Statistics for Research Students

An Open Access Resource with Self-Tests and Illustrative Examples

ERICH C. FEIN; JOHN GILMOUR; TANYA MACHIN; AND LIAM HENDRY

UNIVERSITY OF SOUTHERN QUEENSLAND
TOOWOOMBA

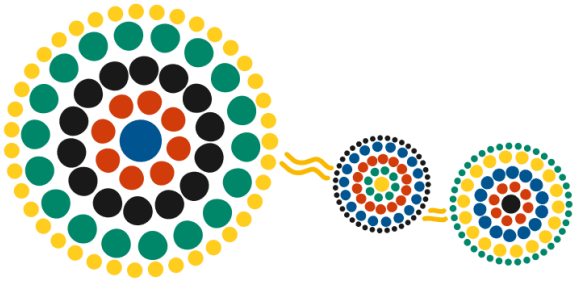
Contents

Acknowledgment of Country

The University of Southern Queensland acknowledges the traditional custodians of the lands and waterways where the University is located. Further, we acknowledge the cultural diversity of Aboriginal and Torres Strait Islander peoples and pay respect to Elders past, present, and future.

We celebrate the continuous living cultures of First Australians and acknowledge the important contributions Aboriginal and Torres Strait Islander people have and continue to make in Australian society.

The University respects and acknowledges our Aboriginal and Torres Strait Islander students, staff, Elders, and visitors who come from many nations.



Accessibility Information

We believe that education should be available to everyone, which means supporting the creation of free, open, and accessible educational resources. We are actively committed to increasing the accessibility and usability of the textbooks and resources we produce.

ACCESSIBILITY FEATURES OF THE WEB VERSION OF THIS RESOURCE

The web version of this resource has been designed with accessibility in mind and incorporates the following features:

- it has been optimised for people who use screen reading technology
  - all content can be navigated using a keyboard
  - links, headings, and tables are formatted to work with screen readers
  - images have alt tags
- Information is not conveyed by colour alone.

OTHER FILE FORMATS AVAILABLE

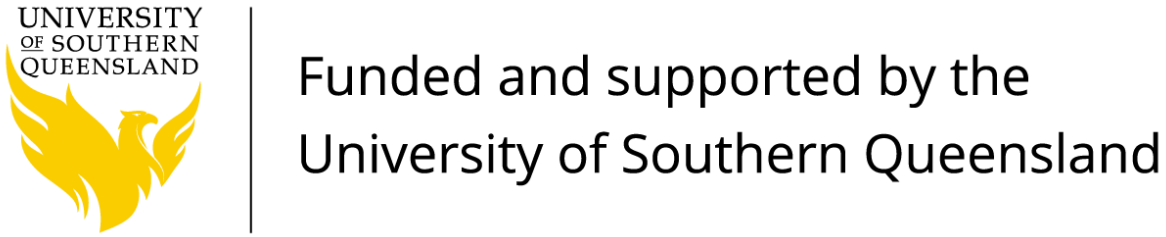
In addition to the web version, this book is available in a number of file formats, including PDF, EPUB (for ereaders), and various editable files. Look for the 'Download this book' drop-down menu on the landing page to select the file type you want.

ACCESSIBILITY IMPROVEMENTS

While we strive to ensure that this resource is as accessible and usable as possible, we might not always get it right. We are always looking for ways to make our resources more accessible. If you have problems accessing this resource, please contact us to let us know so we can fix the issue.

Acknowledgments

This book has been funded and supported by the University of Southern Queensland's open educational grants program. The author team also acknowledges and thanks our Project Officer employed during the duration of the open educational grant, Dr John Gilmour. John wrote the slide text which was use as the source for the book, and he performed all data manipulation and produced the output for the examples throughout the book. His work has been outstanding and essential in creating and compiling this book.



About the Authors

Dr Erich C. Fein is an Associate Professor at the University of Southern Queensland. He received substantial training in research methods and statistics during his PhD program at Ohio State University. He currently teaches four courses in research methods and statistics. His research involves leadership, occupational health, and motivation, as well as issues related to research methods such as the following article: "Safeguarding Access and Safeguarding Meaning as Strategies for Achieving Confidentiality." Click here to link to his Google Scholar profile.

Dr Tanya Machin is a Senior Lecturer and Associate Dean at the University of Southern Queensland. Her research focuses on social media and technology across the lifespan. Tanya has co-taught Honours research methods with Erich, and is also interested in ethics and qualitative research methods. Tanya has worked across many different sectors including primary schools, financial services, and mental health.

Dr Liam Hendry is a Lecturer at the University of Southern Queensland. His research interests focus on long-term and short-term memory, measurement of human memory, attention, learning & diverse aspects of cognitive psychology.

Dr John Gilmour is a Lecturer at the University of Southern Queensland and a Postdoctoral Research Fellow at the University of Queensland, His research focuses on the locational and temporal analyses of crime, and the evaluation of police training and procedures. John has worked across many different sectors including PTSD, social media, criminology, and medicine.

Introduction

Statistics for Research Students: An Open Access Resource with Self-Tests and Illustrative Examples

Why should you read this book?

When students embark on a statistics course, they often feel some amount of concern, ranging from mild anxiety to a sense of dread, prompting "statistics phobia" as part of the student experience. The way that teachers use discipline-specific language in the teaching of statistics – and how teachers can jump from topic to topic before students have the chance to master core ideas – may account for this fear of statistics. Indeed, an understanding and some degree of fluency in foundational terms – such as "central tendency" and "dispersion" – are essential before moving on to more advanced concepts that evolve from these ideas, like how to use standard errors or how to create and interpret confidence intervals.

The purpose of this Open Access Textbook is to provide a scaffolded approach to learning central, main ideas, which are required to use statistics as a research student at university. The basic idea governing this book is to introduce basic concepts first and then to allow students to have the chance to master core ideas through the use of examples and self-tests within the book. The book then progresses on to more advanced concepts as students progress through all parts and chapters. This overall aim reflects intrinsic motivational processes such as autonomy and mastery within Self Determination Theory. For example, we hope that the examples and self-tests within the book permit student autonomy and mastery as students take responsibility to pace themselves in their learning and manage their own development.

This book reflects the reality of statistics as both a method and discipline to guide and structure the collection and interpretation of numerical data, as well as its existence as a type of language or symbol system in itself. The use of statistics is increasing in most tertiary educational institutions, and in recent years the scope of statistics teaching in Australia, where this book was produced, even includes a mandate for significant statistical training in high schools. Accordingly, we hope that this free and accessible textbook will be a source of comfort and confidence for students of statistics in Australia. Whist the book was designed for research students at university, the actual audience may include students and teachers at the secondary level. Finally, it is our aspiration that this book would be used as a free and helpful educational resource for individuals and institutions throughout the world.

PART I

CHAPTER ONE - EXPLORING YOUR DATA

Hello everyone, and welcome to the first chapter of the University of Southern Queensland's online, Open Access Textbook:

Statistics for Research Students: An Open Access Resource with Self-Tests and Illustrative Examples

This book aims to help you understand and navigate statistical concepts and the main types of statistical analyses essential for research students. The aim of the first part of the book is to go over some of the basic concepts related to statistics and data analysis. These basic ideas include the core statistical concepts of *descriptive statistics*, *basic checks for normality*, and *graphing data*. As you progress through the book, you will see each chapter is broken into several sections.

There are some slides that appear via links within most chapters, and you should look for these as you review the current chapter.

Please proceed to the next section to get started.

Section 1.1: Data and Types of Statistical Variables

Learning Objectives

At the end of this section you should be able to answer the following questions:

- Using plain language, how would you define the concept of statistical data?
- What is a statistical variable?
- What are the main types of statistical variables?

So what are statistical data? The term "statistical data" refers to collections or sets of numerical information. The information is located within "cases" or records for separate individual entities, such as different people. However, cases can occur at multiple levels. Therefore, within data sets, you can have cases for people, groups, or larger entities like organisations or regions. For example, if researchers are analysing population health information in Queensland, Australia, they could have data sets representing cases of individual people, data sets representing cases of information for medical emergency teams, and data sets with cases for specific hospitals in which medical emergency teams are located. They could even arrange all this data for Queensland and compare it to other Australian states like Tasmania or Victoria.

Another important idea is the notion of a statistical variable. A statistical variable is a special type of mathematical variable.

As with all mathematical variables, statistical variables represent a conceptual space in a larger set of concepts. The conceptual space may be an abstract concept like a personality trait or it could be a physical concept such as height or weight. The fundamental properties of statistical variables are: 1) they hold the measurement of a particular value for an individual case, and 2) across all cases in a data set, a variable can possibly take on more than one value. If measurement for a "variable" is limited to only one value then it would not vary or change – and it would not be a variable. In this instance, you would have a constant rather than a variable.

Variables can be used to organise observations about many different concepts related to persons, objects, or groups. For example, variables can measure basic demographic information like gender, through to more complex and abstract information like attitudes or mental states. Basically, statistical data can represent a lot of different things.

There are a number of different types of variables.

Categorical variables are variables for which each possible value represents a different distinct category. For example, gender is categorical in most analyses, with people choosing male, female, or other. Another example of categorical data is type of driver's license. A person can be on a provisional license, an open license, or have no license. Australian state of residence is another categorical variable and could be recorded as Queensland, Tasmania, or Victoria. Categorical variables may be grouped into collections of categorical data.

In contrast to categorical variables there are also continuous variables. For continuous variables possible responses will fall on a spectrum. For example, age or height would be a continuous variable. Another type of continuous variable would be reaction time to stimuli.

In addition, although we are discussing categorical and continuous variables as separate types, there are responses to variables that are categorical, but are best thought of as continuous. The most obvious of these are responses to a Likert scale. For a Likert scale, a person can respond to a question like 'how much do you enjoy statistics?' The responses can range from 'greatly enjoy' to 'greatly dislike'. In this case, each response has its

own category, but in practice it is most common to mathematically manipulate the measurements of Likert scale responses as a continuous scale.

Section 1.2: Descriptive Statistics

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What is the concept of central tendency?
- What is the concept of dispersion?

So, let's say you had measured the height of everyone you know. All of those responses by themselves don't tell you much, beyond the height of each individual.

However, what we are after is a way to explain what the height of a typical person within this group might be. That is the concept of *central tendency*.

For example, let's say you wanted to work out what the average height of all your friends is. The most obvious way to do that is to look at the *mean*. The mean is simply all of the numbers (their height measurements) added together and then divided by the number of responses (or number of friends whose height measurements you have collected). In contrast, if the numbers were to be listed in numerical ascending order (from lowest number [shortest person] to highest number [tallest person]), the number in the middle would be the *median*. Another similar measure or statistic that measures central tendency is the *mode*. The mode is just the number that appears most often. Finally, the *range* is the difference between the lowest and highest values.

This information covers the main ways to look at the central numbers of a data set, but how can we tell how the data is distributed across the range? Are the responses all relatively close together, or are they spread widely apart? What we want, when we need to tell how the data is distributed across its range, is a way to explain the general dispersion or scattering of individual responses across the range. That is the concept of *variability* or *dispersion*.

The quickest way to determine how much the responses differ from each other is to look at the *standard deviation*. To work out the standard deviation for the height of all your friends you would first calculate the mean, then for each case or individual response you would subtract the mean and square the result (providing a squared difference from the mean), then you would calculate the mean of those squared differences then take the square root of that value.

If you did these calculations using paper and a pen, it could take a while if you have a lot of friends. Another way to complete the calculations is to use a statistical program to perform all these calculations for you!

Your standard deviation tells you how much the responses generally vary from the mean. A low standard deviation means that most of the numbers are close to the mean. A high standard deviation means that the numbers are more spread out. The *standard error*, which can be found by dividing the standard deviation by the square root of the total number of responses, tells you how accurate the mean of any given sample from that population is likely to be, when compared to the true population mean.

These types of descriptive statistics are the basic information you would provide for describing your data.

Section 1.3: Missing Data

Learning Objectives

At the end of this section you should be able to answer the following questions:

- How would you explain the difference between the concepts of MCAR, MAR, and missing not at random?
- How would you explain the purpose and interpretation of Little's MCAR test?
- What are the two main ways of dealing with missing data?

There are a number of things you will want to check before you commence any serious analyses with your data. The first thing is to check if you have any missing data. Missing data occurs when a response opportunity is missed by someone responding to your survey or questionnaire. Your participants might not respond to an item on your survey for different reasons. They may miss a question on your survey, or they may not want to answer a particular question, or they may get bored and stop filling the survey in! This lack of response to an item or items, creates a bit of a problem for the analyses. Missing data can be missing completely at random (MCAR), missing at random (MAR), or *missing not at random*.

Missing completely at random (MCAR) means the probability of a respondent missing data point is the same for all respondents. Someone might randomly miss a question on your survey, making that missing data point completely random. If the probability of a value being missing is the same only within groups defined by the observed data, then the data are missing at random (MAR). Because of this, MCAR and MAR are closely related concepts.

If the missing data is found not to be MCAR or MAR, it is missing not at random. For example, if a sizeable number of participants decide to skip one particular question on a survey, then that is not random. There is likely a reason for that question being skipped. It could be poorly worded, too personal, or just hidden at the bottom of the page.

There is a test to see if data is missing at random or not, which is called *Little's MCAR test*. Basically, if the test is not significant, any missing data is likely to have occurred at random. If the test is significant, there might be systematic or non-random reason the data is missing.

There are two main ways to deal with missing data.

First, there is a procedure of *mean replacement*. In this instance, you can replace the missing data points with a mean of that variable, though this technique is only recommended if the data is missing at random and is less than 5% of the variable in question.

There is also a technique called *multiple imputation*. This method is when the statistical program you are using, goes through the data and assigns a value for the missing variable of a particular case. This value is based upon previous responses to related variables and other non-missing responses for that variable in other cases. It is recommended that you use this if you have data missing at random of between 5-10% of the total responses of the variable.

Section 1.4: Checking Values

Learning Objectives

At the end of this section you should be able to answer the following questions:

- Is level of measurement for variables a consideration in checking for out of range responses?
- Is the scale of the variables a consideration in checking for out of range responses?

One of the first steps in understanding one's data is to check the values across the cases for indications the responses are what you would expect to see. These expectations should be based on the level of measurement and the scale of the variables.

Therefore, one of the most important data checks you would initially undertake is making sure all of the responses for each variable indicate a value within a range of appropriate values. This is known as checking for *out-of-range responses*.

For example, if you asked participants to write down their age in years, and you find a person who has responded with 512, there is a good chance this was an error made by the participant. Another example would be if you are asking people to rate how happy they are on a 1 to 5 scale, and you get a response of 8. This is another obvious error. These types of responses can occur when a person isn't paying attention to what they are typing or if there is a coding error in the survey software.

If you have found an obvious error in a person's demographic information, there isn't much you can do about it. You can remove that data point, or the whole case, depending on the circumstance. However, if there is an incorrect response in a 1 to 5 scale, you could use mean replacement to correct the inappropriate response.

Section 1.5: Normality

Learning Objectives

At the end of this section you should be able to answer the following questions:

- How would you define the concept of multivariate normality?
- What are three different methods for checking multivariate normality?

There are a number of underlying assumptions that go with parametric statistical testing (which is what we will be focusing on for the majority of this book).

If you are undertaking parametric tests, then one of the key assumptions is *multivariate normality*, or the assumption that the variables in your data are distributed normally.

Chances are you have all encountered an image of the bell curve throughout your academic studies. This bell curve represents a normal distribution.

There are a number of ways you can check for normality.

Your first option is to check multivariate normality by visually examining graphs of the data for each variable. For this type of checking, you will need to create a bar graph or a histogram. A basic visual inspection will often show if the data is normal or near to normal.

You can also check normality by looking at the *skewness* and *kurtosis* (S&K) of the distribution of your variables. Skewness looks at how the data is distributed horizontally. In other words, is the data all bunched up at one end of the graph. Kurtosis is the height of the distribution, and this should be neither too low nor too high. You will need to check if the curve is high and tight or flat and long. In an ideal distribution, the values assigned to S&K would be 0, however, there is nearly always some variance from normality in any dataset. S&K values of less than +/- 2.00 are generally considered to be close enough to normal for you to use parametric statistics.

Finally, a third way to check the normality of a data distribution is to use a dedicated *normality test*, which will be used one variable at a time. There are two main normality tests that researchers would typically use: the Kolmogorov-Smirnov test for samples larger than 50, and Shapiro-Wilk tests for samples less than 50. These tests assume that the distribution is normal. Therefore, if these tests are significant (i.e. p value <.05) it means that the data varies from the normal model and should be considered not normal.

If a continuous variable is found to be non-normal either from visual inspection, skewness and kurtosis values or a normality test, there are a number of ways to deal with this.

Firstly, you would want to check and see if there are any outliers in your data. If there are, it might be worth deleting them from the data set. You can also transform the variable into a logarithmic scale. Finally, if the violations are not too severe, you could just live with the non-normality of the variable and produce bootstrapped results. In any event, you will need to make mention of how you dealt with any non-normal data in the results section of your report or paper or thesis.

