Research Methods and Statistics with jamovi

CATHERINE ORTNER

TRU OPEN PRESS
KAMLOOPS, BC

This book was produced with Pressbooks (https://pressbooks.com) and rendered with Prince.

Contents

Chapter 6: Three or More Means: The One-Way ANOVA

Chapter 7: Repeated Measures ANOVA

Chapter 8: Factorial Designs and Two-Way ANOVA

Chapter 9: Other ANOVAs - Mixed, 3-Way, ANCOVA

Chapter 10: Correlation - Associations Between Pairs of Variables

Chapter 11: Regression

Chapter 12: Chi-Square

Introduction

This book is intended as a resource for undergraduate students in psychology studying statistics and research methods. It was designed for students who have already completed introductory level research methods and statistics classes and who are learning how to use jamovi software to conduct and interpret quantitative data analyses. For each statistical test, there is a step guide to running the test in jamovi, interpreting the results, and writing up the results. There is some review of research methods (with a particular focus on experimental design), but this is by no means intended to be a comprehensive summary of research design.

About the Book

I have been teaching statistics and research methods for about 12 years. I always struggled to find a book for upper level undergraduate students that reviewed some of the basic concepts, meshed research methods and statistics, and also included a guide to using statistical software. I was having to compromise: students either had to purchase multiple books or forgo having a book to support their learning for major components of the course, and even my favourite books often had content that my students would have to skip because it was beyond the scope of the class. Eventually, inspired by my colleague Dr. Dana Wanzer (at the University of Wisconsin-Stout) who had written a book for her class, I decided to create an OER for my class. I drew heavily on Dana's book for the creation of the current book, and I am extremely grateful for the work she did in creating and sharing such an excellent resource.

Acknowledgements

Many sections of this book are drawn from Dr. Dana Wanzer's "Statistics with jamovi" licensed under a Creative Commons BY-SA (CC BY-SA) license version 4.0 and available on github. I am immensely grateful to Dana for creating such an accessible resource for her students and for permitting it to be shared and adapted by others.[You may want to include special adaptations you've made to the book, such as adding local features (**Canadian relevance, or Indigenizing the content – quality-added content**), updating accessibility standards, adding videos or interactive H5P content... Some authors also list specific adaptations for each chapter or section that was modified from an original OER. See the Suggest a Textbook BCcampus criteria to determine what elements that you've added that would be quality-added material or content.] Datasets are from the Learning Statistics with jamovi dataset module in jamovi (Navarro and Foxcroft, 2022).

The Open Press

The Open Press combines TRU's open platforms and expertise in learning design and open resource development. TRU Open Press supports the creation and reuse of open educational resources, while encouraging open scholarship and research.

Land Acknowledgement

Thompson Rivers University (TRU) campuses are situated on the ancestral lands of the Tk'emlúps te Secwépemc and the T'exelc within Secwepemcúl'ecw, the ancestral and unceded territory of the Secwépemc. The rich tapestry of this land also encompasses the territories of the St'át'imc, Nlaka'pamux, Tŝilhqot'in, Nuxalk, and Dakelh. Recognizing the deep histories and ongoing presence of these Indigenous peoples, we express gratitude for the wisdom held by this land. TRU is dedicated to fostering an inclusive and respectful environment, valuing education as a shared journey. TRU Open Press, inspired by collaborative learning on this land, upholds open principles and accessible education, nurturing respectful, reciprocal relationships through the shared exchange of knowledge across generations and communities.

Resource Development Team 2024

Author: Catharine Ortner, PhD

Publications Manager: Dani Collins, MEd

Copy Editing: Kaitlyn Meyers, BA, Christopher Ward, MA

Production: Jessica Obando Almache, BCS

Author support statement and list of CC licensed resources

Example (can be adapted by author): SITE NAME by AUTHOR NAME has been created from a combination of original content and materials compiled and adapted from a number of open text publications, including:

<textbook [hyperlinked]> (Author, Year) <textbook [hyperlinked]> (Author, year)

Example for specific adaptation of larger texts/ chapters:

Unless otherwise noted, Introductory Business Statistics with Interactive Spreadsheets – 1st Canadian Edition is (c) 2010 by Thomas K. Tiemann. The textbook content was produced by Thomas K. Tiemann and is licensed under a Creative Commons Attribution 3.0 Unported licence, except for the following changes and additions, which are (c) 2015 by Mohammad Mahbobi, and are licensed under a Creative Commons Attribution 4.0 International licence.

All examples have been changed to Canadian references, and information throughout the book, as applicable, has been revised to reflect Canadian content. One or more interactive Excel spreadsheets have been added to each of the eight chapters in this textbook as instructional tools. The following additions have been made to these chapters:

Chapter 4

- chi-square test and categorical variables
- null and alternative hypotheses for test of independence

Chapter 8

- simple linear regression model
- least squares method
- coefficient of determination
- confidence interval for the average of the dependent variable
- prediction interval for a specific value of the dependent variable

References

Navarro, D. J. and Foxcroft, D. R. (2022). Learning statistics with jamovi: A tutorial for psychology students and other beginners (Version 0.75). https://doi.org/10.24384/hgc3-7p15

Accessibility

The web version of Research Methods and Statistics with Jamovi has been designed to meet Web Content Accessibility Guidelines 2.0, level AA. In addition, it follows all guidelines in Appendix A: Checklist for Accessibility of the Accessibility Toolkit – 2nd Edition.

Includes:

- **Easy navigation.** This resource has a linked table of contents and uses headings in each chapter to make navigation easy.
- **Accessible videos.** All videos in this resource have captions.
- **Accessible images.** All images in this resource that convey information have alternative text. Images that are decorative have empty alternative text.
- **Accessible links.** All links use descriptive link text.

Accessibility Checklist

Element	Requirements	Pass
Headings	Content is organized under headings and subheadings that are used sequentially.	Yes
Images	Images that convey information include alternative text descriptions. These descriptions are provided in the alt text field, in the surrounding text, or linked to as a long description.	Yes
Images	Images and text do not rely on colour to convey information.	Yes
Images	Images that are purely decorative or are already described in the surrounding text contain empty alternative text descriptions. (Descriptive text is unnecessary if the image doesn't convey contextual content information.)	Yes
Tables	Tables include row and/or column headers with the correct scope assigned.	Yes
Tables	Tables include a title or caption.	Yes
Tables	Tables do not have merged or split cells.	Yes
Tables	Tables have adequate cell padding.	Yes
Links	The link text describes the destination of the link.	Yes
Links	Links do not open new windows or tabs. If they do, a textual reference is included in the link text.	Yes
Links	Links to files include the file type in the link text.	Yes
Video	All videos include high-quality (i.e., not machine generated) captions of all speech content and relevant non-speech content.	Yes
Video	All videos with contextual visuals (graphs, charts, etc.) are described audibly in the video.	Yes
H5P	All H5P activities have been tested for accessibility by the H5P team and have passed their testing.	Yes
H5P	All H5P activities that include images, videos, and/or audio content meet the accessibility requirements for those media types.	Yes
Font	Font size is 12 point or higher for body text.	Yes
Font	Font size is 9 point for footnotes or endnotes.	Yes
Font	Font size can be zoomed to 200% in the webbook or eBook formats.	Yes
Mobile Check	Layout displays properly on smaller screen sizes and is mobile-friendly.	

Known Accessibility Issues and Areas for Improvement

- SAMPLE – Tables use merged cells but they have been structured to work properly with screen readers – make sure tables do NOT have merged cells!
- SAMPLE – These videos do not have edited captions:

Other Formats Available

- SAMPLE – 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. The Digital PDF has passed the Adobe Accessibility Check.

OER Adoption Form

Please consider filling out a survey about this textbook to help us better understand how it's used and fits with the needs of our readers.

CHAPTER 1: RESEARCH DESIGN - A REFRESHER

Muddled Max's New Study

Introduction

Confounds, reliability, interval and ratio data, random assignment, extraneous variables—these terms should all sound familiar to you, but unless you have recently reviewed your notes from your previous research methods courses, you might be feeling a bit rusty on what they actually mean. A lot of this chapter will review what you learned in previous research methods courses. However, there are a few concepts that students often find tricky to fully grasp. Recognizing and defining these key terms is often much easier than applying and generating examples of them.

Furthermore, this chapter will start to highlight the connection between research methods and statistics, which is often overlooked in introductory research methods and statistics classes. Understanding how the two are connected will help you to understand why we do the things we do, both when designing studies and when analyzing our data.

This chapter also will provide a refresher on the scientific research process and on how to critique and design experimental and correlational studies.

But, first, let's meet Muddled Max. Max is a zealous but novice researcher who needs some help with research design. Max joins us today with their first idea. Max has noticed at parties that redheads often have clusters of people around them who want to talk. Max wondered: Do redheads appear more friendly than other people?

Max asks 20 people from their social circle to rate the images below in terms of their friendliness (on a scale from 1 = "not at all friendly" to 5 = "very friendly").

