribution, there are two options:

1. The first option, to transform our data (residuals), seems like a useful option and can work.
2. The second option would be to transform the linear predictor. This enables us to map a non-linear outcome (or response variable) to a linear model. This transformation is done using a link function.

Ultimately, the link function results in the model fitting the data to a pre-defined shape rather than a straight line. One famous example is logistic regression which fits a sigmoidal curve rather than a straight line. This allows the model to make binary predictions.

13. Logistic Regression

Logistic regression is a GLM which allows you to analyse data with a binary outcome. We can deal with binary outcome data is by performing a logistic regression. Instead of fitting a straight line to our data, and performing a regression on that, we fit a line that has an S shape. This avoids the model making predictions outside the [0, 1] range. Here, based on continuous input data, the model will output values between 0 and 1 which will represent the two binary outcome categories.

To clarify the input data that is used here is a continuous independent variable and a binary response or dependent variable which will be predicted by the model.

14. Mixed Effects Models

Mixed effects models are particularly useful in biological and clinical sciences, where we commonly have innate clusters or groups within our datasets. This is because mixed effects models contain random effects in addition to fixed effects (hence the name, “mixed”).

Rather than incorrectly assuming independence between observations, random effects allow us to take into account the natural clusters or structures within datasets, without requiring us to calculate separate coefficients for each group. In other words, this solves the problem of pseudoreplication, without sacrificing as much statistical power. Random effects are always categorical and we often want to control for this and may not be interested in it directly.

Mixed models are referred to as mixed models because they involve fixed effects AND random effects. The random effects come from partially pooling the data across the groups to add more randomness. As a rule of thumb, don’t fit something as a random effect unless there are at least 5 groups otherwise partial pooling isn’t really worth it.