Section 1.6: Outliers

Learning Objectives

At the end of this section you should be able to answer the following questions:

- Explain the differences between a spurious and non-spurious outlier.
- Identify the level of z score for a response that typically indicates an outlier.

One of the major concerns when analysing data is the effect that *outliers* – which are unusually high or low data points – can have on the overall results. For example, if you were asking everyday people how many cups of coffee they consume a day, and most of the responses were between zero to four, that would be a normal spread of responses.

However, if you had one participant who responded that they consumed 17 cups of coffee a day, we would consider this response to be an outlier when compared to the rest of the participants. We could assume this participant has either a caffeine problem or they have incorrectly entered their response. Either way, this response will increase the mean for this particular sample, without being representative of the average coffee drinker.

There are a number of ways to statistically identify outliers in your data set. Participant responses to any variable can be transformed to a "z score," which is a basic transformation allowing you to compare responses across cases to a standardized response, which has a mean of 0 and standard deviation of 1. If a response has a z score of greater than +/- 3.3, it is to be considered to be an outlier. Another way is to graph the data using a box plot or bar graph, and visually identify the outliers. We will run through these options in greater detail later.

Normally, when you find outliers you can do two things: include them in the final analysis if you consider the outliers to be *non-spurious*, or you can remove them if the outliers have occurred for *spurious* reasons, meaning that they do not reflect accurate responses. If the outliers don't make sense in the context of the question, or are extreme without any potential justification, is a good idea to consider these as spurious responses and just remove them from the analysis. However, if you do find some responses that make sense or are only slightly outside the acceptable z score (+/-3.3), it may be worth considering them to be non-spurious outliers and keeping them for analysis.

Section 1.7: Chapter One Self-Test

It is a good idea to test your knowledge after you have completed each chapter. This section presents a Self-Test for Chapter One. Please complete the test now to assess your learning of the ideas inside this chapter.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/statisticsforresearchstudents/?p=473#h5p-6

PART II

CHAPTER TWO - TEST STATISTICS, P VALUES, CONFIDENCE INTERVALS AND EFFECT SIZES

Hello everyone, and welcome to the second chapter of the University of Southern Queensland's online, open access textbook.

The aim of this second chapter is to discuss three statistical tools that allow us to quickly determine if a given estimate of effect (or "effect size") is statistically significant and if it may be of practical use: These tools are p Values and Confidence Intervals as well as the Effect Size itself. A p value is a statement of probability about how often a real association between variables would occur, and it is at the heart of the family of statistics based on frequencies of a particular event or outcome (in this case a real association between variables). Confidence intervals are another way of showing how useful an effect size estimate is, based on how much error has been added into making the estimate. Confidence intervals are important because they provide both an indication of statistical significance (that is, a real association between variables) for an effect as well as an indication of error in its estimation. Effect size estimates are important because they provide an indication of the magnitude of an effect, or how strong an association is between variables.

There are some slides that appear via links within Chapter Two. Please look for these as you work your way through the current chapter.

Section 2.1: p Values

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What is a p value?
- How can you interpret a p value?
- What question can p value answer?

An important area of statistics is probability, and it is the basis for all of the tests we will be reviewing in this textbook. One important kind of probability is a conditional probability. For example, given the weather forecast for today, what is the likelihood that it will rain?

The p value itself is a figure or numeral – typically represented by a number between 0 and 1.00 – that provides the probability of a result (for a particular test statistic) being due to a true effect rather than chance. The p value is a conditional probability and relies on a number of assumptions about the test statistics used.

Here is an example from psychology that provides an illustration of the p value:

Psychological scientists at your university are evaluating a clinical therapy that is believed to reduce anxiety in young adults. In a field study, these scientists use two groups to test the therapy – one group receives the clinical therapy and a second group that does not receive the clinical therapy – which are respectively known as the experimental and control groups. Anxiety in participants is then measured in both groups after the therapy takes place (or not). Using a T-test statistic – which examines the different means for anxiety between two groups – the result of the test statistic is $t(18)= 2.7$, $p = .01$.

The p value is indicated by the statement of $p = .01$ that appears after the 2.7, which is the value of the T-test statistic. You interpret the significance of a p value based on a critical value for p values, which is often designated as .05. You also note that p values less than .05 are considered significant in most research. In our example $p = .01$ which is below .05. This means that the test statistic of $t(18)= 2.7$ provides evidence for a difference between the control and experimental groups.

When concluding there is a difference between the control and experimental groups, a researcher is really referring back to the populations from which the two groups are assumed to be drawn. Hence, there is an inference from the samples back to the populations.

It is critical to remember that a p value does NOT answer "What is the probability that the difference is due to chance?" A p value does answer: 'Assuming that there is no real difference in the populations (that correspond to the two groups), what is the probability that the difference between the means of randomly selected subjects will be as large as or larger than actually observed?' This distinction might sound academic, but it is very important.

Section 2.2: Significance

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What is the main idea underpinning statistical significance?
- Can we interpret a non-significant result as "no difference between means" or "no relationship between variables?"

Understanding statistical significance is important, and sometimes using analogies from other disciplines can help you better understand these ideas.

To illustrate statistical significance using non-statistical jargon, think of a courtroom where people are tried for alleged crimes. In courtrooms using legal systems based on English Common Law, there are generally only two outcomes permitted from the jury after evidence is presented. These outcomes are "GUILTY" or "NOT GUILTY." This outcome is really just a decision, but that decision also must correspond with the true state of reality to be correct.

The decision the jurors have to make are based on the following logic. If the evidence is inconsistent with the assumption of innocence the verdict "GUILTY" can be announced. In contrast, if the evidence is not inconsistent with the assumption of innocence the verdict of "NOT GUILTY" is announced, but the term "INNOCENT" is never used. It is impossible to truly prove innocence.

Now keep this analogy in your mind, while we go back to understanding statistical significance testing.

To assume that the statistical distribution is the same in the two populations of interest, means the null hypothesis is true, which is analogous to assuming that the defendant is innocent in law.

If the evidence from the data of the study is inconsistent with the null hypothesis we can "fail to accept the null hypothesis" and state that the difference is "statistically significant" and conclude from a significant result that the null hypothesis is highly unlikely. In contrast, a non-significant result only tells us that the effect is not big enough to be anything other than a chance finding. It does not tell us that the effect is zero.

Therefore, we never interpret a non-significant result as "no difference between means" or "no relationship between variables." It could mean that the tests run were just unable to detect the association or difference. Accordingly, non-significant results shouldn't be interpreted as NO EFFECT. Non-significant results could be due to many things including a small effect or low statistical power of the experiment.

Additionally, significant results are often interpreted (or over-interpreted) as important results and may be equated or confused with large effect.

Section 2.3: Confidence Intervals

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What is a Confidence Interval?
- How is standard error related to a Confidence Interval?

In many statistical papers, you will see Confidence Intervals (or CIs) reported, usually at the 95% threshold. But what does that actually mean?

The first step in understanding a CI is that there is a "point estimate" or statistic from a sample. The usual statistic first encountered in research is a sample mean, although many other statistics, such as effect size estimates, can be used for point estimates within CIs.

What is good about a CI is that the range of the CI – from the lowest value to the highest value – indicates how much error is used in the measurement of the "point estimate" or statistic from a sample.

Keep in mind that the standard error is a basic measure of variability that accounts for how much error exists when relating the sample results to the true values of a parameter in a population. In our example below, the standard error comes from the standard deviation divided by the square root of the sample size. Therefore, a larger sample of data will result in a smaller amount of standard error in estimating the population value of a parameter. This makes sense because if you have more cases in a sample from a population, you are more likely to estimate the true values of a parameter in a population.

For example, let's say we have a sample of data from 42 fathers, where they provide ratings of their confidence in fathering as well as the time spent with their children. Here are the "point estimates" with standard errors and CIs.

Confidence in Fathering where the mean = 32.50, standard error = .70, CI = [31.20, 33.90]

Time Spent With Children where the mean = 8.00, standard error = .20, CI = [7.70, 8.30]

Notice that the range in the first CI is 2.7, which comes from the distance between the upper and lower limits of 31.20 and 33.90.

In comparison, the range in the second CI is 0.6, which comes from the distance between the upper and lower limits of 7.70 and 8.30.

The basic idea here is that when there is more standard error, then the range of the confidence intervals is bigger, which means that our point estimate is less accurate.

CIs are based on a proportion of confidence in the sample point estimate, that is, related to a set probability of the estimate occurring. This is because, from one sample alone, there is no way to know the value of the true population mean. Therefore, we need to estimate how likely it is that the true population value of the mean lies within the calculated CI.

As a general rule researchers like to state the confidence level as 95%. However, CIs can use other levels of significance such as .01 and .10 as a basis for constructing the interval. Accordingly, if 95% CIs were to be calculated from many repeated samples from the population, the population mean would fall within the limits of the CI in 95% of the samples. However, 5% of the CIs will not capture the population mean. Basically, if you were to sample the population 100 times, 95 times the sample mean would fall within the limits of the 95% CI.

CIs can be also calculated for effect size measures such as the T-test or correlation coefficient. An interesting fact about effect size point estimates based on the normal distribution is that in cases where the interval or the distance

between the upper and lower limits of the CI includes 0, the effect measured by the point estimate is statistically insignificant.

To properly construct a CI, a sample should be randomly selected from the population, or the researchers must assume that a convenience sample adequately represents the population, with the results similar to what would have been observed had a true random sample be used. In addition, it must be assumed that the population has a normal distribution for the variable of interest or at least an approximately normal distribution.

Section 2.4: Effect Sizes

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What is an Effect Size?
- How is an Effect Size related to a p Value?
- How do you determine the magnitude of an Effect Size?
- What is the difference between statistical significance and practical significance?

Effect size is a term used to describe the strength or magnitude of an effect. This effect is usually expressed as a measure of difference or association. Like most statistical tests, effect sizes come in two distinct groups, and effect sizes generally range from 0 to 1.0. The first type of effect size is based on magnitude of difference between groups, and this is known as the d family of effect sizes. The second type of effect size is the measure of association or the variance accounted for by two or more variables, which is known as the r family of effect sizes.

For example, a t-test produces the effect size d, while a correlation coefficient produces the effect size r.

Generally, the effect size values such as d or r are only transformations of the difference between groups or associations between variables, which are then weighted (divided by) the size of the sample and/or its standard error.

Therefore, the higher the difference or association, and the greater the sample, the bigger the effect size.

Cohen (1988) suggests small, medium, and large effects sizes for a T-test would respectively be about .2, .5, and .8, while small, medium, and large effects sizes for the correlation coefficient would respectively be about .1, .3, and .5.

When reporting results from the calculation of a test statistic, it is always a good idea to report more than just the p value. It is far better, and more thorough, to include the effect size and the CI along with the p value result of the test.

As you can no doubt see from the information above, there are a number of effect sizes and they are associated with different kinds of statistical tests.

However, if you find a result that is statistically significant, but has a very small effect size, you must ask yourself if the use of variables and interventions producing those effects would be practical.

It is important to note that just because a researcher finds a statistically significant result, it does not mean the result is sizeable, important, or useful in the real world. Statistical significance is not the only measure of a result. A result should also be practically significant, which means the strength or size of the effect represents a finding that is practically important to others.

Practical significance always involves judgment by other researchers or consumers of research that takes into account factors such as cost and political considerations of interventions tied to the estimated effects.

Section 2.5: Statistical Power

Learning Objectives

At the end of this chapter you should be able to answer the following:

- Explain the idea of Statistical Power.
- Define Type I and Type ll Decision Errors.

As course examiners, we know that students always want to get the 'right' answer. However, it is possible to make errors in the interpretation of statistical effects. When researchers misinterpret the significance of results, these mistakes are often more to do with errors in judgment rather than about mistakes in effect calculations.

You may recall that the essence of significance lies in an inference between a sample and the population to which it corresponds.

Therefore, there are two factors that must be considered.

Firstly, we need to consider the conclusion we reach about a sample result showing a significant or insignificant effect.

Secondly, we need to consider whether there is a true state of reality in the population from which the sample corresponds.

It is important to understand that the conclusion of the sample result and the true state of reality in the population must be aligned.

For example, if we conclude there is no difference between groups in a sample – and at the same time conclude there is no difference between groups in the true state of reality in the population – then our judgment about the sample result and the difference between the true state of the population are aligned.

Misalignment of the sample result and the true state of reality in the population can be thought of either as Type I and Type ll Decision Errors.

A Type l (α) Error is known as a false positive or thinking something is there when it is not. This is where, based on the sample, we say there is a true effect, but in the population, there is in fact no effect. For example, a Type I error occurs when a healthy person is diagnosed with a particular disease like cancer.

A Type ll (β) Error is known as a false negative or thinking something is not there when it is. This is where, based on the sample, we say there is a not a true effect, but in the population, there is in fact an actual true effect. For example, a Type II error occurs when a sick person is diagnosed as being disease-free.

The statistical power of the statistical test is related to the Type ll (β) Error as the probability of rejecting the null hypothesis when it is true.

Power = 1 – β

Statistical power is, therefore, the probability of making a correct decision, or saying there is an effect based on the sample, and at the same time there is in fact an actual true effect in the population.

The smaller the probability of a type II error or a false negative, the bigger the power. However, with large amounts of data, it is possible to have samples that are overpowered and find statistically significant effects that are not practically significant.

There are tests that can be used to check if the sample size you have is large enough to detect a relationship or a difference. These are called apriori and posthoc power analyses. An apriori power analysis allows researchers to understand how large a sample should be so that it has adequate sensitivity to detect true effects. Posthoc power analyses allow researchers to determine if the sample size was inadequate for calculating a statistically significant finding.

Section 2.6: Chapter Two Self-Test

It is a good idea to test your knowledge after you have completed each chapter. This section presents a Self-Test for Chapter Two. Please complete the test now to assess your learning of the ideas inside this chapter.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/statisticsforresearchstudents/?p=476#h5p-4

PART III

CHAPTER THREE - COMPARING TWO GROUP MEANS

Hello everyone, and welcome to Chapter Three of the University of Southern Queensland's online, open access textbook on statistics for research students.

The aim of this chapter is to discuss methods for comparing differences in means across two groups. In other words, would the mean value of some variable in one group be different to the mean value of that same variable in another group. This is a very useful process when two groups differ on some characteristic of importance (such as a psychological treatment or intervention). Hence we can sometimes refer to these groups as "non-equivalent groups."

For more information, please keep reading this chapter.

Section 3.1: Looking at Group Differences

Learning Objectives

At the end of this chapter you should be able to answer the following questions:

- What is the difference between a Treatment Group and a Control Group?
- What types of naturally occurring groups differ on social behaviours?

Examining and understanding how groups of individuals can differ is one of the key goals of many social sciences, including psychology. This understanding is particularly important when you want to examine whether a certain intervention is helpful. For instance, if you wanted to trial a new depression medication or treatment. Often times these interventions are used when conducting a particular type of research design, more often an experiment where one group is assigned to the treatment or intervention, and the other group is not. Let's now unpack this idea a little further.

One group will be known as the *Treatment Group*. A Treatment Group comprises of the group of participants that receives some type of treatment or intervention that is expected to make a difference in one or more outcomes. For example, if a psychologist is using a new method of therapy within a group of clients, that therapy could be considered a treatment or intervention. A different type of group in experimental settings, in a group of participants that receives *no* amount of treatment or intervention, which is known as the *Control Group*.

Most of you would be familiar with treatments administered by researchers in settings such as clinical practice or medical trials. However, it is possible for groups to experience some effect as a "treatment" under natural conditions.

In respect to naturally occurring groups, there are many different groups of people that make up groups and organisations, like classrooms or work groups. Each group may have a distinct environment, including aspects of social differences. It makes sense that many groups would differ on different psychological constructs and behaviour. An example of this would be risk-taking behaviour. Would you expect a group close to retirement to have the same thoughts, feelings and behaviours around taking risks as a group of new employees early in their careers?

There are some slides that appear via links within Chapter Three. Please look for these as you review the current chapter.

Section 3.2: Between Versus Within Groups Analysis

Learning Objectives

At the end of this chapter you should be able to answer the following questions:

- What is the difference between a *Between Groups* test and a *Within Groups* test?
- What is a *Mean Difference* in the context of statistical analysis?

There are two main types of statistical tests: those that look at differences *Between Groups* and those that look at differences *Within Groups*.

Between Groups differences examine how independent groups – groups that are not the same – may differ from each other on a variable. Between Groups difference tests are useful for examining the efficacy of interventions or treatments. For example, if you wanted to see if a new form of anxiety therapy was effective, you could organise two groups of participants, and provide one with the new form of anxiety therapy. This group would be the intervention group. To use a *Between Groups* test you would also need a comparison group that does not receive the treatment, which would be your control group. Both groups would need to receive some form of outcome measure – such as a measure of anxiety taken after the treatment. You would then compare the two and see if there were any differences in mental states.