Images for Hair Colour Study

Figure A. *Red Hair. (Gabriel Silvério/Unsplash License)*

Figure A. *Brown Hair. (Yaroslav Shuraev/Pexels License)*

Ten friends rate image A and ten friends rate image B. Max then analyzes the results.

Activity 1.1

What's the problem? Take a moment to write down all the strengths and weaknesses of Max's study.

Media Attributions

- Figure A. Woman taking photo while showing smile by Gabriel Silvério, via Unsplash, is used under the Unsplash License.
- Figure B. Girl sitting at a table in a library and reading a book by Yaroslav Shuraev, via Pexels, is used under the Pexels License.

References

[reference citations for this section?]

The Research Process

Let's begin by considering an overview of the research process, shown in the figure below:

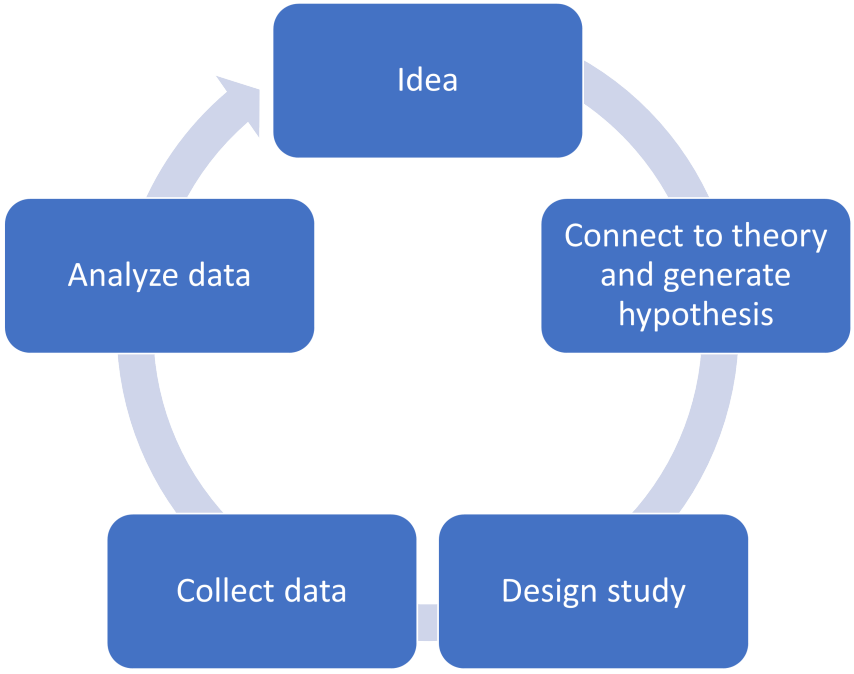


Figure 1.1. *The Research Process*

Generating the Hypothesis

Research usually starts with some research ideas and an observation. Where do those ideas come from? We might get an idea from so-called "common-sense." (Is it true that birds of a feather flock together? Do opposites attract?) We also might get an idea from observations in our daily lives. (Max used an observation from daily life when they wondered if red-heads tended to attract a lot of friends.) However, more typically, theories and past research drive future research. Or, sometimes, research questions come from practical problems that we want to solve. (How can we reduce student stress?)

We then collect data to see whether our expectations are correct. To do this, we need to identify and define our variables. Variables are anything that can be measured, and they can differ across people (assuming we are doing research with humans), or contexts, or time.

After we have selected our general research question, we generate a hypothesis. A hypothesis is usually derived from a **theory**, which is a general principle or set of principles that explains known findings about a particular topic. Note: a theory is not the same as a guess or a hunch. In everyday life when people say things like "I have a theory about ..." (e.g., "presidents of big corporations are all narcissistic!") they usually mean they have a guess about something. It may be based on some observations, but there is not a full set of principles that explain a body of data, which we need for something to be called a theory. A theory allows us to generate predictions about the results of future studies. A **hypothesis** or **prediction** (note that sometimes people distinguish between these terms, but for our purposes we can use them interchangeably) is the expectation of what will happen in the context of a particular study. It is different from a **research question**. For example, when Max asks: "Do redheads appear more friendly than other people?" that is a research question. To

make it into a hypothesis, it should be a statement, specifically in terms of the expected outcomes in the study: "I expect that redheads will receive higher ratings of friendliness than people with brown hair."

Note that we cannot *prove* a hypothesis to be correct. The data we obtained might *fit* the hypothesis, but it is possible that the hypothesis was incorrect and that it was just a coincidence that the data fit the hypothesis. Therefore, we say that a hypothesis or prediction is *supported*, but not proved.

A final note about hypotheses and theories: hypotheses and theories should be falsifiable. Falsification is the act of disproving a theory or hypothesis. Many aspects of Freud's theories about the structure of personality were not falsifiable.Let's consider an example. A self-described psychic comes into the lab to be tested, but they are unable to demonstrate their powers. The person explains it away by saying: "Well, it's just because you don't believe." Then if a believer is present, and the person still cannot demonstrate their powers, they might say: "It's because there is a strange energy in this room that is counteracting my powers." These excuses make the hypothesis unfalsifiable.

Designing the Study

When we design our study, there are various important considerations. What variables will we use, and how will we measure them? How will we address issues of reliability and validity? Will we use a correlational or experimental design?

Identifying the Variables

As we design our study, we need to think about our variables. In an experiment, we have independent and dependent variables. In Max's study of people's ratings of friendliness of redheads versus people with other hair colours, the **independent variable**, the variable that is directly changed or manipulated by the experimenter, is hair colour.

An independent variable should have at least two levels. The **levels** are simply the different values that the independent variable takes on. For example, the people in the photos have either red hair or brown hair, so there are two levels (red or brown). The **dependent variable**, the variable that is measured and that we expect to change as a result of the different values of the independent variable, is friendliness.

In non-experimental contexts (in correlational designs), we are less likely to use the terms independent and dependent variables. Instead, we refer to the **predictor**—the variable that we *think* might be causing a change—and the **outcome** or **criterion**—the variable that we *think* might be changed. Note, I say "think" because unless we actually manipulated an independent variable (as we do in an experiment), we cannot draw conclusions about cause and effect (more on that later).

Design	**Variable**	**Variable**
Correlational	Predictor	Outcome
Experimental	Independent variable	Dependent variable

As we design our variables, we need to consider levels of measurement. You probably recall from previous classes that variables can be considered nominal, ordinal, interval, or ratio. Let's have a quick refresher on what these mean.

Levels of Measurement

Nominal data means any number is simply used to assign a label to the variable (e.g., what is your major? Psychology = 1, History = 2, Geography = 3, etc.). In this case, the numbers are arbitrary. They have no mathematical properties, so the numbers do not mean anything except just to represent a category. If your variable is **binary** or **dichotomous**, there are only two categories (e.g., yes/no; pass/fail; student/non-student; dead/alive). Nominal could refer to when there are two or more than two categories (e.g., whether someone is an omnivore, vegetarian, vegan, or fruitarian).

Ordinal data indicates rank order (e.g., first, second, third fastest in a race, or in class). With ordinal data, there is no information regarding differences between scores (e.g., in a race, the first person could have finished a millisecond before the second, and the second could have finished with a 5 second gap to third—the spaces between the numbers are not the same). Also, with ordinal data, there is no "zero" rank (you cannot have a zeroth position).

If you have **interval data**, each score indicates an actual amount, and there are equal units separating any two adjacent scores. Zero scores is possible, but does not necessarily indicate a zero amount. For example, a score of zero on a memory test does not mean you have "no memory." The distance between a score of one and two is same as between two and three on a memory test, in that one more word was remembered in each case (but note, this could be tricky because the words are not necessarily equivalent in terms of how easily they are remembered).

Ratio data has a true zero that actually means there is zero amount of it present (e.g., if you are counting how many years someone spent living in Kamloops, the number of friends you have, or the number of classes attended). The scores measure an actual amount, and ratio statements are possible (e.g., if Sam attended 2 classes and Amrit attended 6 classes, Amrit attended 3 times as many classes as Sam).

Finally, it is worth mentioning the difference between discrete and continuous variables, because you will hear these terms used as well. **Discrete variables** are things that can only be measured in whole number amounts. Usually nominal and ordinal variables are discrete, but ratio data can also be discrete (e.g., the number of friends you have). **Continuous data** allows for fractions (at least in theory). For example, your feelings of happiness right now on a scale from 1 to 10 could be anything from 1 to 10 (even if I require you to give a whole number for your answer, theoretically your true feelings of happiness might be somewhere around 9.5 (e.g., if the pure joy of embarking on your statistics class is tempered only slightly by the fact that it is a cloudy day).