Within Groups differences are similarly important. For example, if a researcher wants to examine if an exercise program is effective, she could take the BMI of a group of test subjects at the start of the program and again at the end of the program and compare the two. In this case, the researcher is not looking at the differences between two groups, but rather the differences between the same group taken at two time points.

In both differences between groups and differences within groups, we will generally look at differences between means on some variable of interest. When we talk about a *Mean Difference*, we are talking about the difference between the mean of one group and the mean of another group in the case of differences between groups. In the case of differences within groups, we look at differences in means between two or more different points in time when measurements are taken. When looking at the BMI of the trial group we just mentioned, you would want the mean score of the group at pre-program, and the mean score at post-program.

To examine the differences between or within groups, you also need to know the standard deviations of both means you are comparing, as well as the number of participants. With the mean, the measure of variance within the samples is the standard deviation. Once you have the mean difference, the standard deviation, and the number of data points, you can then use the T-test to calculate if the difference between the two means is statistically significant.

There are two main types of t-tests we will be focusing on – the independent samples t-test and the paired samples t-test. When you are examining the difference between two groups, you want to use the independent samples t-test, however, if you are looking into the difference between the same group at two different time points, the paired-sample t-test is the one to use.

Section 3.3: Independent T-test Assumptions, Interpretation, and Write Up

Learning Objectives

At the end of this chapter you should be able to answer the following questions:

- Is the Independent T-test a Between Groups or Within Groups test?
- How many assumptions underpin the Independent Samples T-test?
- What is the first test to examine within the Independent Groups T-test output?
- What is the second test to examine within the Independent Groups T-test output?
- What elements or individual statistics should be reported when writing up an Independent T-test?

An *Independent T-test* or Independent Samples T-test is an important test for Between Groups differences.

Here we will discuss the underlying assumptions of the Independent t-test and explain how to interpret the results of the t-test. There are a number of assumptions that need to be met before performing an Independent t-test:

1. The dependent variable (the variable of interest) needs a continuous scale (i.e., the data needs to be at either an interval or ratio measurement). An example of a continuous dependent variable might be the weight of an athlete. Their weight could be anywhere between 50 and 70 kilograms.
2. The independent variable needs to have two independent groups with two levels. An example of this independent variable could be regional vs metropolitan Australians.
3. The data should have independence of observations. More specifically, there shouldn't be the same participants in both groups.
4. The dependent variable should be normally or near-to-normally distributed for each group. It is worth noting that the t-test is robust for minor violations in normality, however, if your data is very non-normal, it might be worth using a non-parametric test or bootstrapping (see later chapters for more information).
5. There should be no spurious outliers.
6. The data must have homogeneity of variances. This assumption can be tested using Levene's test for homogeneity of variances in the statistics package which is shown in the output included in the next chapter.

Independent T-test Interpretation

The order of interpreting test statistics can be important and there are multiple test statistics to interpret within the Independent Groups T-test output.

Keep in mind that we are examining two groups of individuals – In this example, we are looking at metropolitan versus regional Australians. The dependent or outcome variable is mental distress.

And here we have the output from the T-test.

PowerPoint: Independent T-test Output

You will need to click on the below link to access the output:

- Chapter Three Independent T-test Output

Group Statistics

	Region	N	Mean	Std. Deviation	S.E. Mean
MentalDistress	Metropolitian	209	35.90	12.10	.84
	Regional	158	38.87	12.69	1.01

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
MentalDistress	Equal variances assumed	2.51	.114	-2.27	365.00	.024	-2.96	1.30	-5.52	-.40
	Equal variances not assumed			-2.26	329.59	.025	-2.96	1.31	-5.54	-.38

Green: Levene's test

Red: Test statistics

Blue: Means and standard deviations

Green: The first thing you should examine is Levene's test. If this test is nonsignificant, that means you have homogeneity of variance between the two groups on the dependent or outcome variable. If Levene's test is significant, this means that the two groups did not show homogeneity of variance on the dependent or outcome variable. In our example, Levene's test is nonsignificant so we can move on to the statistics for the tests under the condition of equal variances assumed.

You should notice that there are two lines or rows of statistics given in the output. The first row, which we are using, provides statistics for the tests under the condition of equal variances assumed. The second row, which we are not using, provides statistics for the tests under the condition of equal variances not assumed.

Red: The next thing you should look at is the t value, the degrees of freedom, and the p value statistics in the first or top row of the output. The p-value of .024 shows that there is a significant difference in levels of mental distress reported by metropolitan and regional Australians. If we look at the mean scores, we can tell that regional Australians reported higher levels of mental distress (38.867) than the Australians who live in major cities (35.904).

You will also notice that there is a 95% CI presented, which is a 95% Confidence Interval of the difference. This CI has a lower limit at -5.525 and an upper limit at -.401. Because the CI does not include 0 we can infer that the difference between the two groups does exist in the population.

Blue: Next, make sure you have a look at the mean, standard deviation, and sample size (N) for both groups. You can get the effect size (Cohen's D) by using an effect size calculator.

You may find an effect size calculator here: https://www.socscistatistics.com/effectsize/default3.aspx

If you enter the mean, standard deviation, and sample size for both groups, you should get a Cohen's D of .239.

Independent T-test Write-Up

You will need to report the Means and SD for each group, along with the t test statistic (*t*), its *p* value, and its effect size *d*.

It is common in many formats to round your decimal places to two. Therefore, a Write-Up for an Independent T-test should look like this:

An independent samples t-test showed that the metropolitan sample (M = 35.90, SD = 12.10) reported lower levels of mental distress (*t*=-2.27, *p*=.024, *d*=.24) than the regional sample (M = 38.87, SD = 12.69).

Section 3.4: Paired T-test Assumptions, Interpretation, and Write Up

Learning Objectives

At the end of this chapter you should be able to answer the following questions:

- How does the Paired t-test differ from the Independent t-test?
- What assumptions that need to be met before performing a Paired-samples t-test?
- What elements or individual statistics should be reported when writing up a Paired T-test?

The Paired t-test is also known as the Paired Samples t-test or the Dependent t-test.

Paired T-test Assumptions

Just like in the Independent t-test from our previous chapter, there are a number of assumptions that need to be met before performing a Paired-samples t-test:

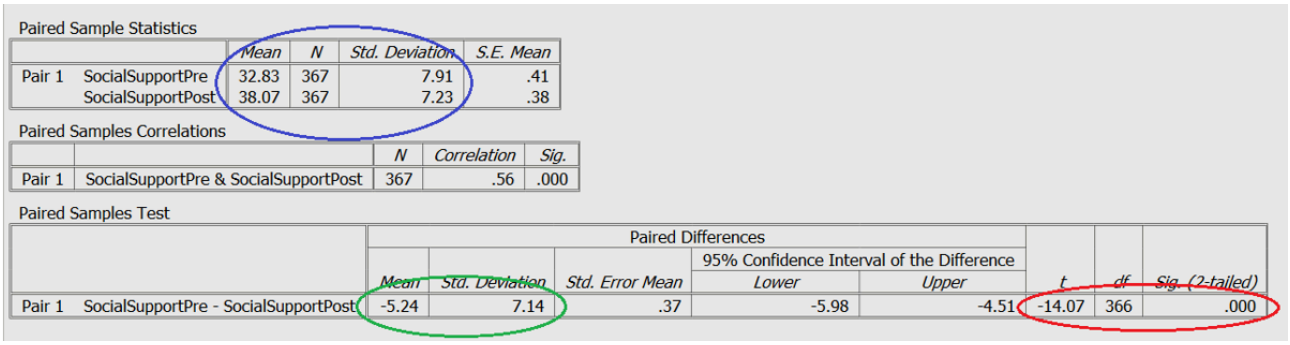
1. The dependent variable (the variable of interest) needs a continuous scale (i.e., the data needs to be at either an interval or ratio measurement). An example of this variable could be weight or level of anxiety.
2. The independent variable must have two groups that are related or have "matched pairs". Matched pairs, as mentioned above, mean that the same participants are present in both groups. More specifically, we are measuring the same persons twice.
3. There should be no spurious outliers across either of the groups being used in the test.
4. There should be an approximate normal distribution of differences across the two related groups.

Paired T-test Interpretation

A Paired t-tests examines the within-group differences of a single group. So the same subjects are the respondents for both pairs of measurements, and this indicates that you have related groups of scores.

This means that the "related groups" or "matched pairs" are the same participants who have been measured in each of the two groups. Specifically this means the two groups both contain the same people, who are producing the same scores or measurements in both groups. Unlike the independent samples t-test, researchers can measure the same participants in each level of the independent variable because the measurement or scoring for each participant will be taken at two distinct time points.

In the worked example for this test, we will be looking at the levels of perceived social support a group of Australians reported before engaging with a social skills building program and after completing the program.



Paired Sample Statistics

		Mean	N	Std. Deviation	S.E. Mean
Pair 1	SocialSupportPre	32.83	367	7.91	.41
	SocialSupportPost	38.07	367	7.23	.38

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	SocialSupportPre & SocialSupportPost	367	.56	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	SocialSupportPre - SocialSupportPost	-5.24	7.14	.37	-5.98	-4.51	-14.07	366	.000

Red: Test statistics

Green: Differences in means for effect size

Blue: Means and standard deviation

This example shows perceived social support as an outcome variable. This outcome is expected to change after the intervention of a social skills building program.

The output shows levels of perceived social support across two groups of scores: Scores from a group of Australians reported before engaging with a social skills building program, and scores the same people reported after completing the program.

As you can see, the t-test was significant as circled in red, showing a change from the scores before completing the social skills program, and after. By looking at the means as circled in blue you can tell that the mean level of perceived social support was higher following the completion of the program. To calculate the effect size, you divide the mean difference by the standard deviation of the difference as circled in green.

Paired T-test Write Up

When you are writing your report, you will need to include the Means and SD for each group, along with the t test statistic (*t*), its *p* value, and its effect size *d*.

A Write-Up for a Paired T-test should look like this:

A paired samples t-test showed that the participant's level of perceived social support increased from pre-program (M = 32.83, SD = 7.91) to post-program (M = 38.07, SD = 7.23; $t = -14.07$, $p < .001$, $d = -.73$).

Section 3.5: Chapter Three Self-Test

It is a good idea to test your knowledge after you have completed each chapter. This section presents a Self-Test for Chapter Three. Please complete the test now to assess your learning of the ideas inside this chapter.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/statisticsforresearchstudents/?p=544#h5p-7

PART IV

CHAPTER FOUR - COMPARING ASSOCIATIONS BETWEEN TWO VARIABLES

Hello everyone, and welcome to the fourth chapter of the University of Southern Queensland's online, open access textbook on statistics for research students. The aim of this chapter is to discuss associations between variables. As we have seen, there are various formulas that allow us to determine if variables are statistically associated, and such formulas produce a *test statistic*.

Such test statistics are produced by showing how the variables of interest change together across different cases of paired responses. For instance, say we measured reading ability and test performance for 100 students. According to this design, each of the 100 students would produce paired variables for reading ability and test performance. A test statistic will reflect to what extent reading ability and test performance change together for each student, and the test statistic reflects this change across all 100 student cases.

Therefore, if the two variables – reading ability and test performance change together or covary, we would have a significant test statistic, which would reflect the aggregated change across the 100 pairs of responses.

These ideas will be explored in more depth within this chapter.

There are some slides that appear via links within Chapter Four. Please look for these as you review the current chapter.

Section 4.1: Examining Relationships

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What is a statistical relationship?
- What is the difference between a positive and negative relationship?
- What does a Pearson correlation coefficient indicate?

When we discuss relationships or associations between variables, the terms "relationship" and "association" mean that two variables change, or vary, together. However, just because two variables change together does not mean that this change is statistically significant. Thus, there are two types of variation or relationships – significant and nonsignificant.

As a researcher, looking into the relationships between psychological constructs can tell me a great deal about how certain mental health concerns and behaviours effects individuals with respect to behavioural and emotional levels. For example, does a person with greater levels of social connection and support have greater feelings of well-being? Do people with greater levels of mindfulness have lower levels of perceived stress? These are questions that can be answered using correlational analyses.

So what is a statistical relationship? A statistical relationship is the association between two variables that is statistically significant. This significance is based on the level of a probability test, which is a *p-value* in the case of Pearson correlation coefficients. If one variable increases or decreases, an associated variable will also show an increase or decrease, and it is statistically significant if this variability can be attributed to more than chance. An example is calorie intake and weight, as more calories are consumed, weight will likely increase. This type of example shows a positive relationship between the variables which means that as one variable increases the other also increases (e.g., as height increases, weight usually increases). However, relationships can be either positive or negative. A negative relationship is present when one variable increases, the other decreases (e.g., as stress levels increase, health will likely decrease).

A Pearson correlation coefficient (represented as an r value statistically) is a very useful tool in psychological research. However, there are many other types of correlation coefficients, such as the Spearman rank-order correlation coefficient, which is a nonparametric measure of association between two variables.

A Pearson correlation draws on a "line of best fit" that will be imposed through the two variables in the data to establish the relationship between two variables. Using the linear model, the Pearson's correlation coefficient (which is represented by an *r*), represents the strength of the association. This means that the distance from the data points show the line of best fit and how strongly the two variables are related. Mathematically, the Pearson correlation is calculated from the central tendency statistic of the mean and the standard deviation for each of the variables. Have a look at the below illustration by clicking on the link labelled "Chapter Four – Line of Best Fit," which displays a graph with the Line of Best Fit for the two variables mental distress and physical illness. The variables are in fact correlated with a significant Pearson correlation coefficient ($r = .472$, $p < .000$).

PowerPoint: Line of Best Fit

Take a look at the following PowerPoint slides:

- Chapter Four – Line of Best Fit

The Pearson correlation coefficient can range from +1 to -1, with positive values indicating that as one value increases (e.g., as height increases, weight increases) the other also increases, or negative values which show that as one value increases, the other decreases (e.g., as stress increases health decreases). The stronger the association between the two variables, the closer to 1 the value will be.

	Coefficient, r	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

Section 4.2: Correlation Assumptions, Interpretation, and Write Up

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What assumptions should be checked before performing a Pearson correlation test?
- What is the relationship between correlation and causation?

Correlation Assumptions

There are four assumptions to check before performing a Pearson correlation test.

1. The two variables (the variables of interest) need to be using a continuous scale.
2. The two variables of interest should have a linear relationship, which you can check with a scatterplot.
3. There should be no spurious outliers.
4. The variables should be normally or near-to-normally distributed.

Correlation Interpretation

For this example will be looking at two relationships: levels of physical illness and mental distress, and physical illness and life satisfaction.

PowerPoint: Correlation Output

Have a look at the following slides while you are reviewing this chapter:

- Chapter Four Correlation Output

Correlations

		PhysicalIllness	MentalDistress	LifeSatisfaction
PhysicalIllness	Pearson Correlation	1.00	.47	-.34
	Sig. (2-tailed)		.000	.000
	N	367	367	367
MentalDistress	Pearson Correlation	.47	1.00	-.49
	Sig. (2-tailed)	.000		.000
	N	367	367	367
LifeSatisfaction	Pearson Correlation	-.34	-.49	1.00
	Sig. (2-tailed)	.000	.000	
	N	367	367	367

As you can see from the output (**circled in green**), the relationship between physical illness and mental distress is both a positive and statistically significant. As a person reports higher levels of physical illness, they are more likely to report higher levels of mental distress. The reverse is true of the relationship between physical illness and life satisfaction, with these showing a significant negative relationship (**shown in red**). The greater levels of physical illness, the more likely the person is to have lower levels of life satisfaction. Both relationships are significant (see p values) and are medium in strength.

Although we have two significant relationships or associations between these variables, it does not mean that one variable causes the other. Researchers need to remember that there may be third variable (or covariates) present that affects both variables accounted for in a correlation coefficient. Therefore, the existence of a significant correlation coefficient for a pair of variables by itself does not imply causation between the two variables.

Correlation Write Up

A write-up for a Correlation Analyses should look like this:

Among Australian Facebook users, the levels of reported physical illness and mental distress showed a medium positive relationship, $r(366) = .47$, $p < .001$, whereas the levels of physical illness and life satisfaction showed a medium negative relationship, $r(366) = -.34$, $p < .001$.

Section 4.3: Chapter Four Self-Test

It is a good idea to test your knowledge after you have completed each chapter. This section presents a Self-Test for Chapter Four. Please complete the test now to assess your learning of the ideas inside this chapter.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/statisticsforresearchstudents/?p=481#h5p-5

PART V

CHAPTER FIVE - COMPARING ASSOCIATIONS BETWEEN MULTIPLE VARIABLES

Hello everyone, and welcome to the fifth chapter of the University of Southern Queensland's online, open access textbook on statistics for research students. The aim of this chapter is to discuss associations between three or more variables.

As we have discussed in previous chapters, associations are mathematical relationships between variables. Generally, relationships are framed around a key pair of variables that explain a central effect of interest. For example, we have discussed that reading ability and test performance may be associated. However, there are other variables that could be associated with *either* reading ability, or test performance, or there may be other variables that are associated with *both* reading ability and test performance. Such associated variables are often termed covariates. Covariates are often used as additional variables of interest in Multiple Regression and Hierarchical Regression.

There are three main methods of regression analysis that correspond to three different types of models:

- **Simple or Basic Regression**: a regression model with one independent variable and one dependent variable.
- **Multiple Regression**: a regression model with two or more independent variables and one dependent variable.
- **Hierarchical Regression**: a regression model with two or more independent variables entered within two or more blocks of sequential predictors, and one dependent variable.

There are some slides that appear via links within Chapter Five. Please look for these as you review the current chapter.

Section 5.1: The Linear Model

Learning Objectives

At the end of this section you should be able to answer the following questions:

- Explain the difference between Correlation and Regression Analyses.
- Explain why the linear model is important for Regression.

As we discussed in the chapter on correlations, correlational analyses focus on the relationship between two variables. What we will now be exploring is the association of multiple variables with a single dependent variable via a series of models known as Regression Models. In linear regression, the logic of the linear "line of best fit" that we discussed in correlation analyses is also used.

When we are looking at any linear regression model, we are producing and examining the straight line of best fit for predicting the relationship between two variables. This is the same standardised line that we examine when using correlation.

However, rather than simply plotting a line of best fit for cases of two variables, there is can be a predictive model based on more than one variable that is associated with a single dependent variable. Have a look at the link below, which presents the graph of a linear relationship from Chapter Four. The mathematical formula that produces this type of relationship, which is characterised by a straight line that is always increasing, is a linear model.

PowerPoint: Line of Best Fit

Have a look at the following slides while you are reviewing this chapter:

- Chapter Four – Line of Best Fit

Section 5.2: Simple Regression Assumptions, Interpretation, and Write Up

Learning Objectives

At the end of this section you should be able to answer the following questions:

- Explain the Assumptions for Simple Regression.
- Explain what R Squared means.

Like in our previous chapters, it is important to understand that simple regression also has assumptions. In this case, Simple Regression Assumptions include:

1. The two variables (the variables of interest) need to be using a continuous scale.
2. The two variables of interest should have a linear relationship, which you can check with a scatterplot.
3. There should be no spurious outliers.
4. The variables should be normally or near-to-normally distributed.

Simple Regression Interpretation

PowerPoint: Simple Regression

For this example, you can examine the output for the simple regression model by opening the link below.

- Chapter 5 – Simple Regression

The first slide provides you with an example output in which physical illness is regressed on mental distress scores. In other words, we are using mental distress to predict your physical illness score.

Model Summary (PhysicalIllness)

R	R Square	Adjusted R Square	Std. Error of the Estimate
.47	.22	.22	5.29

ANOVA (PhysicalIllness)

	Sum of Squares	df	Mean Square	F	Sig.
Regression	2922.56	1	2922.56	104.38	.000
Residual	10219.66	365	28.00		
Total	13142.22	366			

Coefficients (PhysicalIllness)

	Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
(Constant)	7.25	.87	.00	8.31	.000	5.54	8.97
MentalDistress	.23	.02	.47	10.22	.000	.18	.27

Green: Model statistics

Light Blue: Degrees of Freedom

Blue: R value

Orange: R^2 value

Red: Variable test statistcs

If you go to the second slide, we have circled certain important elements of the statistical output. For example, you can see the statistics for the overall model, that is, the F statistic and the overall significance of the model which have been circled in green. The degrees of freedom (*df*) have been circled in light blue (see the ANOVA (Physical Illness) box). Although these statistics are more important in multiple or hierarchical regression, they are useful here because they provide you with an indication of the significance of the overall model.

The next thing to examine is the significance of the individual predictor variable (circled in red), with the standardised b value (known as beta), the t value, and the significance shown.

Don't forget to look at the R Squared value (circled in orange), which can be interpreted as a percentage of variance explained in the outcome by the predictor. In this example, mental distress accounts for 22% of the shared variance with physical illness.

Finally, the *r* value (circled in a darker blue) which is the Pearson correlation coefficient, represents the standardised line of best fit for the model.

Correlations

		PhysicalIllness	MentalDistress
PhysicalIllness	Pearson Correlation	1.00	.47
	Sig. (2-tailed)		.000
	N	367	367
MentalDistress	Pearson Correlation	.47	1.00
	Sig. (2-tailed)	.000	
	N	367	367

As you can see the *r* value, shows the same value as the correlation between the two variables, which we discussed in Chapter Four. This is expected in simple regression, but the values for predictors in regression models will change as we include more predictor variables to make a more complex model.

Simple Regression Write Up

Here is an example of how you can write up the results of a simple regression analysis:

In order to test the research question, a simple regression was conducted, with mental distress as the predictor, and levels of physical illness as the dependent variable. Overall, the results showed that the utility of the predictive model was significant, $F(1,365) = 104.38$, $R2 = .22$, $p < .001$. Mental distress explained a large amount of the variance

between the variables (22%). The results showed that mental distress was a significant positive predictor of physical illness (β=.47, t= 10.22, p< .001).

Section 5.3: Multiple Regression Explanation, Assumptions, Interpretation, and Write Up

Learning Objectives

At the end of this section you should be able to answer the following questions:

- Explain the difference between Multiple Regression and Simple Regression.
- Explain the assumptions underlying Multiple Regression.

Multiple Regression is a step beyond simple regression. The main difference between simple and multiple regression is that multiple regression includes two or more independent variables – sometimes called predictor variables – in the model, rather than just one.

As such, the purpose of multiple regression is to determine the utility of a set of predictor variables for predicting an outcome, which is generally some important event or behaviour. This outcome can be designated as the outcome variable, the dependent variable, or the criterion variable. For example, you might hypothesise that the need to belong will predict motivations for Facebook use and that self-esteem and meaningful existence will uniquely predict motivations for Facebook use.

Before beginning your analysis, you should consider the following points:

- Regression analyses reveal relationships among variables (relationship between the criterion variable and the linear combination of a set of predictor variables) but do not imply a causal relationship.
- A regression solution – or set of predictor variables – is sensitive to combinations of variables. Whether a predictor is important in a solution depends on the other predictors in the set. If the predictor of interest is the only one that assesses some important facet of the outcome, it will appear important. If a predictor is only one of several predictors that assess the same important facet of the outcome, it will appear less important. For a good set of predictor variables – the smallest set of uncorrelated variables is best.

PowerPoint: Venn Diagrams

Please click on the link labeled "Venn Diagrams" to work through an example.

- Chapter Five – Venn Diagrams

In these Venn Diagrams, you can see why it is best for the predictors to be strongly correlated with the dependent variable but uncorrelated with the other Independent Variables. This reduces the amount of shared variance between the independent variables. The illustration in Slide 2 shows logical relationships between predictors, for two different possible regression models in separate Venn diagrams. On the left, you can see three partially correlated independent variables on a single dependent variable. The three partially correlated independent

variables are physical health, mental health, and spiritual health and the dependent variable is life satisfaction. On the right, you have three highly correlated independent variables (e.g., BMI, blood pressure, heart rate) on the dependent variable of life satisfaction. The model on the left would have some use in discovering the associations between those variables, however, the model on the right would not be useful, as all three of the independent variables are basically measuring the same thing and are mostly accounting for the same variability in the dependent variable.

There are two main types of regression with multiple independent variables:

- Standard or Single Step: Where all predictors enter the regression together.
- Sequential or Hierarchical: Where all predictors are entered in blocks. Each block represents one step.

We will now be exploring the single step multiple regression:

All predictors enter the regression equation at once. Each predictor is treated as if it had been analysed in the regression model after all other predictors had been analysed. These predictors are evaluated by the shared variance (i.e., level of prediction) shared between the dependant variable and the individual predictor variable.

Multiple Regression Assumptions

There are a number of assumptions that should be assessed before performing a multiple regression analysis:

1. The dependant variable (the variable of interest) needs to be using a continuous scale.
2. There are two or more independent variables. These can be measured using either continuous or categorical means.
3. The three or more variables of interest should have a linear relationship, which you can check by using a scatterplot.
4. The data should have homoscedasticity. In other words, the line of best fit is not dissimilar as the data points move across the line in a positive or negative direction. Homoscedasticity can be checked by producing standardised residual plots against the unstandardized predicted values.
5. The data should not have two or more independent variables that are highly correlated. This is called multicollinearity which can be checked using Variance-inflation-factor or VIF values. High VIF indicates that the associated independent variable is highly collinear with the other variables in the model.
6. There should be no spurious outliers.
7. The residuals (errors) should be approximately normally distributed. This can be checked by a histogram (with a superimposed normal curve) and by plotting the of the standardised residuals using either a P-P Plot, or a Normal Q-Q Plot .

Multiple Regression Interpretation

For our example research question, we will be looking at the combined effect of three predictor variables – perceived life stress, location, and age – on the outcome variable of physical health?

PowerPoint: Standard Regression

Please open the output at the link labeled “Chapter Five – Standard Regression” to view the output.

- Chapter Five – Standard Regression

Slide 1 contains the standard regression analysis output.

Model Summary (PhysicalIllness)

R	R Square	Adjusted R Square	Std. Error of the Estimate
.50	.25	.24	5.22

ANOVA (PhysicalIllness)

	Sum of Squares	df	Mean Square	F	Sig.
Regression	3240.86	3	1080.29	39.61	.000
Residual	9901.35	363	27.28		
Total	13142.22	366			

On Slide 2 you can see in the red circle, the test statistics are significant. The F-statistic examines the overall significance of the model, and shows if your predictors as a group provide a better fit to the data than no predictor variables, which they do in this example.

The R^2 values are shown in the green circle. The R^2 value shows the total amount of variance accounted for in the criterion by the predictors, and the adjusted R^2 is the estimated value of R^2 in the population.

Coefficients (PhysicalIllness)

	Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
(Constant)	3.03	1.67	.00	1.81	.071	-.26	6.31
Gender	2.11	.65	.15	3.23	.001	.83	3.40
Age	-.01	.02	-.02	-.49	.625	-.05	.03
PercievedStress	.40	.04	.47	9.96	.000	.32	.48

Moving on to the individual variable effects on Slide 3, you can see the significance of the contribution of individual predictors in light blue. The unstandardized slope or the B value is shown in red, which represents the change caused by the variable (e.g., increasing 1 unit of perceived stress will raise physical illness by .40). Finally, you can see the standardised slope value in green, which are also known as beta values. These values are standardised ranging from +/-0 to 1, similar to an *r* value.

We should also briefly discuss dummy variables:

Coefficients (PhysicalIllness)

	Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
(Constant)	3.03	1.67	.00	1.81	.071	-.26	6.31
Gender	2.11	.65	.15	3.23	.001	.83	3.40
Age	-.01	.02	-.02	-.49	.625	-.05	.03
PercievedStress	.40	.04	.47	9.96	.000	.32	.48

A dummy variable is a variable that is used to represent categorical information relating to the participants in a study. This could include gender, location, race, age groups, and you get the idea. Dummy variables are most often represented as dichotomous variables (they only have two values). When performing a regression, it is easier for interpretation if the values for the dummy variable is set to 0 or 1. 1 usually resents when a characteristic is present. For example, a question asking the participants "Do you have a drivers license" with a forced choice response of yes or no.

In this example on Slide 3 and circled in red, the variable is gender with male = 0, and female = 1. A positive Beta

(B) means an association with 1, whereas a negative beta means an association with 0. In this case, being female was associated with greater levels of physical illness.

Multiple Regression Write Up

Here is an example of how to write up the results of a standard multiple regression analysis:

In order to test the research question, a multiple regression was conducted, with age, gender (0 = male, 1 = female), and perceived life stress as the predictors, with levels of physical illness as the dependent variable. Overall, the results showed the utility of the predictive model was significant, $F(3,363) = 39.61$, $R2 = .25$, $p < .001$. All of the predictors explain a large amount of the variance between the variables (25%). The results showed that perceived stress and gender of participants were significant positive predictors of physical illness ($\beta = .47$, $t = 9.96$, $p < .001$, and $\beta = .15$, $t = 3.23$, $p = .001$, respectively). The results showed that age ($\beta = -.02$, $t = -0.49$ $p = .63$) was not a significant predictor of perceived stress.

Section 5.4: Hierarchical Regression Explanation, Assumptions, Interpretation, and Write Up

Learning Objectives

At the end of this section you should be able to answer the following questions:

- Explain how hierarchical regression differs from multiple regression.
- Discuss where you would use "control variables" in a hierarchical regression analyses.

Hierarchical Regression Explanation and Assumptions

Hierarchical regression is a type of regression model in which the predictors are entered in blocks. Each block represents one step (or model). The order (or which predictor goes into which block) to enter predictors into the model is decided by the researcher, but should always be based on theory.

The first block entered into a hierarchical regression can include "control variables," which are variables that we want to hold constant. In a sense, researchers want to account for the variability of the control variables by removing it before analysing the relationship between the predictors and the outcome.

The example research question is "what is the effect of perceived stress on physical illness, after controlling for age and gender?". To answer this research question, we will need two blocks. One with age and gender, then the next block including perceived stress.

It is important to note that the assumptions for hierarchical regression are the same as those covered for simple or basic multiple regression. You may wish to go back to the section on multiple regression assumptions if you can't remember the assumptions or want to check them out before progressing through the chapter.

Hierarchical Regression Interpretation

PowerPoint: Hierarchical Regression

For this example, please click on the link for Chapter Five – Hierarchical Regression below. You will find 4 slides that we will be referring to for the rest of this section.

- Chapter Five – Hierarchical Regression

For this test, the statistical program used was Jamovi, which is freely available to use. The first two slides show the

steps to get produce the results. The third slide shows the output with any highlighting. You might want to think about what you have already learned, to see if you can work out the important elements of this output.

Linear Regression

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.202	0.0408	0.0356	7.75	2	364	< .001
2	0.497	0.2466	0.2404	39.61	3	363	< .001

Slide 2 shows the overall model statistics. The first model, with only age and gender, can be seen circled in red. This model is obviously significant. The second model (circled in green) includes age, gender, and perceived stress. As you can see, the F statistic is larger for the second model. However, does this mean it is significantly larger?

To answer this question, we will need to look at the model change statistics on Slide 3. The R value for model 1 can be seen here circled in red as .202. This model explains approximately 4% of the variance in physical illness. The R value for model 2 is circled in green, and explains a more sizeable part of the variance, about 25%.

Linear Regression

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.202	0.0408	0.0356	7.75	2	364	< .001
2	0.497	0.2466	0.2404	39.61	3	363	< .001

Model Comparisons

Comparison							
Model		Model	ΔR^2	F	df1	df2	p
1	-	2	0.206	99.1	1	363	< .001

The significance of the change in the model can be seen in blue on Slide 3. The information you are looking at is the R squared change, the F statistic change, and the statistical significance of this change.

Model Specific Results Model 1

Model Coefficients - PhyisicalIllness

Predictor	Estimate	SE	t	p	Stand. Estimate
Intercept	16.5620	1.0953	15.12	< .001	
Gender	1.9883	0.7378	2.69	0.007	0.138
Age	−0.0661	0.0238	−2.78	0.006	−0.143

Model Specific Results Model 2

Model Coefficients - PhyisicalIllness

Predictor	Estimate	SE	t	p	Stand. Estimate
Intercept	3.0281	1.6711	1.812	0.071	
Gender	2.1133	0.6550	3.227	0.001	0.1471
Age	−0.0107	0.0218	−0.489	0.625	−0.0231
PercievedStress	0.4019	0.0404	9.957	< .001	0.4692

On Slide 4, you can examine the role of each individual independent variable on the dependant variable. For model one, as circled in red, age and gender are both significantly associated with physical illness. In this case, age is negatively associated (i.e. the younger you are, the more likely you are to be healthy), and gender is positively associated (in this case being female is more likely to result in more physical illness). For model 2, gender is still positively associated and now perceived stress is also positively associated. However, age is no longer significantly associated with physical illness following the introduction of perceived stress. Possibly this is because older persons are experiencing less life stress than younger persons.

Hierarchical Regression Write Up

An example write up of a hierarchal regression analysis is seen below:

In order to test the predictions, a hierarchical multiple regression was conducted, with two blocks of variables. The first block included age and gender (0 = male, 1 = female) as the predictors, with difficulties in physical illness as the dependant variable. In block two, levels of perceived stress was also included as the predictor variable, with difficulties in perceived stress as the dependant variable.

Overall, the results showed that the first model was significant $F(2,364) = 7.75$, $p = .001$, $R2 = .04$. Both age and gender were significantly associated with perceived life stress ($b = -0.14$, $t = -2.78$, $p = .006$, and $b = .14$, $t = 2.70$, $p = .007$, respectively). The second model ($F(3,363) = 39.61$, $p < .001$, $R2 = .25$), which included physical illness ($b = 0.47$, $t = 9.96$, $p < .001$) showed significant improvement from the first model $\Delta F(1,363) = 99.13$, $p < .001$, $\Delta R2 = .21$, , Overall, when age and location of participants were included in the model, the variables explained 8.6% of the variance, with the final model, including physical illness accounted for 24.7% of the variance, with model one and two representing a small, and large effect size, respectively.