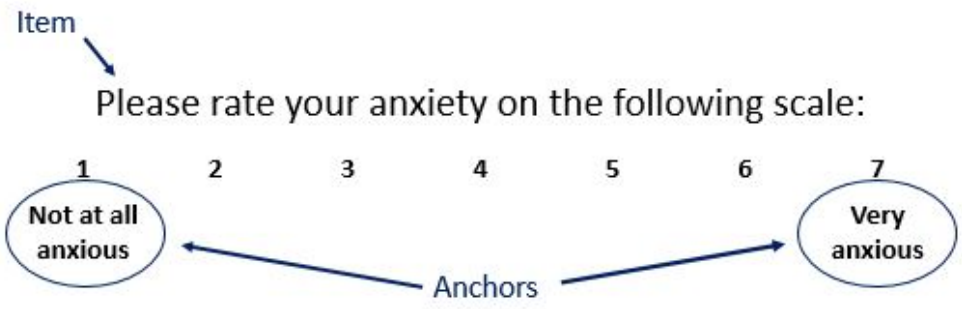


Figure 1.2. *Anatomy of a Rating Scale. Note. Words attached to numbers at each end of the scale are called anchors." The item is the wording of the rating scale itself.*

Why Do Levels of Measurement Matter?

Levels of measurement determine our ability to detect differences or change. You will have different degrees of **precision** or **sensitivity** according to the level of measurement. You will get a lot more information from

an interval or ratio scale than nominal scale. Similarly, you can get much more information on a continuous variable than on a dichotomous variable. Think about asking "Are you anxious: yes or no?" which is dichotomous, versus "How anxious are you on a scale from 1 to 10?" which is a continuous variable. The former may not allow you to detect changes in anxiety from before to after an intervention, for example. Perhaps people feel very anxious before the intervention and only mildly anxious after the intervention, but at both timepoints they are going to answer "Yes" to that question! Thus, we would say that the latter item is more **sensitive** than the former. At the same time, simply adding more points to the scale is not always better. We could ask "How anxious are you on a scale from 1 to 100?" and your participants might find it very difficult to answer. What is the difference between 64 and 65 on this scale when rating subjective anxiety?

Levels of measurement impact what kinds of statistical tests we can use (we will learn more on this as we get into each statistical test). Note that there is often a fair bit of confusion about the difference between interval and ratio data. Note that for the purposes of choosing a statistical test, we do not need to distinguish between interval and ratio data.

Reliability and Validity

An important consideration when designing our measures is measurement error. This is the difference between the actual value we are trying to measure, and the number we use to represent that value. Let's say you are going to treat your roommates to a home-baked chocolate cake. You are weighing some flour. You put flour into the bowl on your scales. The scale tells you that the weight is only 200 g. In fact, your scales are not very accurate, and you already have 300 g of flour in the bowl. However, you do not know this, so you add more flour. Sadly, your cake will be very dry and dense. There was a lot of measurement error here! To keep measurement error to a minimum, we need to increase validity and reliability.

Validity

There are several different types of validity. Very generally, **validity** refers to the extent to which an instrument measures what it set out to measure. More specifically, **content validity** is the extent to which the measure actually reflects the variable of interest (e.g., if I have developed a test of memory, does it measure memory alone, and not something else, like attention or fatigue; if I have a scale assessing clinical depression, does it assess only depression and not anxiety). **External validity** is the extent to which the results generalize to other situations or settings (**generalizability**). **Ecological validity** is a special type of external validity which relates to the extent to which research can be generalized to common real-world behaviours and natural situations. So, if you were studying the effects of study time on memory for a list of abstract nouns (like justice, hate, self-esteem, curiosity) you might ask yourself: Do we ever do this in everyday life? Would memory operate similarly if someone is trying to memorize a shopping list?

Later we shall talk about **internal validity**, which is a special type of validity related to the relationship between variables.

Reliability

Reliability is the ability of the measure to produce the same results under the same conditions. **Test–retest reliability** refers to whether the measurements are consistent. Is the same score produced each time a particular behaviour is measured? Are the scores stable over repeated testings? For reliability, I would expect

that if you rate a particular photo as attractive now, you will feel the same way about it 10 minutes from now, and tomorrow, and next week. Otherwise, it would be a pretty unreliable measurement that could have big effects on your results.

Why should we care about reliability? Reliability is an important consideration in our research design because often we want to see a change in scores as a result of our manipulation of the independent variable. We need to know that any change in scores is because of the manipulation, and not because the test is unreliable. Perfect reliability is very hard to achieve when measuring psychological variables. However, it is important to increase reliability as much as we can, as it affects our ability to detect change or differences, as you will see when we start to learn about statistical tests.

Research Design

Do you remember the difference between correlational and experimental research? Let's have a brief review here. The main differences are that with correlational research, we simply measure variables. In **correlational research**, the variables are measured to determine if there is an association; no variables are manipulated. Of course, this yields two big problems. The first problem is the **direction of cause and effect problem**. We do not know which variable caused which. The second problem is the **third variable problem**. Perhaps the variables are not causally related at all, and some uncontrolled third variable may be responsible for the observed association between the variables. (More on these issues in Chapter 10, Correlational Design.)

In **experimental research**, one or more variable(s) is systematically **manipulated** to see their effect (alone or in combination) on an outcome variable. Researchers vary the presence or strength of an independent variable and determine whether those variations have an impact on the behaviour or mental process in question (the dependent variable). The goal is to determine cause and effect. Does a certain variable influence a particular behaviour. Does watching TV violence affect aggression levels in children? Does mood affect memory performance? Experimental research allows us to determine cause and effect. Note that experimental research can take place in the laboratory or outside of the lab (in the "field").

So why would we ever use correlational research? There are a few situations when it is very useful. For example, sometimes it is unethical or impossible to manipulate a variable of interest (e.g., if you want to study the effects of maternal smoking on infant health outcomes; or when the focus of the research is on participant characteristics such as age or personality, which cannot be manipulated). Sometimes, the goal of the research is to describe variables and their potential relationships, which is particularly important in the early stages of research. Or, you may wish to predict future behaviour, such as when you are studying the relation between a predictor variable and some criterion (e.g., you might measure beliefs about intelligence at the beginning of university and look at how this predicts academic performance). Finally, correlational research is easier to conduct outside of the lab than experiments, so there is the potential for greater external and ecological validity. (And with better external validity, you may be better able to predict behaviours.)

But we must consider the direction of cause and effect problem. Consider this example: on the CBC there was a news item about research showing that people with dementia were more likely to be living close to highways. Did the dementia cause the people to make such choices, or does living near a highway somehow cause dementia? We cannot conclude from the (correlational) research that living near the highway caused dementia. Also, remember the third variable problem. Perhaps socioeconomic status is a causal variable both in where people live (houses closer to the highway might be cheaper) *and* in increasing the risk of dementia. Socioeconomic status would therefore be a third variable, and dementia and living location would not be causally related at all.

I hope by now that you would agree that correlational designs do not allow us to infer cause and effect. To infer cause and effect, we need to rule out all other explanations of the potential cause-effect relationship. We need to show that the effect is present when the cause is present and that when the cause is absent, the effect

is also absent. The way to do this is to compare two controlled situations: in one condition, the cause is present, and in the other condition, the cause is absent. To do this, we can run an experiment.

We shall have a closer look at research design for correlational studies later in the book (see Chapter 10, Correlational Design). For now, we are going to focus on experiments.

The Experiment

In this section, we shall focus on key aspects of experimental design. We are also going to begin to look at the crucial connections between research design and statistics. Briefly, we shall look at the distinction between between- and within-subjects design; next we shall consider the various influences on the dependent variable.

Some influences we are interested in (the independent variable[s]), but others we really do not want (extraneous variables: nuisance/noise variables and confounds). So we shall discuss how to maximize the effect of the independent variable (IV) and minimize the effects of the extraneous variables. And then you will see why this is important—in particular how it relates to our statistical test.

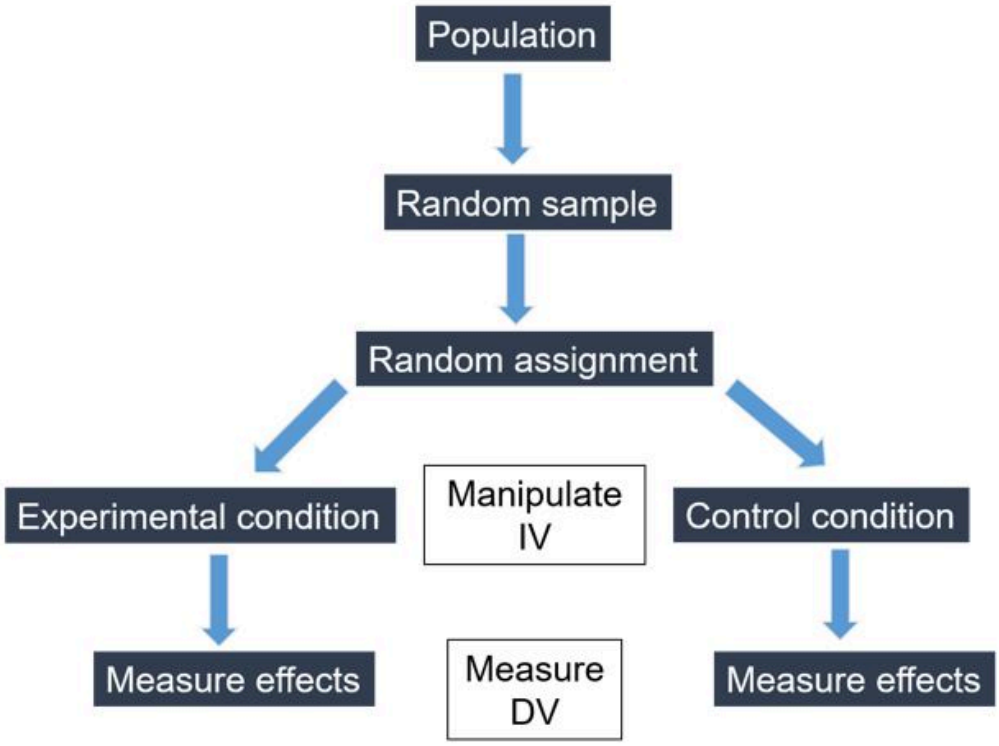


Figure 1.3. *The Experiment. (Adapted from Vermeulen, n.d.). Note. IV = independent variable; DV = dependent variable*

In the ideal experiment shown in Figure 1.3, we start by drawing a random sample from the population. We randomly assign our participants to an experimental condition or a control condition (note that in some experimental designs, there is no random assignment because participants complete both the experimental and the control conditions—more on that later).