Section 5.5: Chapter Five Self-Test

It is a good idea to test your knowledge after you have completed each chapter. This section presents a Self-Test for Chapter Five. Please complete the test now to assess your learning of the ideas inside this chapter.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/statisticsforresearchstudents/?p=548#h5p-8

PART VI

CHAPTER SIX - COMPARING THREE OR MORE GROUP MEANS

Hello everyone, and welcome to the sixth chapter of the University of Southern Queensland's online, open access textbook for statistics for research students.

The aim of this chapter is to discuss the testing of differences between mean values on a central variable of interest between three or more groups. When we examined differences between two groups we used a form of the t-test. However, when moving to an examination of differences between three or more groups we can extend the logic of the t-test to a method known as Analyses of Variance or ANOVA.

There are some slides that appear via links within Chapter Six. Please look for these as you review the current chapter.

Section 6.1: Between Versus Within Group Analyses

Learning Objectives

At the end of this section you should be able to answer the following questions:

- Explain how an ANOVA test differs from the t-test.
- Explain the purpose of planned contrasts and post hoc tests.
- Explain the difference between planned contrasts and post hoc tests.

As discussed before, examining and understanding how groups of individuals can differ is one of the key goals of psychology. There are many different groups of people that make up society, each with their own biology, environment, and values. It makes sense that many of these groups would vary on different psychological constructs and behaviour. In psychology, examining these differences can be key to understanding differing mental and social processes. This is important when considering if treatments for mental health concerns actually work, or if there are long-term trends in behaviours across ages and groups.

There are two main types of difference tests: those that look at differences between groups, and those that look at differences within groups. Between groups differences examine how two groups can differ from each other across a variable. Within-groups differences are similarly important. You're not looking at the differences between two groups, but rather the differences between the same group taken at two time points. Remember that when we talk about differences, we are talking about the difference between the mean of one group/time point and the mean of another. To examine the differences between or within groups you also need to know the standard deviations of both means you are comparing, as well as the number of participants. With the mean, the measure of variance within the samples (standard deviation, and the number of data points or participants, you can use the t-test to calculate if the difference between the two means is statistically significant.

One-Factor ANOVA

You may have seen in journal articles, reference to an ANOVA and asked yourself what is an ANOVA test, and how does it differ from the t-tests discussed in previous chapters? Historically, ANOVA is an analysis for experimental design (like t-tests), that aims to determine the influence of the effect of one Independent Variable (IV) on the Dependent Variable (DV). The main aim of an ANOVA is to evaluate if the means of the samples are sufficiently different from each other to suggest they are representative of different populations. The main difference between an ANOVA and a t-test, is that you can use a one-way ANOVA to test if there are differences in more than two groups. An example would be if you wanted to check to see if there were differences in class attendance rates for first, second, and third-year university students in a particular course.

The ANOVA is a two-stage test.

Stage 1:

The first stage is an overall or omnibus test. This stage tests the overall hypothesis. In this example, you would be looking to test whether attendance rates differ across first, second, and third-year university students. If a significant result is found, that would indicate that rates do differ, and a second stage is required.

Stage 2:

Stage two introduces specific procedures for testing group differences. In this example, if a main effect is found, how do you know which year level differs from which year level? A main effect can be described as the effect of an independent variable on a dependent variable averaged across the levels of any other independent variables. For example, do first-year university students attend class at a higher rate than second or third years university students, and do second and third-year university students also differ?

Ts and Fs

The t-test is an examination of the mean difference and the standard error of the mean difference, and only examines the difference between two groups relative to the spread of their scores.

The F-test examines the variability of sample means (explained variance) to an estimate of error variance (unexplained variance). This can tell you if there is a difference between multiple groups, though the test can only tell if there is a difference, not tell which groups are different when there are more than 2 groups.

Planned contrasts or Posthoc Tests

Planned contrasts/comparisons are performed by a researcher who has a specific hypothesis in mind. Going back to the class attendance example, if you hypothesised that the first-year students would have a higher class attendance rate than second or third students, with no hypotheses relating to comparing second and third-year students, you would plan to compare or contrast first-year class attendance with second and then third-year class attendance. It would involve breaking down the variance accounted for by the model into component parts.

Posthoc tests are used when a researcher is expecting to find a difference, but isn't sure in which group that difference will be found. Posthoc tests aim to compare every group (similar to carrying out several t-tests) but familywise error is controlled (depending on the choice of posthoc procedures). Familywise error rate (FWER) refers to the probability of making at least one Type I error within the family of tests under consideration. Another alternative is to use t-tests, but you need to be aware that can inflate the chances of a type 1 error.

Data Screening

The usual data screening methods are applied including having no out-of-range responses; correct coding; and missing data has been dealt with. Being aware of outliers is important for ANOVAs but, assuming there are no major outliers, the F model is still interpretable with data showing minor violations of normality (Tabachnick & Fidell).

Section 6.2: One-Way ANOVA Assumptions, Interpretation, and Write Up

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What are assumptions that need to be met before performing a Between Groups ANOVA?
- How would you interpret a Main Effect in a One-Way ANOVA?

One-Way ANOVA Assumptions

There are a number of assumptions that need to be met before performing a Between Groups ANOVA:

1. The dependent variable (the variable of interest) needs to be a continuous scale (i.e., the data needs to be at either an interval or ratio measurement).
2. The independent variable needs to have two independent groups with two levels. When testing three or more independent, categorical groups it is best to use a one-way ANOVA, The test could be used to test the difference between just two groups, however, an independent samples t-test would be more appropriate.
3. The data should have independence of observations (i.e., there shouldn't be the same participants who are in both groups.)
4. The dependent variable should be normally or near-to-normally distributed for each group. It is worth noting that while the t-test is robust for minor violations in normality, if your data is very non-normal, it would be worth using a non-parametric test or bootstrapping (see later chapters).
5. There should be no spurious outliers.
6. The data must have homogeneity of variances. This assumption can be tested using Levene's test for homogeneity of variances in the statistics package. which is shown in the output included in the next chapter.

Sample Size

A consideration for ANOVA is homogeneity. Homogeneity, in this context, just means that all of the groups' distribution and errors differ in approximately the same way, regardless of the mean for each group. The more incompatible or unequal the group sizes are in a simple one-way between-subjects ANOVA, the more important the assumption of homogeneity is. Unequal group sizes in factorial designs can create ambiguity in results. You can test for homogeneity in PSPP and SPSS. In this class, a significant result indicates that homogeneity has been violated.

Equal cell Sizes

It is preferable to have similar or the same number of observations in each group. This provides a stronger model that tends not to violate any of the assumptions. Having unequal groups can lead to violations in normality or homogeneity of variance.

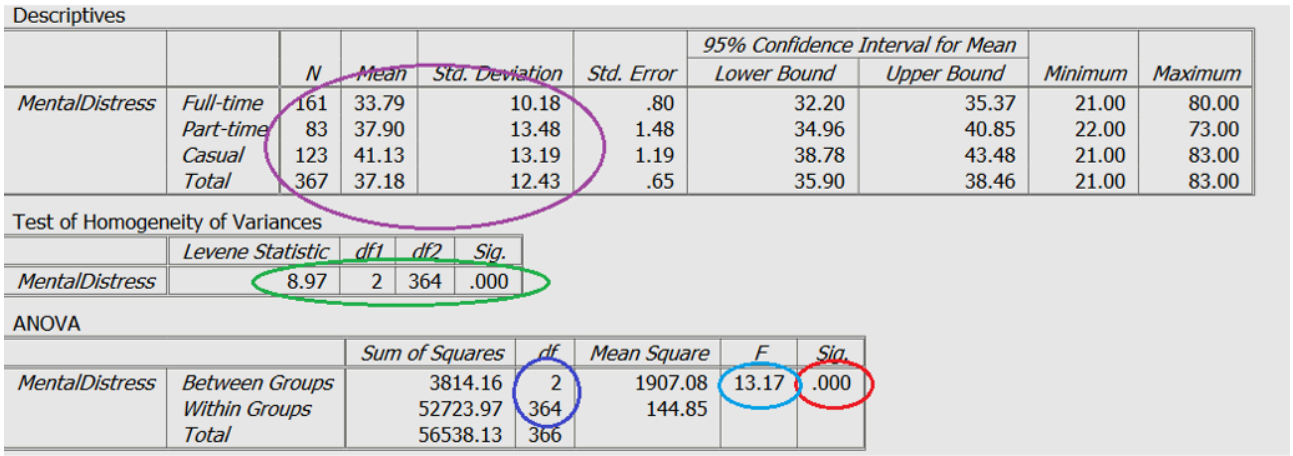
One-Way ANOVA Interpretation

Below you click to see the output for the ANOVA test of the Research Question, we have included the research example and hypothesis we will be working through is: Is there a difference in reported levels of mental distress for full-time, part-time, and casual employees?

PowerPoint: One Way ANOVA

Please have a look at the following slides:

- Chapter Six – One Way ANOVA



Descriptives

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
MentalDistress	Full-time	161	33.79	10.18	.80	32.20	35.37	21.00	80.00
	Part-time	83	37.90	13.48	1.48	34.96	40.85	22.00	73.00
	Casual	123	41.13	13.19	1.19	38.78	43.48	21.00	83.00
	Total	367	37.18	12.43	.65	35.90	38.46	21.00	83.00

Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
MentalDistress	8.97	2	364	.000

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
MentalDistress	Between Groups	3814.16	2	1907.08	13.17	.000
	Within Groups	52723.97	364	144.85		
	Total	56538.13	366			

Main Effects

As can be seen in the circled section in red on Slide 3, the main effect was significant. By looking at the purple circle, we can see the means for each group. In the light blue circle is the test statistic, which in this case is the F value. Finally, in the dark blue circle, we can see both values for the degrees of freedom.

Posthoc Tests

In order to run posthoc tests, we need to enter some syntax. This will be covered in the slides for this section, so please do go and have a look at the syntax that has been used. The information has also been included on Slide 4.

Posthoc Test Results

These are the results. There are a number of different tests that can be used in posthoc differences tests, to control for type 1 or type 2 errors, however, for this example none have been used.

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
MentalDistress	Between Groups	3814.16	2	1907.08	13.17	.000
	Within Groups	52723.97	364	144.85		
	Total	56538.13	366			

Multiple Comparisons (MentalDistress)

	(I) Are you employed?	(J) Are you employed?	Mean Difference (I - J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
LSD	Full-time	Part-time	-4.11	1.63	.012	-7.31	-.92
		Casual	-7.34	1.44	.000	-10.18	-4.51
	Part-time	Full-time	4.11	1.63	.012	.92	7.31
		Casual	-3.23	1.71	.060	-6.59	.14
	Casual	Full-time	7.34	1.44	.000	4.51	10.18
		Part-time	3.23	1.71	.060	-.14	6.59

Results

As can be seen in the red and green circles on Slide 6, both part-time and casual workers reported higher mental distress than full-time workers. This can be cross-referenced with the means on the results slide. As be seen in blue, there was not a significant difference between casual and part-time workers.

One-Way ANOVA Write Up

The following text represents how you may write up a One Way ANOVA:

A one-way ANOVA was conducted to determine if levels of mental distress were different across employment status. Participants were classified into three groups: Full-time (n = 161), Part-time (n = 83), Casual (n = 123). There was a statistically significant difference between groups as determined by one-way ANOVA ($F(2,364) = 13.17$, $p < .001$). Post-hoc tests revealed that mental distress was significantly higher in participants who were part-time and casually employed, when compare to full-time ($Mdiff = 4.11$, $p = .012$, and $Mdiff = 7.34$, $p < .001$, respectively). Additionally, no difference was found between participants who were employed part-time and casually ($Mdiff = 3.23$, $p = .06$).

Section 6.3 Repeated Measures ANOVA Assumptions, Interpretation, and Write Up

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What are the assumptions that need to be met before performing a Repeated Measures ANOVA?
- What different hypotheses are testable when using an ANOVA repeated measures design?

Let's look now at a repeated measures ANOVA. A repeated measures ANOVA is used to compare three or more group means when the participants are the same in each group. Usually, this would occur when a participant is repeated tested, particularly if you are evaluating an intervention. For example, you could use a repeated measures ANOVA to better understand whether there is a difference in anxiety levels after a group therapy program. You could measure anxiety levels at three time points, such as at the start of the program, one month after the program is completed, and then six months after the program is completed.

Repeated Measures ANOVA Assumptions

For a Repeated Measures ANOVA there are two or more independent variables (factors) that can be denoted by the levels of each Independent Variable (IV).

For example, in a design with 2 IVs, the ANOVA is described as A X B ANOVA

(A = Number of levels of IV1; B = Numbers of levels of IV2)

Meanings of the levels of factors can change when researchers shift between between-subjects designs and within-subjects designs.

For between-subjects design, levels can be thought of as different groups of the factor.

For within-subjects design, levels can be thought of as different conditions of the factor.

There are also several assumptions that go with Repeated Measures ANOVA:

1. The dependent variable (the variable of interest) needs a continuous scale (i.e., the data needs to be at either an interval or ratio measurement).
2. The "within-groups" variable must have two or more groups that are related or have "matched pairs". As in the Paired-samples T-test, matched pairs mean that the same participants are present in both groups (i.e. measuring the same persons twice).
3. There should be no spurious outliers.
4. The dependent variable should be normally or near-to-normally distributed for each group. It is worth noting that the t-test is robust for minor violations in normality, however, if your data is very non-normal, it would be worth using a non-parametric test or bootstrapping (see later chapters).
5. The data must have what is known as "sphericity". This means that amount of variable across the differences for all of the groups (both within and between) must be equal or near-to-equal the variances of the differences between all combinations of related groups must be equal. This can be tested using statistical software.

Repeated Measures ANOVA Interpretation

In our example below, the researchers are interested in the effects of more than 1 IV on a DV, which in this case are the effects of a social support program and gender on perceived levels of life satisfaction.

In the scenario, there are 2 factors or IVs:

Life satisfaction (pre- and post-program: 2 levels) X gender (male and female: 2 levels)

PowerPoint: Repeated Measures ANOVA

To view the output for the example Repeated Measures ANOVA output, please click on the following link:

- Chapter Six – Repeated Measures ANOVA

For this test, the statistical program used was Jamovi, which is freely available to use. The first two slides show the steps to get produce the results. The third slide shows the output with any highlighting.

Hypothesis

Often when using an ANOVA repeated measures design, three or more different hypotheses are testable. A main effect of factor 1, a main effect of factor 2, and an interaction effect of both factor 1 and factor 2. As you can see in the figure, there is a 3-way design to this ANOVA = A X B X C

Interaction

Answers the question similar to "moderation effect" in regression

Determines whether differences that can be attributed to a factor are consistent at all levels of the other factor/s or differences that can be attributed to a factor depending on the level of the other factor.

Run the analysis

In this example, we are interested in the effects of more than 1 IV on the DV. In this case, we want to know the effects of a social support program and gender on perceived levels of life satisfaction. We want to know if the program works in improving perceptions of life satisfaction (Factor 2), and if gender can play a role in levels of life satisfaction (Factor 2). Finally, we want to know if both males and females improve at the same rate in the intervention.

In the scenario, there are 2 factors or IVs

Program (pre- and post-program: 2 levels) X gender (male and female: 2 levels)

Effects to test in ANOVA :

- Main program effect (Time)
- Main gender effect (Gender)
- Interaction effect of time and gender

Jamovi Interface

As you can see, there is a number of things to enter into Jamovi, which we will cover in the slides.

Homogeneity

In this case, as we are only texting two groups, and two time points, Homogeneity isn't a major consideration in the overall analysis. However, as you can see in red, you do want these values to be greater than .05 when testing multiple groups/timepoints.

Repeated Measures ANOVA

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_p
Time	5690.4	1	5690.39	1473.02	< .001	0.801
Time ✻ Gender	29.0	1	28.98	7.50	0.006	0.020
Residual	1410.0	365	3.86			

Note. Type 3 Sums of Squares

Within groups

The within-groups results can be seen here. In green is the main effect for time, which is our primary within groups variable. The results are significant, which shows a change from pre- to post-scores. The large red circle shows the interaction of gender and time, which is of interest as well. As this interaction is significant, this shows that when moving from pre- to post-intervention, male and female participants scores' show a different rate of change. The small red circle is the overall degrees of freedom for the model, which you will need when reporting the results.

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_p
Gender	1301	1	1301.2	19.4	< .001	0.050
Residual	24482	365	67.1			

Note. Type 3 Sums of Squares

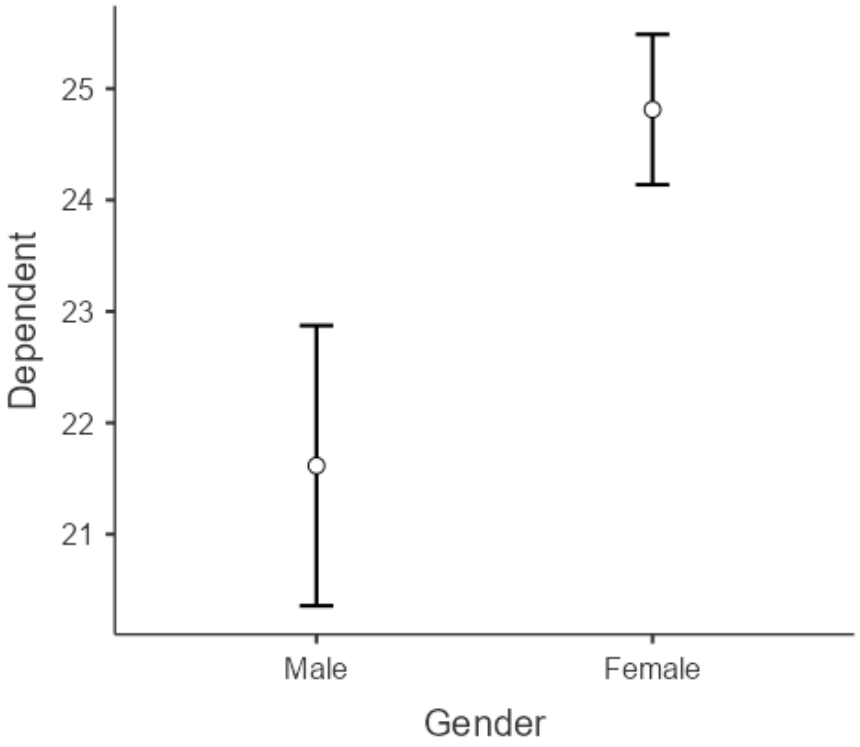
Between groups

As you can see in the large red circle, the results of the between-groups analysis shows that there is a difference between males and females.

Post Hoc Comparisons - Gender

Comparison							
Gender		Gender	Mean Difference	SE	df	t	p
Male	-	Female	−3.20	0.726	365	−4.40	< .001

Gender



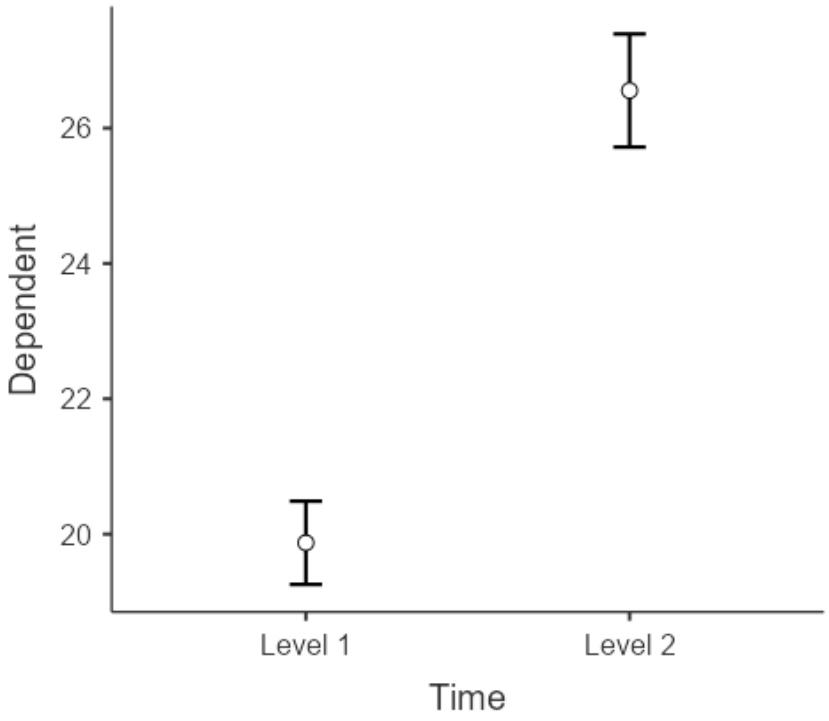
Estimated Marginal Means - Gender

			95% Confidence Interval	
Gender	Mean	SE	Lower	Upper
Male	21.6	0.640	20.4	22.9
Female	24.8	0.343	24.1	25.5

Post Hoc Comparisons - Time

Comparison							
Time		Time	Mean Difference	SE	df	t	p
Level 1	-	Level 2	−6.68	0.174	365	−38.4	< .001

Time



Estimated Marginal Means - Time

			95% Confidence Interval	
Time	Mean	SE	Lower	Upper
Level 1	19.9	0.313	19.3	20.5
Level 2	26.6	0.425	25.7	27.4

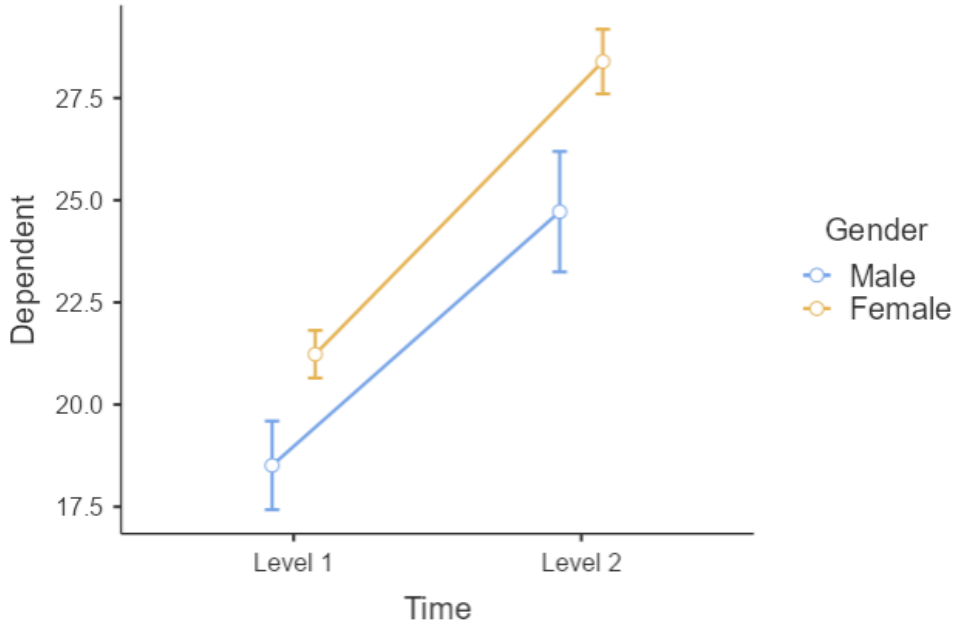
Main effects

The results of the main effects can be seen here. In the green circles are the means for the groups/time points, and in the red is the actual comparison tests. As we can see here, there was an improvement in scores from pre- to post-intervention, and females tended to score more highly.

Post Hoc Comparisons - Time ✻ Gender

Comparison									
Time	Gender		Time	Gender	Mean Difference	SE	df	t	p
Level 1	Male	-	Level 1	Female	−2.72	0.626	365	−4.35	< .001
		-	Level 2	Male	−6.21	0.307	365	−20.22	< .001
		-	Level 2	Female	−9.88	0.682	365	−14.48	< .001
	Female	-	Level 2	Male	−3.49	0.805	365	−4.33	< .001
		-	Level 2	Female	−7.16	0.165	365	−43.49	< .001
Level 2	Male	-	Level 2	Female	−3.67	0.850	365	−4.32	< .001

Time ✻ Gender



Estimated Marginal Means - Time ✻ Gender

				95% Confidence Interval	
Gender	Time	Mean	SE	Lower	Upper
Male	Level 1	18.5	0.551	17.4	19.6
	Level 2	24.7	0.749	23.2	26.2
Female	Level 1	21.2	0.296	20.6	21.8
	Level 2	28.4	0.402	27.6	29.2

Interactions

In this case, we have used follow-up post hoc t-tests to test the difference across gender at each timepoint. This is to further examine the interaction of gender and time. The means for each group at each timepoint can be seen in green, and the results of the posthoc are in red. In this case, it is apparent from the results, that while both males

and females improved pre to post-intervention, female participants started with greater life satisfaction, and saw greater improvement following the intervention than males.

Repeated Measures ANOVA Write Up

The following presents a write up for a Repeated Measures ANOVA:

The analysis showed both a significant main effect for time, $F(1,365) = 1473.02$, $p < .001$, $\eta p2 = 0.80$, and gender, $F(1,365) = 19.40$, $p < .001$, $\eta p2 = 0.05$, as well as a significant interaction of time and gender $F(1,365) = 7.50$, $p = .006$, $\eta p2 = 0.02$. Two follow up independent samples t-tests were conducted comparing life satisfaction for male and female participants at time 1 and time 2; $t(365) = -4.35$, $p < .001$, $d = 0.25$, and $t(365) = 4.32$, $p < .001$, $d = 0.27$, respectively. The results showed that female participants reported higher scores than males at both pre- and post-intervention. (Graphs should be included when reporting these results.)

Section 6.4: Chapter Six Self-Test

Please review your learning by completing the below self-test:

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/statisticsforresearchstudents/?p=739#h5p-9

PART VII

CHAPTER SEVEN - MODERATION AND MEDIATION ANALYSES

Hello everyone, and welcome to the seventh chapter of the University of Southern Queensland's online, open access textbook.

The aim of this seventh chapter is to discuss the use of two specific types of covariates: moderator and mediator variables. Moderator and mediator variables function in respect to a relationship between two other variables, such as reading ability and test performance. In a sense, moderator and mediator variables are simply special types of covariates. However, in the case of a moderator variable, the moderator functions by changing the strength or magnitude of an existing effect between two variables. For example, in the case of reading ability and test performance a third variable, such as anxiety, may change the pre-existing relationship between reading ability and test performance. In the case of a mediator variable, there is some amount of the effect between the two existing variables that is transmitted through the mediator. For example, in the case of reading ability and test performance, perhaps reading ability must function at least in part through a third variable such as attention capacity to effect test performance.

There are some slides that appear via links within Chapter Seven. Please look for these as you review the current chapter.

Section 7.1: Mediation and Moderation Models

Learning Objectives

At the end of this section you should be able to answer the following questions:

- Define the concept of a moderator variable.
- Define the concept of a mediator variable.

As we discussed in the lesson on correlations and regressions, understanding associations between psychological constructs can tell researchers a great deal about how certain mental health concerns and behaviours affects us on an emotional level. Correlation analyses focus on the relationship between two variables, and regression is the association of multiple independent variables with a single dependant variable.

Some predictor variables interact in a sequence, rather than impacting the outcome variable singly or as a group (like regression).

Moderation and mediation is a form of regression that allows researchers to analyse how a third variable effects the relationship of the predictor and outcome variable.

PowerPoint: Basic Mediation Model

Consider the Basic Mediation Model in this slide:

- Chapter Seven – Basic Mediation Model

We know that high levels of stress can negatively impact health, we also know that a high level of social support can be beneficial to health. With these two points of knowledge, could it be that social support might provide a protective factor from the effects of stress on health? Thinking about a sequence of effects, perhaps social support can mediate the effect of stress on health.

Mediation is a more complicated extension of multiple regression procedures. Mediation examines the pattern of relationships among three variables (Simple Mediation Model), and can be used on four or more variables.

Examples of Research Questions

Here are some examples of research questions that could use a mediation analysis.

- If an intervention increases secure attachment among young children, do behavioural problems decrease when the children enter school?
- Does physical abuse in early childhood lead to deviant processing of social information that leads to aggressive behaviour?

- Do performance expectations start a self-fulfilling prophecy that affects behaviour?
- Can changes in cognitive attributions reduce depression?

PowerPoint: Three Mediation Figures

Consider the Three Figures Illustrating Mediation from the following slides:

- Chapter Seven – Three Mediation Figures

Looking at this conceptual model, you can see the direct effect of X on Y. You can also see the effect of M on Y. What we are interested in is the effects of X on Y, accounting for the effects of M.

An example mediation model is that of the mediating effect of health-related behaviours on conscientiousness and overall physical health. Conscientiousness, or the personality trait associated with hardworking has relationship with overall physical health, but if an individual is hardworking, but does not perform health-related behaviours like exercise or diet control, then they are likely to be less healthy. From this, we can assume that health-related behaviours mediates the relationship between conscientiousness and physical health.

Section 7.2: Mediation Assumptions, The PROCESS Macro, Interpretation, and Write Up

Learning Objectives

At the end of this section you should be able to answer the following questions:

- Explain the assumptions that should be met before performing a mediation analysis.
- Explain the PROCESS Macro.
- What are the main ideas to focus on in mediation interpretation?

Mediation models focus on two effects – the direct effect and the indirect effect – and these can be combined into a measure of the model's total effect.

Effects in a Simple Mediation Model

Using the prior example of the effects of conscientiousness and physical health, the indirect effect is the product of a and b = ab, from the previous figure. This is the indirect effect of the pathway from X to M, and M to Y. The total model effect is the combined direct effect and the indirect effect. The total effect quantifies how much two cases that differ by one unit on X are estimated to differ on Y.

Mediation Assumptions

There are a number of assumptions that should be met before performing a mediation analysis.

1. The dependent, independent, and mediator variables (the variables of interest) need to be using a continuous scale.
2. The variables of interest (the dependent variable and the independent and mediator variables) should have a linear relationship, which you can check with a scatterplot
3. The data must not show multicollinearity (see Multiple Regression).
4. There should be no spurious outliers, and the distribution of the variables should be approximately normal.

The MedMod Macro

The advent of affordable personal computers with statistical software has prompted researchers to develop new tools for analyses. Jamovi provides a number of free modules for more advanced analyses, including the MedMod Macro for meditation and moderation. Another tool for mediation analyses is the PROCESS Macro, which is available as a free extension for SPSS.

PowerPoint: Hayes PROCESS Macro

The following slide provides information on MedMod by illustrating where it appears in the Jamovi menu, and by showing menu option:

- Chapter Seven – MedMod Macro

Mediation Interpretation

PowerPoint: Mediation Menu and Results

The linked slides provide an example of mediation output:

- Chapter Seven – Mediation Menu and Results

Mediation

Mediation Estimates

			95% Confidence Interval				
Effect	Estimate	SE	Lower	Upper	Z	p	% Mediation
Indirect	0.0494	0.0240	0.00228	0.0965	2.05	0.040	3.71
Direct	1.2823	0.0652	1.15450	1.4102	19.66	< .001	96.29
Total	1.3317	0.0614	1.21142	1.4520	21.69	< .001	100.00

Path Estimates

					95% Confidence Interval			
			Estimate	SE	Lower	Upper	Z	p
PercievedStress	→	FacetoFaceSocialSupport	−0.365	0.0505	−0.464	−0.2664	−7.24	< .001
FacetoFaceSocialSupport	→	MentalDistress	−0.135	0.0631	−0.259	−0.0115	−2.14	0.032
PercievedStress	→	MentalDistress	1.282	0.0652	1.155	1.4102	19.66	< .001

The total effect of the model can been seen in blue, with the direct effect (i.e. X and Y) in green. The indirect effect can be seen in purple, with the p value for the indirect effect can be found in orange. Now the interpretation of many of these statistics (p values etc) has been explained in previous lessons, but the main thing to focus on is the direct and indirect effects. If the direct effect is significant, then X does effect Y, and if there is a significant indirect effect then M does indeed mediate the relationship between X and Y.

Mediation Write Up

This mediation output results can be written up as follows:

A mediation analysiswas conducted to examine the mediating effect of social support on perceived stress and mental distress. The total effect of the model was found to be significant, *b*=1.33, z=21.69, BCa CI [1.21, 1.45], *p*<.001. It was found that there was a statistically significant direct effect, *b*=1.28, z=19.66, BCa CI [1.15, 1.41], *p*<.001. A statistically significant indirect effect was also found, *b*=0.05, *z*= 2.05, *p*=.040. These results suggest that social support partially mediated the relationship between perceived stress and mental distress

Section 7.3: Moderation Models, Assumptions, Interpretation, and Write Up

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What are some basic assumptions behind moderation?
- What are the key components of a write up of moderation analysis?

Moderation Models

Difference between Mediation & Moderation

The main difference between a simple interaction, like in ANOVA models or in moderation models, is that mediation implies that there is a causal sequence. In this case, we know that stress causes ill effects on health, so that would be the causal factor.

Some predictor variables interact in a sequence, rather than impacting the outcome variable singly or as a group (like regression).

Moderation and mediation is a form of regression that allows researchers to analyse how a third variable effects the relationship of the predictor and outcome variable.

Moderation analyses imply an interaction on the different levels of M

PowerPoint: Basic Moderation Model

Consider the below model:

- Chapter Seven – Basic Moderation Model

Would the muscle percentage be the same for young, middle-aged, and older participants after training? We know that it is harder to build muscle as we age, so would training have a lower effect on muscle growth in older people?

Example Research Question:

Does cyberbullying moderate the relationship between perceived stress and mental distress?