With Max's study of hair colour and perceived friendliness (see earlier in this chapter), Max would randomly assign participants to view either image A or image B. Each participant would rate the friendliness of the person viewed in the photo. In this case, we just have two levels of the independent variable (red hair = experimental condition, or brown hair = control condition), but Max could add further levels (e.g., blond hair, black hair, bald).

The Choice of Between- or Within-Subjects Experiment Design

We have two different ways to collect data in an experiment. In the **between-groups** (or **between-subjects, independent**) **design**, we have different people in the experimental conditions. In the example above, Max would randomly assign participants to view the photo of the person with red hair *or* the person with brown hair. One of the benefits of this approach is that it reduces **demand characteristics** compared to the within-subjects design.

Demand characteristics are when cues in the experiment inform the participant how the experimenter expects them to behave. As a result, participants may conform to those expectations. Demand characteristics can become a **confound** (i.e., they threaten internal validity because it looks like the independent variable influenced the dependent variable, but really the difference in scores between the conditions is due to demand characteristics).

Alternatively, we could use a **within-subjects** (or **repeated measures design**) where the same people take part in all conditions—in other words, participants would be exposed to all levels of the IV. The same participants would view both the image of the person with red hair and the image of the person with brown hair. There are various benefits to this approach, including that it can be more economical (it should require fewer participants because it controls for participant variables). However, it may lead to practice effects, fatigue (now the experiment is twice as long), and demand characteristics.

Later in the book, we shall look a bit more closely at how to choose between a between- and a within-subjects design.

Influences on the Dependent Variable

We have reviewed independent variables (IV) and dependent variables (DV). However, these are not the only variables at work in our experiment. Two other types of variables influence the dependent variable, so we need to pay close attention to them. The figure below is probably the most important (if the most mundane) figure in this whole book. Usually, when we think about experiments, we think about the very left part of the figure (systematic variation of the independent variable reflecting the experimental effect and causing observed differences in the dependent variable). However, it is critical to consider that both confounds and extraneous variables also cause differences in the dependent variable and will impact your ability to draw conclusions from your experiment, albeit in different ways.

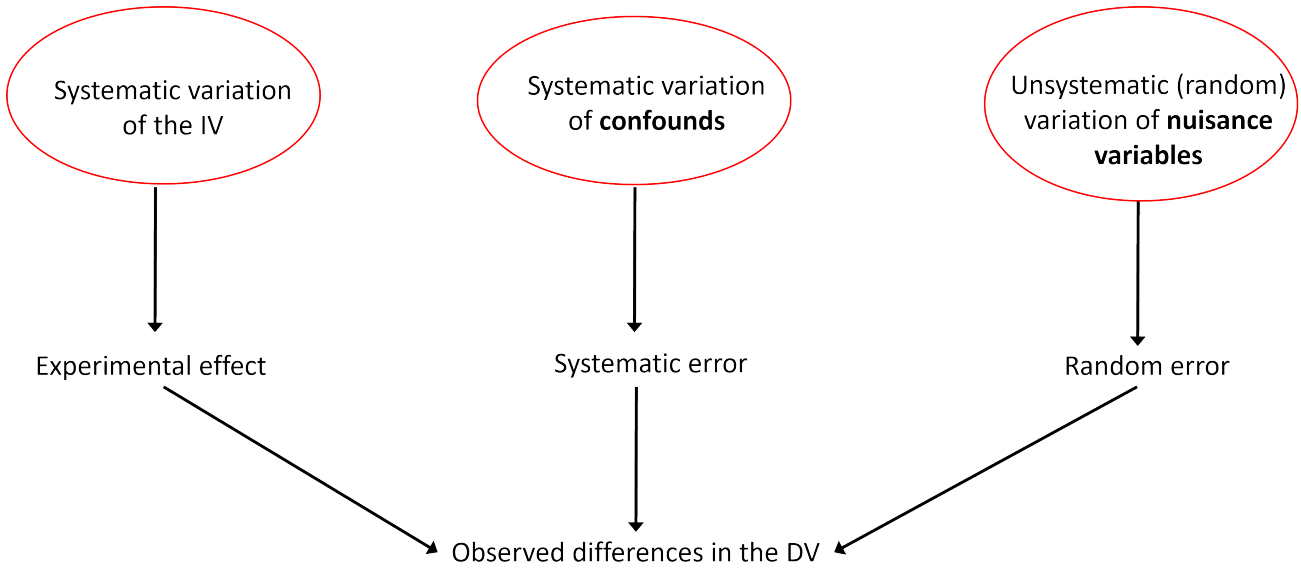


Figure 1.4. Sources of Variation in the Experiment. (Adapted from Vermeulen, n.d.)

As you can see in Figure 1.4, there are three causes of variation in the scores on the dependent variable: systematic variation of the independent variable; systematic variation of confounds; and unsystematic variation of nuisance (noise) variables. We want systematic variation of the independent variable—this is what leads to our experimental effect. On the other hand, we do not want systematic variation of confounds or unsystematic variation of nuisance variables, both of which represent error.

So back to Max's experiment, we would expect there to be some differences in friendliness scores due to Max's experimental manipulation (the person in the photo has red hair or brown hair). These differences are due to **systematic variation**. We call this systematic because we deliberately and methodically manipulated the independent variable (hair colour). The variability in the dependent variable due to the manipulation of the independent variable are what we want to observe.

Extraneous Variables: Nuisance Variables

The behaviour of participants is also influenced by other variables that the researcher was not interested in manipulating. Nuisance or noise variables are a type of extraneous variable that lead to **unsystematic variation**. By unsystematic, we mean that it is something we have not controlled or manipulated but that just happens to vary from one experimental condition to the next, or from one participant to the next. For example, imagine that one of Max's participants was recently dumped in a most awful way (imagine the worst possible way to end a relationship) by a person with red hair. Based on that experience, that participant might not feel positively inclined towards people with red hair at the moment. Another participant might have a best friend with red hair and, therefore, thinks that people with red hair are the friendliest on the planet! So, the different participants in different conditions are going to vary in their general feelings towards people with red hair due to their idiosyncratic experiences, which is going to influence their perceptions of the friendliness of people in the photos with red or brown hair. Also, the mood that each participant is in when they take part in the experiment will influence how friendly they perceive the person in the photo to be (regardless of the hair colour). These variables (recent experiences with people with red hair, current mood) are going to contribute to random variation in the scores on the dependent variable.

We call these variables **nuisance or noise variables**, because they were not manipulated, and they vary *at random* in the experiment. Nuisance variables can be characteristics of the participant (we call those

participant variables, as in the examples above), or they could be characteristics of the experiment (e.g., time of day, experimenter characteristic, etc.). You can probably think of some other nuisance variables that might be at play here. The key thing to remember is that nuisance variables are (usually) unknown variables that we have not measured but that influence the dependent variable in some way. We do not really want them because, as we shall see later, they make it harder, *statistically*, for us to detect the experimental effect. Imagine the static on the radio when you are trying to pick up a signal: nuisance variables are like the static and the signal is your experimental effect.

Extraneous variables: Confounds

What can also be a problem is if these extraneous variables start to vary in systematic ways. In other words, if they vary *along with* the independent variable. In that case, they can be confounds! A **confound** is a variable that varies *systematically* with (at the same time as) the independent variable.

So, for example, let's say Max has two research assistants working with them. One research assistant is really friendly to the participants and is the researcher who happens to administer the experiment to all the participants who view the photo of the person with red hair. The other researcher is grumpy, and they administer the experiment to all the participants who view the photo with of the person with brown hair. Now the research assistant's demeanour (friendly or grumpy) *covaries* with the manipulation of the independent variable (red hair or brown hair), so we would say that the research assistant's demeanour is a confound. Why is this a problem? Well, if Max finds that the person with red hair was rated as more friendly than the person with brown hair, we do not know if that is due to the manipulation of the independent variable (red or brown hair colour) or the effect of the confound (research assistant's friendly or grumpy demeanour). The confound has compromised internal validity.

The **internal validity** of our experiment refers to the extent to which the relationship between the independent variable and the dependent variable reflects only the relationship between them, and not effects due to other factors. Another way of saying this is to say that internal validity reflects the extent to which the apparent effect of the independent variable on the dependent variable is indeed due to our intended manipulation, and not due to the effect of confounding variables. We say that internal validity is good when we are confident that we have ruled out potential confounds.

What could be some other confounds in Max's experiment? Let's imagine that all the participants in the brown hair condition have to come into the lab for the experiment really early in the morning, but participants in the red hair condition come in later in the afternoon. Perhaps participants in the brown hair group perceive the person in the photo to be less friendly because they are tired and grumpy about coming in so early. Take a look back at the photos that Max used in the experiment (images A and B) earlier in this chapter. Do these give you a clue as to another potential confound?

In a nutshell:

Type of variable	**Do we want it?**	**Why or why not?**
Independent variable	Yes	It is the effect of interest
Nuisance variable	No	Causes random variation in the dependent variable, making it harder to detect differences in groups/conditions that are due to the manipulation of the independent variable (i.e., reduce power – see chapter 4 on power)
Confounding variable	No	Compromises internal validity, meaning that we might think that the difference between groups/conditions is due to the manipulation of the independent variable, but really it is due to the confounding variable

If you want further examples of confounding variables, see this great explanation by Karen Grace-Martin, "confusing statistical terms #11: confounder" from The Analysis Factor website.

So, what next? We need to maximize the effect of the independent variable and minimize the effects of extraneous variables, i.e., nuisance variables and confounds. Next, we shall look at how to do each of these things in turn.

Maximizing the Effect of the Independent Variable

We can maximize the effect by having a **strong manipulation**. In other words, we select amounts of the independent variable that are substantially different from one another. Max should ensure that the red-haired person has *really* red hair and that the brown-haired person has *clearly* brown hair, with no hints of red. We also need to ensure that exposure is sufficient for the manipulation to influence behaviour: a 20 second display of the photo should be more effective than a 500 ms display.

There are of course caveats to having a strong manipulation. First we have to consider whether the manipulation is ecologically valid. Sometimes, a manipulation done in the lab would not be as strong in the real world. Second, it might be unethical to use a very strong manipulation (e.g., if I want to look at the effects of hunger on cognitive performance, it would not be ethical to starve my participants for five days before they complete a cognitive task). Finally, we should try to use a **manipulation check**. This is a measurement to determine if an intended manipulation of an independent variable actually occurred. Did each condition of the independent variable actually have its intended effect? Usually we assess this after participants have completed the dependent variable (note, the manipulation check is *not* the measurement of the dependent variable).

For example, imagine you are investigating if a cold room temperature influences attention when studying. Participants are randomly assigned to sit in a cold room or a warm room while they study a list of words, and later they are tested on their recall. The dependent variable is the number of words recalled. The manipulation check might entail asking participants in all conditions to rate how cold or warm they felt while studying the word list. If you only vary the temperature of the room by a little bit (e.g., warm room = 21 degrees; cold room = 20 degrees), the change might not be detectable to participants (manipulation check!) and may not be sufficient to influence their performance on the task (the effect on the dependent variable!). Then you might say that it was a weak manipulation.

Manipulation checks are particularly useful in a pilot study, which is when you are just beginning your research and testing out your measures. They can also be used at the end of an actual experiment to demonstrate that the manipulation of the independent variable did have the desired effect (again, note that this is not the same as the effect on the dependent variable). What are the advantages of using a manipulation check? If, in a pilot study, the manipulation did not work, we can save the expense of running the actual experiment. Alternatively, if we run our study and get non-significant results (no relationship between our independent variable and dependent variable), we can check whether this is due to a problem with the manipulation (for example, perhaps it was not strong enough).

Reducing the Influence of Confounds

To maximize internal validity, thereby reducing the possibility of confounds, the main goal is to keep everything as consistent as possible from one condition of the experiment to another except for the variable that is being manipulated (the independent variable). Thus, you need to think carefully about your research stimuli and other aspects of your study design. You may have noticed that Max had different people pose for the photos with red

and brown hair. This is a design confound. It relates to an error in the experimental design. Max should have used the same person in each image presented to participants and used Photoshop to change the colour of the hair.

A summary of the many types of possible confounds is shown in the table below. You will sometimes see confounds described as "threats to internal validity," because they are exactly that!

Threat to internal validity	**Definition**
History effects	Non-manipulated external event occurs between two measurement timepoints
Maturation	Participants change over time and so produce different scores at two timepoints
Testing effect	Pretest sensitizes participants
Instrument decay	Characteristics of measurement instrument changes over time
Regression to the mean	Participants initially selected for extreme scores will be closer to mean on second measurement (statistical phenomenon)
Attrition	Participants with certain characteristics more likely to drop out of the study
Selection differences	Pre-existing differences in groups of participants
Demand characteristics	Subtle cues influences participants' behaviour to respond in hypothesized direction
Design confound	Unintended characteristic of experiment covaries with IV
Order effects	Exposure to one level of IV changes response to other level of IV (e.g., practice, fatigue, carryover effects)
Placebo effect	Scores on DV change/improve because participants expect the IV to have an effect

These above definitions are very brief, so you might want to look back at your notes from your introductory research methods class for a more detailed refresher. As you review these potential confounds: (a) for each one, consider whether it would be more likely to be an issue in a between-subjects design, a within-subjects design, or both; and (b) think of examples of what these might look like in Max's hair colour and friendliness study (you might need to imagine both between- and within-subjects designs).

So, with these varied threats to internal validity in mind, how can we reduce them?

Using a between-subjects design will often reduce some threats to internal validity (e.g., maturation). Randomly assigning participants to groups is essential and will further eliminate some threats (e.g., selection differences). There is an important caveat about random assignment, however. Random assignment simply helps reduce the likelihood of there being systematic differences in extraneous variables between the two groups. Remember that randomization does not guarantee equivalent groups.

Demand characteristics are usually reduced in a between-subjects design. However, they may still be present. Thus, our comparison group ideally has a "placebo" of some kind. For example, if you want to test the effects of alcohol consumption on reading speed, participants in the placebo group would receive a drink that looks and tastes like the alcoholic drink, but without the critical ingredient (alcohol). At the very least, all participants should be *blind* to the condition they are in.

What if you are using a within-subjects design for other reasons (e.g., you may know that there are strong influence of participant variables on the dependent variable, so you wish to control for them by using a within-subjects design)? Then you will need to find ways to minimize practice effects, boredom or fatigue, and demand characteristics! Ideally, you would **counterbalance** the order in which participants take part in the different conditions. Note that participants should be randomly assigned to the order. For example, if using a within-

subjects design, Max should have half of the sample rate the red-haired person first and the brown-haired person second, and vice versa for the other half of the sample.

Reducing the Influence of Nuisance Variables

Unsystematic variation of extraneous variables (nuisance variables) is present in both between- and within-subjects designs. However, for the most part, the unsystematic variation is reduced in within-subjects designs because we have the same people participate in both conditions (**participant variables** like age, IQ, personality are held constant). So this reduces the variability due to nuisance variables (and also reduces the chance that they might become confounds).

Another way to hold these participant variables constant is to do something called **matching**. In this case, you pair up participants by the extraneous variables you are concerned about, and then randomly assign one person from each pair to either the experimental or control group using a between-subjects design. For example, people's pre-existing anxiety levels might influence how well they respond to an intervention to reduce test anxiety. You might measure test anxiety in all participants prior to randomly assigning them to an intervention or control group. Then, you would match them in pairs according to their test anxiety scores and randomly assign each participant from within each pair to one of the two conditions.

There are also some things you can do, regardless of whether you have a within- or between-subjects design, to reduce the effect of nuisance variables. First, it is important to have **clear scoring criteria**. We should know when a response is correct or incorrect, how to distinguish responses, and where a response begins and ends. This is easier in some situations than others. Consider, if you want to measure children's aggressive behaviours in the playground, how would you score behaviours? What behaviours would you describe as aggressive: nudging, yelling, pushing, hitting, or kicking? You would need to develop some very clear scoring criteria so that the behaviours could be quantified and recorded. You might want to video the children playing so that multiple researchers (raters) can watch the recordings and code the responses. You could then look for agreement amongst the different raters. Second, we can use a **sensitive** dependent variable. Think back to levels of measurement. For example, if you ask people "how much do you like raw oysters on a scale from 1 – 10?" this is more sensitive than asking "do you like raw oysters: yes or no?" With a more sensitive measure, we can get a range of responses and detect smaller changes across levels of the independent variable. Third, we can address **reliability** of behaviours: on any given occasion, different extraneous variables might influence behaviour. We may have to test several times. For example, in Max's study to assess perceived friendliness of red- and brown-haired people, features of the target faces might elicit idiosyncratic preferences for different participants. Max might want to consider presenting multiple different faces, with brown or red hair, for participants to rate.