Moderation Assumptions

1. The dependent and independent variables should be measured on a continuous scale.
2. There should be a moderator variable that is a nominal variable with at least two groups.
3. The variables of interest (the dependent variable and the independent and moderator variables) should have a linear relationship, which you can check with a scatterplot.
4. The data must not show multicollinearity (see Multiple Regression).
5. There should be no significant outliers, and the distribution of the variables should be approximately normal.

Moderation Interpretation

PowerPoint: Moderation menu, results and output

Please have a look at the following link for the Moderation Menu and Output:

- Chapter Seven – Moderation Output

Moderation

Moderation Estimates

			95% Confidence Interval			
	Estimate	SE	Lower	Upper	Z	p
PercievedStress	1.2276	0.0574	1.11504	1.3401	21.38	< .001
Cyberbullying	1.0516	0.1673	0.72362	1.3795	6.28	< .001
PercievedStress ✻ Cyberbullying	0.0494	0.0229	0.00456	0.0943	2.16	0.031

Simple Slope Analysis

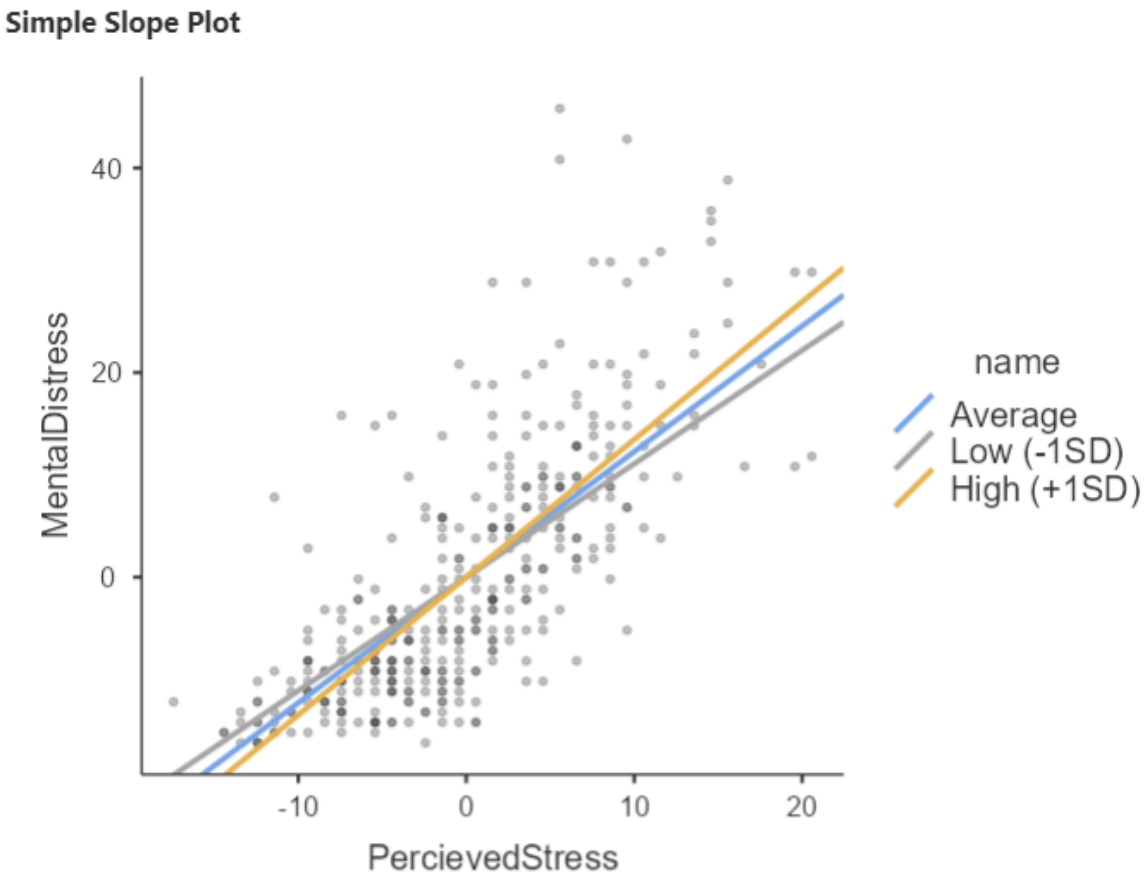
Simple Slope Estimates

			95% Confidence Interval			
	Estimate	SE	Lower	Upper	Z	p
Average	1.23	0.0577	1.114	1.34	21.3	< .001
Low (-1SD)	1.11	0.0806	0.951	1.27	13.8	< .001
High (+1SD)	1.35	0.0789	1.191	1.50	17.1	< .001

Note. shows the effect of the predictor (PercievedStress) on the dependent variable (MentalDistress) at different levels of the moderator (Cyberbullying)

Interpretation

The effects of cyberbullying can be seen in blue, with the perceived stress in green. These are the main effects of the X and M variable on the outcome variable (Y). The interaction effect can be seen in purple. This will tell us if perceived stress is effecting mental distress equally for average, lower than average or higher than average levels of cyberbullying. If this is significant, then there is a difference in that effect. As can be seen in yellow and grey, cyberbullying has an effect on mental distress, but the effect is stronger for those who report higher levels of cyberbullying (see graph).



Moderation Write Up

The following text represents a moderation write up:

A moderation test was run, with perceived stress as the predictor, mental distress as the dependant, and cyberbullying as a moderator. There was a significant main effect found between perceived stress and mental distress, *b*= -1.23, BCa CI [1.11, 1.34], z=21.38, *p* <.001, and nonsignificant main effect of cyberbullying on mental distress *b*= 1.05, BCa CI [0.72, 1.38], z=6.28, *p* <.001. There was a significant interaction found by cyberbullying on perceived stress and mental distress, *b*= -0.05, BCa CI [0.01, 0.09], z=2.16, *p* =.031. It was found that participants who reported higher than average levels of cyberbullying experienced a greater effect of perceived stress on mental distress (*b*= 1.35, BCa CI [1.19, 1.50], z=17.1, *p* <.001), when compared to average or lower than average levels of cyberbullying (*b*= 1.23, BCa CI [1.11, 1.34], z=21.3, *p* <.001, *b*= 1.11, BCa CI [0.95, 1.27], z=13.8, *p* <.001, respectively). From these results, it can be concluded that the effect of perceived stress on mental distress is partially moderated by cyberbullying.

Section 7.4: Chapter Seven Self-Test

It is a good idea to test your knowledge after you have completed each chapter. This section presents a Self-Test for Chapter Seven. Please complete the test now to assess your learning of the ideas inside this chapter:

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/statisticsforresearchstudents/?p=552#h5p-10

PART VIII

CHAPTER EIGHT - FACTOR ANALYSIS AND SCALE RELIABILITY

Hello everyone, and welcome to the eighth chapter of the University of Southern Queensland's online, open access textbook.

The aim of this eighth chapter is to discuss two methods to determine if individual "questions" or "test item" variables that measure a common concept or construct work together in a mathematically connected fashion. The first method we examine is factor analysis. If you recall from chapter four that correlations will estimate the amount of change shared by two variables, you can extend this logic to factor analysis because factor analysis is a way to estimate the shared change or variability between a much larger set of variables. Scale reliability analysis is another method to estimate the shared change or variability between a set of variables although the set of variables examined in scale reliability analysis is generally much smaller than what is used in factor analysis.

There are some slides that appear via links within Chapter Eight. Please look for these as you review the current chapter.

Section 8.1: Factor Analysis Definitions

Learning Objectives

At the end of this section you should be able to answer the following questions:

- How would you explain the aim of Factor Analysis?
- How is Factor Analysis related to measure development?

In psychology, we use many measures to capture psychological constructs. Many of you in Psychology would have encountered measures like the Depression Anxiety Stress Scale or the Satisfaction with Life Scale, or other such measures. These measures use many items to capture constructs like depression, well-being, or intelligence. These measures go through a development process, in which a number of items (i.e., test questions) are tested with a population, and the items are tested to see if they cluster together around a construct. For example, questions like 'I fell down', 'I often feel unhappy' or 'I find it hard to get excited about life' could measure depression. An item like 'I often feel happy' would not go with such items.

So how do you justify this statistically? Generally, one step is the use of Factor Analysis, which is a form of analysis that aims to "*summarise the interrelationships among the variables in a concise but accurate manner as an aid in conceptualisation*" (Gorsuch, 1983; p2.). This analysis method can be used to help develop scales and measures by removing items and developing factors. Therefore, at the heart of Factor Analysis is the reduction of a set of items, which is based on removing items that do not share a sufficient amount of variability with the other items in the set.

Section 8.2: EFA versus CFA

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What types of questions can be answered with EFA?
- What is the difference between EFA and CFA?

EFA vs CFA

There are two main schools of factory analyses: one that aims to explore a new measure and determine the factors within an unfactorized measure, and one that aims to confirm a pre-existing factor structure that has already been established. In this lesson we will be focusing on the first type, known as an exploratory factor analysis. As you can see here, there are differences between the EFA and the confirmatory factor analysis:

Exploratory FA	Confirmatory FA
(theory generating)	(theory testing)
Theory-weak literature base	Strong theory and/or strong empirical base
Determine the number of factors	Number of factors fixed a priori
Determine whether the factors are correlated or uncorrelated	Factors fixed a priori as correlated or uncorrelated
Variables are able to load on all factors	Variables must load on a specific factor or factors

Research Questions

You use exploratory factor analyses when you have questions like:

- How many reliable and interpretable factors/components are there in a set of variables?
- How many factors/components should be extracted?
- How much variance in a set of variables is accounted for by the retained factors/components?
- How are the factors/components interpreted?

This is a preliminary test, and is used primarily in the early stages of measurement or inventory development.

Variables & Level of Measurement

EFAs are different from most of the analyses we have covered here. “Independent variables” and “dependent

variables" are not terms used in EFA. The set of variables are the set of items that need to be reduced for the final measures, as well as establishing factors for this measure.

The items should be measured on an interval, ordinal scale, or nominal scales. The level of measurement determines the correlation matrix (matrix of association for decomposition).

A few important words about Likert response format/rating scale: Likert scales are usually best used in EFA.

Sample Size

Sample size is very important in EFAs, with the minimum recommended sample generally being at least cases from 100 individuals (Kline, 1994). Some sources recommend at least five cases per item, so if you have a scale with 30 items you need at least 150 participants (Hatcher, 1994).

Section 8.3: EFA Steps with Factor Extraction

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What are the two types of rotation?
- What is the difference between a Covariance Matrix and a Correlation Matrix?

There are a number of decisions that need to be made before running the analysis. Some of which we will discuss in a second, some of which we will discuss as we go through the SPSS commands. A short list of decision points within EFA are: 1) generating a Matrix of Association, 2) Method of Extraction, and 3) Method of Rotation.

Matrix of Association

When generating a Matrix of Association this refers to a Covariance Matrix of your input items, or alternatively a Correlation Matrix of those items, which is a simple transformation of the Covariance Matrix. Some software programs will generate the matrix for you from raw data. Other programs may require you to enter the matrix in some way, such as pasting it into syntax.

Method of Extraction

Methods of extraction refer to means of estimating the variability explained by the input items by generating a parsimonious set of factors. There are several traditional methods of extraction of factors within EFA. There are many theoretical paper written about this, but for most purposes you will mainly be using the maximum likelihood method.

Rotation Methods

Rotation refers to changing the scaling of the factor data vectors (or sets of information corresponding to each factor) according to geometric axis – like an X-Y Cartesian Axis or a space of greater dimension such as a 3 dimensional axis space. There are two main categories of rotation options – or how to express the factors as vectors in a dimensional space (a very mathematical thing to say). These categories of rotation are *oblique rotation*, which allows for small-to-moderate correlation of factors, and *orthogonal rotation*, which assumes that the factors are uncorrelated. In psychology, it is very rare to find concepts that are unrelated to each other. For example, depression is often related to things like well-being, anxiety, or health. As such we generally want to use an oblique rotation choice, such as ProMax rotation.

Example

Now we will focus on our example: How many unique factors are there for items that examine social support in person and on Facebook?

We will be discussing analysis methods as we go, hopefully, this will inform you on the process of EFAs.

PowerPoint: Exploratory Factor Analysis Menu

Have a look at the below slides, which illustrate how to run an EFA:

- Chapter Eight – Exploratory Factor Analysis Menu

For this test, the statistical program used was Jamovi, which is freely available to use. We select the KMO and Bartlett's test of sphericity. The reason why we want to use these tests is to make sure that an EFA is useful with your data: if significant, it indicates that the data is appropriate for a factor analysis.

As seen in the tslide we want to choose ProMax, and the load plots to see how our items load on to factors. On the bottom right, we want to suppress lower factor values (which will be explained shortly).

Sampling Adequacy

To measure and judge if we have adequate sampling adequacy we have the results of the KMO and Bartlett tests, which tests if an EFA is useful with your data: if significant, it indicates that the data is appropriate for a factor analysis.

Communalities

The next we want to check in our EFA results is communalities. These indicate the proportion of common variance in an item, relative to all the factors. The method of figuring this out is the sum of squared factor loadings for that variable (e.g., item) across all the factors. This result keeps track of how much of the original variance that was contained in a particular variable is still accounted for by all retained factors.

Section 8.4: EFA Determining the Number of Factors

Learning Objectives

At the end of this section you should be able to answer the following questions:

- How do you determine the number of factors suggested by a Scree Plot?
- What is Kaiser's rule for eigenvalues?

PowerPoint: Communalties Scree Plots and Number of Factors

Please have a look at the SPSS Output below for Communalities:

- Chapter Eight – Communalities Scree Plots and Number of Factors

Here is the communalities table from Jamovi. It is recommended that you reproduce this in your write-up.

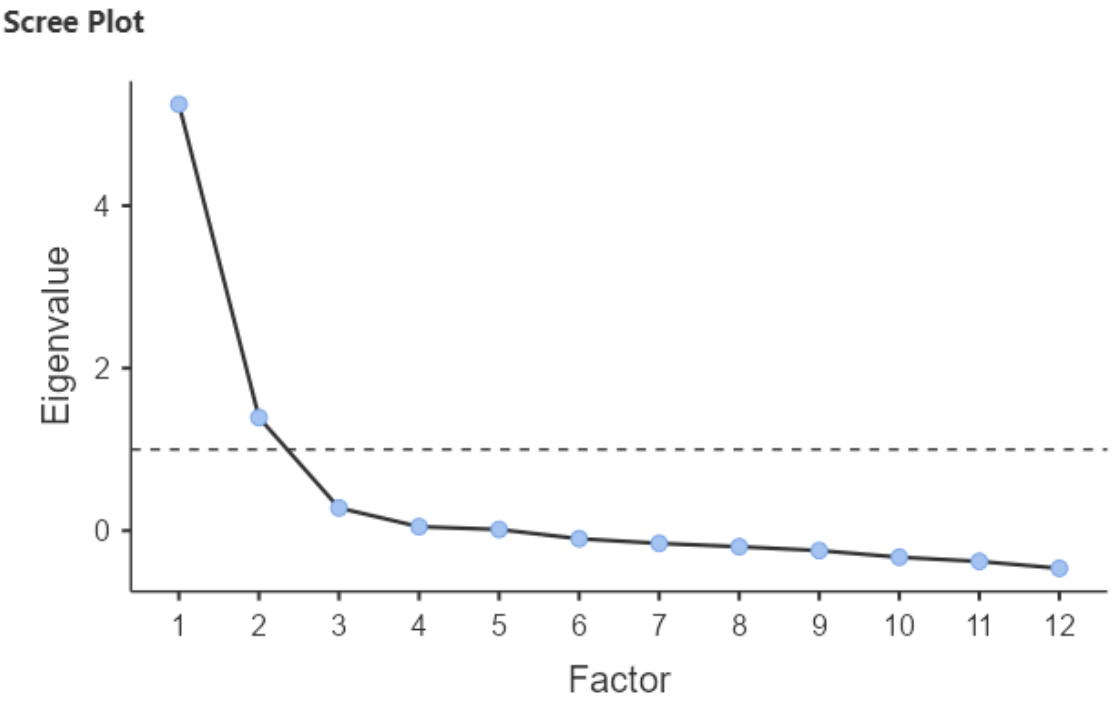
Factor Loadings

	Factor		
	1	2	Uniqueness
ISELSF03		0.550	0.702
ISELSF04		0.814	0.373
ISELSF05		0.685	0.475
ISELSF06		0.791	0.416
ISELSF09		0.651	0.457
ISELSF10		0.742	0.481
ISELFB03	0.732		0.441
ISELFB04	0.783		0.366
ISELFB05	0.845		0.309
ISELFB06	0.826		0.356
ISELFB09	0.817		0.317
ISELFB10	0.800		0.346

Note. 'Maximum likelihood' extraction method was used in combination with a 'promax' rotation

Number of Factors

Determining the number of factors measured by the items used can often be more of a judgment call than a simple yes or no. The number of factors can be determined using two things in the results: Scree plot, by using visual examination, and Kaiser's rule, which requires eigenvalues of greater than 1. Kaiser's rule is a useful indicator but needs to be supplemented with other types of information like the Scree Plot.