If we are using questionnaires or scales, we need to maximize reliability of these measures too. Instructions should be clear, and we should use items that are not going to elicit idiosyncratic responses. Multiple items should be used to assess the same construct; and items should be worded carefully to be unambiguous (e.g., no double-barrelled items).

Practice trials can be help for some tasks, too. Finally, we should aim for consistency in the experimental procedures. This can be achieved by some of the following: (a) having a detailed experimental protocol (e.g., typed-out instructions that are clear, with no jargon, and the same every time [no ad libbing!]); (b) using a double-blind procedure (neither the participant nor the experimenter interacting with the participant knows what condition the participant is in); (c) automatising the experiment (e.g., using a computerized task if possible); (d) ensuring that any confederates are blind to the testing condition; (e) testing participants in groups so that they all get the same instructions (but beware if behaviour might change in a group setting); and (f) pilot testing the instructions to ensure that they are clear.

Connection to Statistics

Why does all this matter? Why should we try to maximize the effect of the independent variable on the dependent variable, while minimizing confounds and the influence of extraneous variables? Here is a really cool thing: *statistical tests* look at the variation in scores on the dependent variable, and compare how much variation in the dependent variable is systematic versus how much is unsystematic. It goes something like this:

$$Test\ statistic = \frac{Systematic\ variation}{Unsystematic\ variation}$$

If you have confounds, the variation due to the independent variable and the variation due to the confounds are lumped together (because they are both systematic) so you cannot tell them apart! In that case, internal validity is compromised, and you will not know what caused the changes in the dependent variable! However, if you can rule out potential confounds—by being very careful in the design of your study—you are now left with just variation due to the manipulation of the independent variable and any unsystematic variation (variation due to nuisance variables).

Statistical tests compare how much variation in the dependent variable is systematic versus how much is unsystematic. You should be aiming to get a large amount of systematic variation (due to the manipulation of the independent variable)—i.e., the radio signal—relative to the amount of unsystematic variation—the static. This will allow you to obtain a large value for the test statistic and hence detect the experimental effect.

External Validity

Finally, as noted earlier in this chapter, external validity is the degree to which the results accurately generalize to other individuals and situations. A special type of external validity is ecological validity, which is the extent to which research can be generalized to natural situations. For example, will the effects of red or brown hair on perceived friendliness extend to real world settings?

It is important to note that there is often a trade-off between internal validity and external validity. We can increase internal validity through greater control, but this produces a more unnatural situation and, therefore, it lowers external validity! Generally, lab experiments have more internal validity, but field experiments have greater external validity. For some kinds of research (e.g., fundamental research to understand basic perceptual processes), you need lots of control, so you need to do this in the lab. For other kinds of research, it might be more important to focus on external validity. For example, I study emotion regulation. Lots of research either has people come into the lab, where we ask them to regulate their emotions in different ways and observe the effects on their behaviour, or we give them questionnaires where they respond by saying how they would regulate their emotions. None of this really matters if, in the former case, the strategies we ask them to use in the lab are not the ones they would use in the real world anyway and, in the latter case, if they do not do what they say they would do!

Given that external validity is a limitation in a lot of psychological studies, researchers have to be very careful how they write their conclusions to ensure that they do not extend beyond the participants and settings tested in the study.

CHAPTER 2: STATISTICS

This chapter is largely an excerpt from “Statistics with jamovi” by Dana Wanzer (Creative Commons BY-SA (CC BY-SA) license version 4.0), with some minor changes and some additions (addition of the equation for the mean and for sum of squares, description of what measures of variation represent, an explanation of sampling distributions, interpretation of *p*-values, confidence intervals, and parametric vs. non-parametric tests, and other minor changes).

Descriptive vs. Inferential Statistics

There are basically two different types of statistics:

1. **Descriptive statistics** are used to summarize, organize, and overall *describe* our sample data. Typically, we do so using measures of central tendency (e.g., mean, median, mode), measures of dispersion (e.g., range, standard deviation, variance), and shape (e.g., skew, kurtosis). We may also visualize the data using tables or graphs.
2. **Inferential statistics** are what we use when we collect data about a sample and see how well that sample *infers* things about the population from which the sample comes from. Typically, we do so with statistical tests like the *t*-test, ANOVA, correlation, chi-square, regression, and more.

We can visualize the relationship between the population, sample, descriptive statistics, and inferential statistics (see figure below). We are typically interested in a **population** of interest but may not be able to collect data from the entire population because of budget, time, access, or other constraints. We therefore **sample** from the population; ideally, we do so randomly, but there are other types of sampling methods available. We then use **descriptive statistics** to describe our sample data and **inferential statistics** to make generalizations about the population from which they were selected.

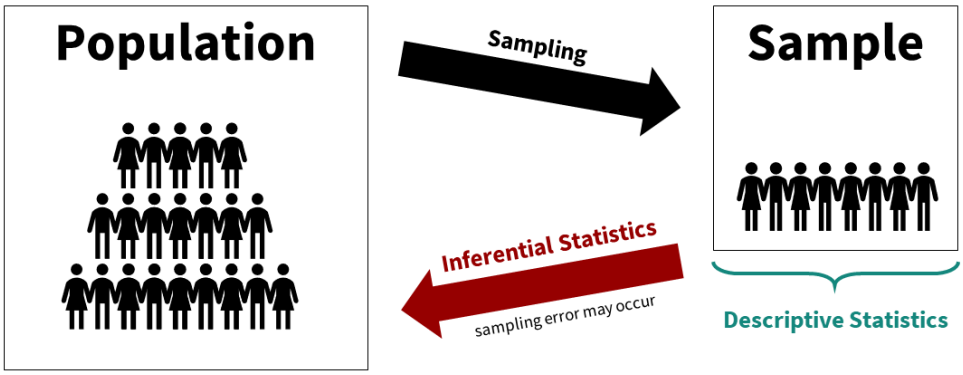


Figure 2.1. *Population Samples: Inferential and Descriptive Statistics. (Wanzer, 2022)*

An Example of Using Descriptive and Inferential Statistics in a Study

This has been pretty abstract so far. Let's go through a fairly simple research study to walk through all of this.

Imagine we're conducting an experimental study examining whether watching *Community* –a very good show–versus watching video lessons on studying techniques[1] –useful, but maybe a bit boring–improved test performance in TRU students.

1. Interested in better techniques for studying? Check out The Learning Scientists.

Our population of interest is therefore all TRU students, several thousand students. We cannot include them all in our study; it wouldn't be feasible for us to collect all that data and probably not possible to get the university to get on board with the study of the entire student body. Therefore, we smartly decide to only collect data from a sample of the student body.

Who might our sample be? Ideally, we'd gather a random sample of students. However, to do that we'd likely need to still get university approval and get a list of a portion of student emails for recruitment purposes (oversampling because our response rate is unlikely to be 100%). I just want to do this study to show what descriptive and inferential statistics are, so I just use students in two different sections of introduction to psychology classes (around 80 students total) as my population. This is definitely not a random sample, but a fine study for our illustrative purposes.

We conduct our study–let's assume we're fabulous researchers and it worked out perfectly. We randomly assign half our students to watch *Community* as part of their studying, and the other half watch video lessons on studying techniques. They have an exam a week later and we measure their accuracy on that exam. We then want to know: which group performed better on the exam?

First, let's describe the sample. We would likely visualize our results, perhaps as a histogram of all test scores, maybe separated by which group they were in. This would help us look at whether our data is normally distributed (more on this in a subsequent chapter on assumption checking). We would get the descriptive statistics: probably the mean, maybe the median if our data is skewed, the standard deviation and variance, and the range. These would provide a summary of our data. If we wrote up our results and didn't share a visualization, this information would give a good sense of our data to our readers.

But what we really want to know is: which group performed better on the test? For that, we need to use inferential statistics. In this particular example, we would use our mean, standard deviations, and sample sizes for both groups. We then plug the numbers into the equation for this particular inferential statistic (in this case, an independent samples *t*-test, more on that later) or–even better–we perform the statistic in our statistical software (jamovi). It spits out our statistical value, confidence intervals, effect size, and our *p*-value and we can then infer what the results mean for our population and answer our research question. Thus, the inferential test allows us to say something about what we think is going on in the real world, based on the statistical model (in this case the *t*-test) that we build using our data.

Central Tendency, Dispersion, and Shape

Central Tendency

There are multiple **measures of central tendency** (these are *all* averages so you must be careful when you say that word to explain which type you mean!):

- **Mode:** the most frequent score; the only measure suitable for nominal data; not stable across samples-subject to substantial sampling fluctuation; ignores much information in the data set
    - **Multimodal** or **bimodal**: when two or more values are the most frequent score
- **Median**: the middlemost value; less susceptible to outliers and best used when interval/ratio data are skewed; used for ordinal data; moderately stable across samples; loses information (only reflects frequency of scores in the lower half of the distribution)
- **Mean**: the sum of all points divided by the total number of points; susceptible to outliers; stable across samples; see below, where x_i represents an individual score in the dataset, Σ represents sum, and n is the number of scores in the dataset.

The mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

This formula can be a bit overwhelming, at first glance! What you need to know is that this symbol Σ (sigma) means "sum of." And

$$\sum_{i=1}^{n}$$

means: add these things together, from $i = 1$ (the first participant) to n = the last participant in the dataset.

Dispersion

The mean on its own does not tell us very much. Perhaps everyone scored the mean, or perhaps there was lots of variability in scores around the mean. Therefore, we also look at something called **dispersion**, which relates to how much our scores are spread out in our dataset. There are multiple measures of dispersion.

- **Range**: the difference between the maximum and minimum value (e.g., if the minimum score is 17 and the maximum is 49, then the range is 32)
- **Quartile**: when a dataset is divided into four equal parts, the first quartile (Q1) is at the 25th percentile, the second quartile (Q2) is at the 50th percentile, and the third quartile (Q3) is at the 75th percentile.
    - **Interquartile range**: the middle 50% (Q1 to Q3)
- **Variance**: the sum of the squared deviations from the mean. This means first (a) calculating the mean, (b) subtracting each score from the mean (aka deviations from the mean), (c) squaring each of those deviations values, and (d) summing all those squared deviations. This is represented by the equation:

$$s^2 = \frac{\sum(x - \mu)^2}{N}$$

- **Standard deviation**: is the square root of the variance. This is represented by the equation:

$$s = \sqrt{\frac{\sum(x - \mu)^2}{N}}$$

    - Note that these two equations are only used if we are examining the whole population. If we only have a sample, we replace the denominator *N* with *N*-1.
- You should also be aware of the **sum of squared errors (SS)**. This reflects the total spread of scores around the mean:

$$SS = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Note that we do not use SS as a measure of descriptive statistics because the size of SS depends on the number of scores in the dataset. If we take an average of SS, we get variance. Variance is not that useful as a descriptive statistic because it is measured in units squared. Therefore, we usually report the standard deviation (square root of the variance). The standard deviation can be used to compare variation across different sized samples because it represents a "standardized" amount of deviation from the mean.
- However, we do use SS in inferential statistics, so you will see it again later!
- Finally, SS, variance, and standard deviation all tell us:
    - How well the mean "fits" the data – smaller values indicate better fit
    - Variability of the data
    - How well the mean represents the observed data
    - How much "error" there is

Shape

Finally, there are two main **measures of shape** that describe the shape of the distribution of our data:

- **Skew**: in a non-normal distribution, it is when one tail of the distribution is longer than another, producing an asymmetric distribution.
    - **Negative skew**: when the tail points to the negative end of the spectrum; in other words, most of the values are on the right side of the distribution
    - **Positive skew**: when the tail points to the positive end of the spectrum; in other words, most of the values are on the left side of the distribution
- **Kurtosis**: the weight of the tails relative to a normal distribution.
    - **Leptokurtic**: light tails; values are more concentrated around the mean
    - **Platykurtic**: heavy tails; values are less concentrated around the mean

There are other terms we use to describe data:

- **Frequency distribution**: overview of the times each value occurs in a dataset; often portrayed visually like with a histogram
- **Histogram**: a visual depiction of the frequency distribution using bars to depict a range of the distribution
- **Normal distribution**: a special distribution in which the data are symmetrical on both sides of the mean; under a normal distribution, the mean is also equal to the median

Inferential Statistics 101

There are some fundamental principles that all the test statistics (*t*, *r*, *F*, etc.) draw on. Let's go back to Max's research question from the first chapter: Do redheads appear more friendly than other people? If we had infinite capacity for testing participants, we could gather up all the people in the world and present them all with photos of redheads and people with other coloured hair and ask them to rate the friendliness of each photo. Of course, we cannot do this, so we just test a sample of participants from the population and use those data to make inferences about the population. Hence, we use inferential statistics, to *infer* the characteristics of the population from our sample.

Null Hypothesis Significance Testing

How do we do that? We are going to use **null hypothesis significance testing (NHST)**. First, Max sets up his **null** and **alternative hypotheses**. Assuming there is some prior research and theory to support Max's idea, we might suggest:

Alternative hypothesis (H_1): Participants will rate redheads as more friendly than non-readheads.

Null hypothesis (H_0): There will be no difference in ratings of friendliness for redheads and non-redheadeads OR participants will rate redheads as less friendly than non-readheads.

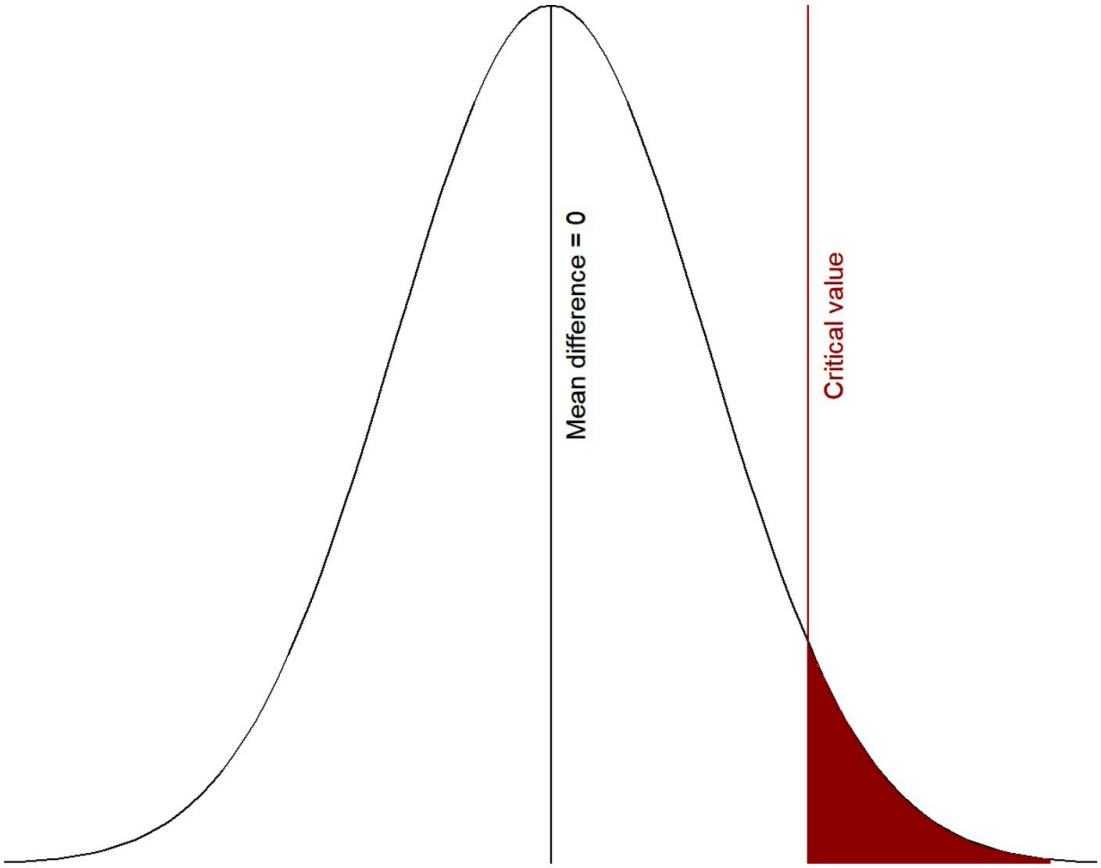
Note that I have set this up as **one-tailed** or **directional hypotheses**. If I had **two-tailed** or **non-directional hypotheses**, what might they be?[1] (Note that, either way, I have to ensure that all possible results are captured.)

Now, Max goes out and collects some data with 30 participants (N = 30) and finds that the participants do indeed rate the redheads as more friendly than the non-redheads. However, it's important to realize that the difference in ratings in our sample is unlikely to be exactly zero, even if the null is true. What we need to find out is how big the difference in ratings should be for us to reject the null hypothesis. How do we do that?

From Sampling Distributions to *p*-Values

First, we generate a **sampling distribution**: for this, we start by assuming that the null hypothesis is true, and that there is no difference between the friendliness ratings for redheads and nonredheads (mean difference = 0). We imagine going into the population an infinite number of times, collecting an infinite number of random samples the same size as our sample (N = 30 in this case), calculating the mean difference for each sample, and plotting the frequency of those means differences. We might get something like this:

1. Alternative hypothesis: participants will rate redheads differently from non-redheads. Null hypothesis: there will be no difference in ratings of friendliness for redheads and non-redheads.



As you can see from the figure, we would get many samples where the mean difference = 0 (high frequency), quite a few samples where the mean difference is a bit above or below 0, and far fewer samples where the mean difference is much larger or smaller than 0.