Scree Plot

Using the scree plot, you look for the point at which the line graph begins to 'flatten', and that will tell the number of factors present. "Flatness" on a curve can be usually defined as the portion of the curve following the last, large drop. You choose the point before the last, large drop on the Scree Plot to indicate the number of factors. As can be seen here, the line flattens at about 3, indicating that there are two major factors.

Eigenvalues

Initial Eigenvalues

Factor	Eigenvalue
1	5.2479
2	1.3926
3	0.2797
4	0.0494
5	0.0145
6	−0.0990
7	−0.1560
8	−0.1974
9	−0.2446
10	−0.3267
11	−0.3782
12	−0.4629

Factor Statistics

Summary

Factor	SS Loadings	% of Variance	Cumulative %
1	3.90	32.5	32.5
2	3.06	25.5	58.0

Eigenvalues

The finding of the scree plot is supported by the eigenvalues. The number of factors with eigenvalues greater than 1 is two. This explains 67.71% of the variance of the data.

Section 8.5: EFA Interpretation

Learning Objectives

At the end of this section you should be able to answer the following questions:

- How do we assign a name or label to factors?
- How many variables or items can be recommended per factor?

Pattern Matrix & Structure/Factor Matrix

The next piece of information is the pattern and the structure matrixes. These provide useful information on the factor loadings of the items. The pattern matrix shows the unique contribution of a variable to a factor, and is generally simpler to interpret. The structure matrix shows the shared variance that is ignored in the pattern matrix, and is more complicated to interpret. Note: Jamovi automatically produces the pattern matrix.

PowerPoint: Matrices

Please have a look at the below output for a Pattern Matrix:

- Chapter Eight – Matrices

Number of Variables per Factor

The replicability and strength of a component are determined by the number of variables per factor/component. A good rule is that a minimum of 4 variables is recommended per factor. All items that load on to factor should have a score of greater than .40 on the pattern matrix. Just be aware that items can "crossload" (i.e. have scores great than .40 on more than one factor).

Factor Loadings

	Factor		
	1	2	Uniqueness
ISELSF03		0.550	0.702
ISELSF04		0.814	0.373
ISELSF05		0.685	0.475
ISELSF06		0.791	0.416
ISELSF09		0.651	0.457
ISELSF10		0.742	0.481
ISELFB03	0.732		0.441
ISELFB04	0.783		0.366
ISELFB05	0.845		0.309
ISELFB06	0.826		0.356
ISELFB09	0.817		0.317
ISELFB10	0.800		0.346

Note. 'Maximum likelihood' extraction method was used in combination with a 'promax' rotation

Pattern Matrix

As you can see on the output, the pattern matrix clearly shows the loadings of each item onto the two factors. The two factors show that there are siz items for each, with one factor measuring in-person support, and the other Facebook-based support. This makes sense, both statistically, and theoretically. It is important to note that this is an optimal factor loading as there are two distinct factors, with no poor loadings (i.e. < .40) and crossloadings (have scores great than .40 on more than one factor). If there were any poor loadings/crossloadings, these items would be removed for the next round of EFAs.

Interpretation of Factors/Components

When naming the factors found it is usual to characterise the factor by assigning a name or label related to the semantic topic the items measure. This involves not only the knowledge of the area (science) but is an artistic pursuit sometimes. For this example we would have one factor labeled in-person support, and the other Facebook-based support. And we can see that they have a moderate correlation with each other.

Section 8.6: EFA Write Up

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What are some common elements of a Results Section in an EFA write up?
- What are some common Tables included in an EFA write up?

Guide (only) to Writing the Results section

A brief guide for the result write-up is: What analysis was conducted and for what purpose (include extraction method, number of items, and number of participants or sample size, including a test of sampling adequacy). What were the outcomes from data screening. You should present results from the analysis (not describe the analysis), i.e. answer the research questions:

- The criterion for determining the number of components to extract
- Method of rotation
- Cut-off used for retaining items for interpretation
- Appraise the solution (e.g. are there distinct components or are there many items with cross pattern coefficients?)
- Describe the components and name the components
- Estimate the internal consistency of each component
- A full example write-up will be provided.

Guide to Table/s to be Included

For the tables you should include:

A summary of eigenvalue, the total variance accounted by each component, and the cumulative percentage of total variance accounted by the four components.

Also the items, pattern coefficients (in descending order), and communalities (indicate before rotation since they are values from the output) of the items.

See the EFA Example Write-Up.

Section 8.7: Scale Reliability

Learning Objectives

At the end of this section you should be able to answer the following questions:

- How would you explain *internal consistency* reliability?
- What is a good range for alpha?

When researchers have a final set of items that form a functional scale, reliability calculations for all the scale items must be conducted by analysing the total number of items as a set.

This is most often done by using a type of reliability called *internal consistency* reliability, which is based on formulas that give an index of how much variability is shared and accounted for by the set of items, thus reflecting their degree of interrelationship. High internal consistency reliability reflects that items are consistent with other items in the set, and that the items are measuring the same construct.

Cronbach's alpha is a common statistic used to measure internal consistency, and it measures the correlation between multiple items in a factor. When using Cronbach's alpha, it is important to make sure that all items are related and measured in a similar way – but not with exact similarity in wording or in regard to aspects of the construct measured.

If dealing with multiple constructs or factors that cluster within a higher order factor, Cronbach's alpha should be run for both the total scale, and the items in each factor.

When conducting EFA procedures, scale reliability should be tested for each factor following the last EFA and the finalisation of the factor structure.

PowerPoint: Alpha Output

Please have a look at the link below for SPSS Output on Internal Consistency calculations:

- Chapter Eight – Alpha Output

As can be seen here, the alpha for the total scale used is 'good' as seen by a .85 value. You can also check and see if some items were deleted, would the alpha improve. If removing an item improves the score by .01-.02 it might be worth removing the item.

In general, "good" alpha estimates range from .7 – .9 (George & Mallery, 2003), with the following intreptations:

<.50 = Unacceptable
.51-.60 = Poor
.61-.70 = Questionable
.71-.80 = Acceptable
.81-.90 = Good
.91-.95 = Excellent

If the alpha is greater than .95, it is likely that there are a number of items that ask very similar, or the same question. For example, “I often feel down” and “I am often down”.

Section 8.8: Chapter Eight Self-Test

It is a good idea to test your knowledge after you have completed each chapter. This section presents a Self-Test for Chapter Eight. Please complete the test now to assess your learning of the ideas inside this chapter.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/statisticsforresearchstudents/?p=554#h5p-11

PART IX

CHAPTER NINE - NONPARAMETRIC STATISTICS

Hello everyone, and welcome to the ninth chapter of the University of Southern Queensland's online, open access textbook.

The aim of this ninth chapter is to discuss the idea of nonparametric statistics. Nonparametric statistics are types of test statistics with related formulas that can be used to estimate associations between two or more variables without basing these associations on changes from the mean. The arithmetic mean can be seriously influenced by extreme values and values that are dispersed in non-normal ways. Essentially if collections of data are not arranged according to the *normal distribution*, and when researchers can be reasonably sure that the actual distribution of variable values in a population is *not* normal, nonparametric statistics can then be used to better estimate associations between variables.

There are some slides that appear via links within Chapter Nine. Please look for these as you review the current chapter.

Section 9.1: Nonparametric Definitions

Learning Objectives

At the end of this section you should be able to answer the following questions:

- How would you define non-parametric methods?
- What types of assumptions are made for non-parametric methods?

Non-Parametric Methods

What can be done when the assumptions we have discussed in past lessons (t-tests, correlation etc.) are not maintained? There are tests used when a number of assumptions are not maintained for regular tests like t-tests or correlations (e.g. nonnormal distribution or small sample sizes). These tests – called non-parametric tests – use the same type of comparisons but with different assumptions.

Parametric Assumptions

Parametric statistics is a branch of statistics that assumes that sample data comes from a population that follows parameters and assumptions that hold true in most, in not all, cases. Most well-known elementary statistical methods are parametric, many of which we have discussed on this webpage.

Parametric Assumptions and the Normal Distribution

Normal distribution is a common assumption for many tests, including t-tests, ANOVAs and regression. Recall that parametric tests we have discussed here met the following assumptions of the normal distribution: minimal or no skewness and kurtosis of variables and error terms are independent across variables.

These assumptions allow us to infer a normal distribution in the population.

Non-Parametric Methods

Statistical methods which do not require us to make distributional assumptions about the data are called non-parametric methods. Non-parametric, as a term, actually does not apply to the data, but to the method used to analyse the data. These tests use rankings to analyse differences. Non-parametric methods can be used for different types of comparisons or models

Nonparametric Assumptions

1. Nonparametric tests make assumptions about sampling (that it is generally random).
2. There are assumptions about the independence or dependence of samples, depending on which nonparametric test is used, there are no assumptions about the population distribution of scores.

Nonparametric Tests and Level of Measurement

Variables at particular categorical levels of measurement may require Nonparametric Tests

Consider variables like autonomy, skill, income. Would such variables always follow a normal distribution? It is possible that when looking at income, you would expect the data to be skewed, as there are a small minority of the population who earn extremely high salaries.

Mean vs Median

When a distribution is highly skewed, the mean is affected by the high number of relative outliers. For example, when measuring something like income, where there are few high-income earners but many middle and low-income earners, the center of the distribution is quite skewed. This means that the median (i.e., the middle amount with 50% above and below this amount) is best used.

Sample Size

Sample size is another consideration when deciding if one should use a parametric or nonparametric test. Often, researchers will want to run a certain type of parametric test, but might not have the recommended minimum number of participants. Additionally, if the sample is very small, tests of normality often cannot be run. This is due to the lack of power needed to provide an interpretable result. When this is coupled with non-normal distributions of data, researchers might decide to use nonparametric tests.

Outliers

As discussed in previous chapters, parametric tests can only use continuous data for the dependant variable. This data should be normally distributed and not have any spurious outliers. However, some nonparametric tests can use data that is ordinal, or ranked for the dependant variable. These tests may also not be impacted severely by non-normal data or outliers. Each parametric test has its own requirements, so it is advisable to check the assumptions for each test.

Section 9.2: Choosing Appropriate Tests

Learning Objectives

At the end of this section you should be able to answer the following questions:

- To what extent do nonparametric equivalents exist for common parametric tests?
- What factors are considered when you decide to use nonparametric statistics?

Multiple Considerations Required

When deciding to use nonparametric statistics, an examination of whether the mean or the median is the best representation of the center of the data distribution is needed. If it is found that the median is the best representation of the data's center, then nonparametric tests are most likely to be appropriate, even with a larger sample of participants. If you have a small sample, then nonparametric statistics may be appropriate either way.

Different Tests

Each parametric test of difference we have discussed previously has a nonparametric equivalent, which can be used in cases where there is nonnormal data or a small sample size.

PowerPoint: Nonparametric Analogues

Please click on the link below to see slides with a chart of parametric tests with a nonparametric equivalent.

- Chapter Nine – Nonparametric Analogues

Section 9.3: Comparing Two Independent Conditions: The Mann– Whitney U Test

Learning Objectives

At the end of this section you should be able to answer the following questions:

- When examining differences between two independent groups, which nonparametric test can be used?
- When examining differences between two dependent groups, which nonparametric test can be used?

The Mann-Whitney U Test for two Independent Samples

When examining differences between two groups, Mann-Whitney U Test is best. This test examines the differences in median scores, as well as the size of the differences. Example: Is there a difference in the median number of Facebook Friends for male and female internet users? If a researcher wanted to compare Two Related Conditions, the test to use would be the Wilcoxon Signed-Rank Test.

Ranks

	Gender	N	Mean Rank
FacebookFriends	Male	82	159.46
	Female	285	191.06
	Total	367	

Test Statistics

	FacebookFriends
Chi-Square	5.65
df	1
Asymp. Sig.	.017

Interpretation for the Mann-Whitney U Test

As can be seen in the blue, there is a statistically significant difference, note the p value. The chi-squared value, and degrees of freedom are also needed for reporting. The median ranks indicate that female internet users have more Facebook Friends than male users.

Write-up

The results of the Mann-Whitney U Test indicate that female internet users reported having a statistically significantly higher number of Facebook Friends (Median = 191.06) than male users (Median = 159.46; U = 5.65, p = .017).

PowerPoint: Mann-Whitney

Please click on the slides below to see an example of interpretation for the Mann-Whitney U Test.

- Chapter Nine – Mann-Whitney

Section 9.4: Comparing Two Dependent Conditions or Paired Samples – Wilcoxon Sign-Rank Test

Learning Objectives

At the end of this section you should be able to answer the following questions:

- How would you interpret a Wilcoxon Sign-Rank Test?
- How is a Wilcoxon Sign-Rank Test related to the size of the differences in scores?

The Wilcoxon Test for Paired Samples

When examining within groups differences, Wilcoxon Signed Ranks Test is best. This test examines the differences in scores, as well as the size of the differences.

Example: The levels of perceived social support a group of Australians reported before engaging with a social skills building program and after completing the program.

Ranks

		N	Mean Rank	Sum of Ranks
SocialSupportPre - SocialSupportPost	Negative Ranks	259	184.30	47732.50
	Positive Ranks	68	86.70	5895.50
	Ties	40		
	Total	367		

Test Statistics

	SocialSupportPre - SocialSupportPost
Z	-12.24
Asymp. Sig. (2-tailed)	.000

Interpretation of the Wilcoxon Test

Using the same example from the t-test module, the levels of perceived social support a group of Australians reported before engaging with a social skills building program and after completing the program. As can be seen in red, the Z score, and in green the p value. These indicate that there is a difference in median pre- vs post-test rank score. The scores appear to improve from time 1 to time 2, which we can infer by the negative Z score, and the number of positive ranks in time 2.

Write-up

An example write up: A Wilcoxon Sign-Rank Test indicated that median post-test ranks for social support were statistically significantly higher than the pre-test ranks ($Z = -12.24$, $p < .001$).

PowerPoint: Wilcoxon Test

Please click on the slides below to see an example of interpretation for the Wilcoxon Sign-Rank Test.

- Chapter Nine – Wilcoxon Test

Section 9.5: Differences Between Several Independent Groups: The Kruskal–Wallis Test

Learning Objectives

At the end of this section you should be able to answer the following questions:

- What is the parametric antilog to the Kruskal-Wallis Test?
- What test would you use to compare differences between several related groups?

The Kruskal-Wallis H test for three or more Independent Samples

When examining the differences between three or more groups, Kruskal-Wallis H Test is best. This test examines the differences in median scores, as well as the size of the differences. This test examines the main effect of your variable, similar to an ANOVA. Example: Is there a difference in the median reported levels of mental distress for full-time, part-time, and casual employees? If one wanted to compare differences between several related groups, the test to use would be Friedman's ANOVA.

Ranks

	Are you employed?	N	Mean Rank
MentalDistress	Full-time	161	157.01
	Part-time	83	185.11
	Casual	123	218.59
	Total	367	

Test Statistics

	MentalDistress
Chi-Square	23.53
df	2
Asymp. Sig.	.000

Interpretation of the Kruskal-Wallis H test

As can be seen in the blue, there is a statistically significant difference, note the p value. The chi-squared value, and degrees of freedom are also needed for reporting. The median ranks indicate that casual employees have the highest

scores of mental distress. It is important to note that follow-up tests are required for individual group differences (like Mann-Whitney U Tests), similar to posthoc tests in ANOVA.

Write-up

A Kruskal-Wallis H test showed that there was a statistically significant difference in levels of mental distress, $\chi2(2) = 23.53$, $p < .001$, for full-time (Median = 157.01) , part-time (Median = 185.11) , and casual employees (Median = 218.58).

PowerPoint: Kruskal Wallis

Please click on the slides below to see an example of interpretation for the Kruskal-Wallis H test.

- Chapter Nine – Kruskal Wallis

Section 9.6: Chapter Nine Self-Test

It is a good idea to test your knowledge after you have completed each chapter. This section presents a Self-Test for Chapter Nine. Please complete the test now to assess your learning of the ideas inside this chapter.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://usq.pressbooks.pub/statisticsforresearchstudents/?p=556#h5p-12

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2nd Ed. Hillsdale, NJ, Erlbaum.

Field, A. P. (20). Discovering statistics using SPSS: (and sex, drugs and rock 'n' roll) (5th ed.). Los Angeles: SAGE Publications.

Gorsuch, R. L. (2014). *Factor analysis: Classic edition*. Routledge.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.)*. Boston: Allyn & Bacon.

Hatcher, L. (1994) *A Step-by-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling*. SAS Institute, Inc., Cary.

Kline, P. (1994). *An Easy Guide to Factor Analysis*. Abingdon-on-Thames: Routledge.

Tabachnick, B., & Fidell, L. (2019). *Using Multivariate Statistics (7th Ed)*. Boston, MA: Pearson Education Inc.