Next, we decide on our critical value, shown on the figure as a red line. This is determined by our alpha level, which we have set as our level of significance. Most studies you read use the arbitrary $\alpha = .05$ (5%). In the visualization above, we set the alpha to 5% and so the area shaded in red is exactly 5% of the area under the curve of the normal distribution.

Our critical value is the level at which we are saying we would consider it "surprising" (versus "not surprising") if we got a mean difference that large or larger (i.e., the red region in the graph) *assuming that the null hypothesis is true*. Basically, if we assume there truly is a mean difference of 0 in the population (i.e., the null hypothesis), values beyond the critical value would be considered surprising enough that we would say that we reject the null hypothesis. This is why it is called *null hypothesis significance testing*.

In other words, *the area in red are values that are unlikely to occur if the null hypothesis (in this case, mean difference <= 0) were true*. In fact, we would get a mean difference in the red region 5% of the time, in the long run, if the null hypothesis were true.

A bit more about sampling distributions

There are many different kinds of sampling distributions. Every statistic that we use has sampling distributions associated with it: there are sampling distributions for *t*, *r*, *F*, and so on, and in each case they vary in shape according to the degrees of freedom.

Also, you should be aware that we do not actually go out and create the sampling distribution. It is a theoretical distribution. In addition, we do not have to calculate the critical value ourselves, thanks to the work of some fancy mathematicians in the days before computers! Based on the incredible properties of the normal distribution, they figured out what the sampling distributions of all the test statistics we need to use would look like. They determined how the shape of those distributions would change according to sample size (or, to be more precise, degrees of freedom). And they calculated the critical values for each test statistic depending on the direction of the hypothesis and alpha level. In your lower level stats class, you probably looked these things up in a table in a book, but in this class, our statistical software will generate the information we need.

When you run your inferential statistics using statistical software (jamovi, SPSS, or whatever software you plan to use), you will obtain the actual value for the test statistic in your sample and a *p*-value. This ***p*-value** represents the probability or the likelihood of getting that value for the test statistic or more extreme, in the long run, when the null hypothesis is true. If $p < .05$, we figure we are pretty unlikely to get a value that extreme or more extreme when the null is true, and so we reject the null hypothesis.

So to sum up:

1. We generate a sampling distribution for our test statistic;
2. We calculate the test statistic for our particular sample; and
3. We find out how likely we are to get that value or a more extreme value for that test statistic, when the null hypothesis is true, in the long run, i.e., the *p*-value.

Type 1 and Type 2 errors

There is one more thing to keep in mind: we can be wrong! Just because we get a result does not automatically mean that result is 100% accurate. There are many things that could lead us to an inaccurate interpretation!

We can use the table below when discussing errors. On the far left column, we have our results: were they statistically significant ($p < .05$) or not ($p > .05$)? On the top row, we have whether *in the real world* the null or alternative hypothesis is true. In reality, we can *never* truly know whether the null or alternative hypothesis is true. We can at best approximate our understanding of the real world through replication!

	H_0 is true	**H_1 is true**
***p* < .05** (statistically significant)	Type 1 error (alpha)	Correct interpretation
***p* > .05** (statistically non-significant)	Correct interpretation	Type 2 error (beta)

Therefore, any time we get a statistically significant result ($p < .05$), then *either* we made a correct interpretation *or we made a Type 1 error*!

Similarly, any time we get a statistically non-significant result ($p > .05$) then *either* we made a correct interpretation *or we made a Type 2 error*!

What does a p-value tell you?

The *p*-value does *not* tell you:

1. The importance of an effect
2. When the null hypothesis is true
3. When the null hypothesis is false

It *does* tell you: the probability (in the long run) of getting a test statistic this extreme, or more extreme, if the null hypothesis were true.

Confidence Intervals

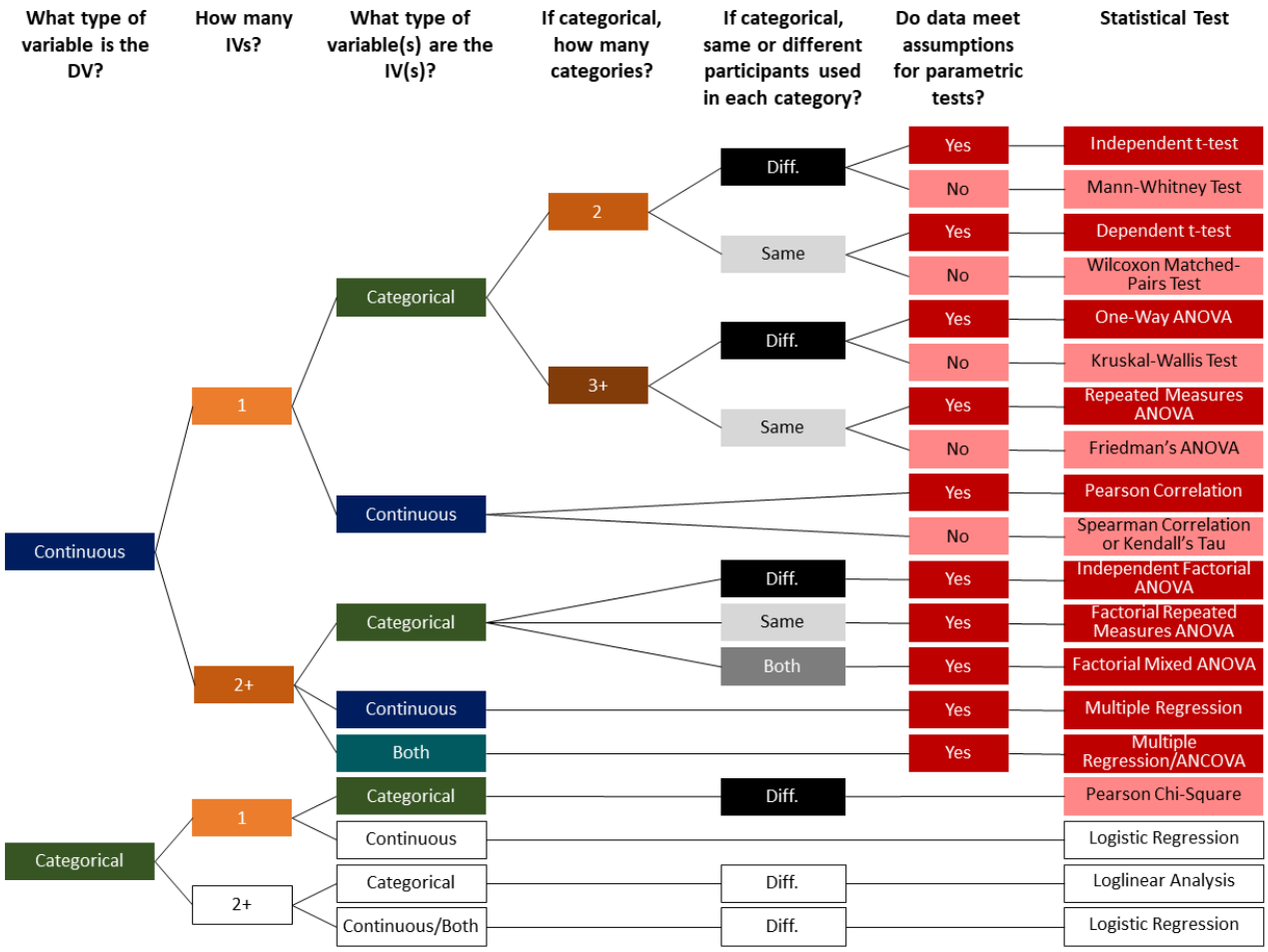
Again, when we use inferential statistics, we are aiming to infer something about the population from our sample. In a very simple model, we can use our sample mean as an estimate of the population mean. This is called **point estimation**. However, it may not be very accurate. We might be better off estimating the population mean by giving a range of values within which the population mean might fall. To do this, we use **confidence intervals (interval estimation)**. Confidence intervals (CIs) are a statistical way of saying "around." Specifically, a 95% CI for a mean tells us that 95% of the time, in the long run, the CI will contain the true value of the population mean. We can compute confidence intervals around not just means, but also around our statistical test values (e.g., *t*, *r*, etc.) and around effect sizes.

Parametric and Non-Parametric Tests

As you learn more about statistics (and later in this book) you will notice that there is a distinction between parametric and non-parametric tests. These are both kinds of inferential statistics. However, parametric tests make certain assumptions about the data (e.g., that the variances for different groups in the experiment should be homogenous) and so we cannot use them in all situations. On the other hand, non-parametric tests make fewer assumptions and so can be useful when our data violate the assumptions of a parametric test. More about assumptions and the non-parametric alternatives to parametric tests as we proceed through the books.

Choosing the Correct Statistical Test

It is important that you learn how to identify *which* inferential statistic you should perform. This chart can help you determine what statistical test to perform. Note that on the right dark red boxes are parametric tests, light red boxes are non-parametric tests, and white boxes will not be covered in this class at all (in fact, there are many others not even shown that we won't cover!). Data types are indicated in either blue (continuous), green (categorical), or teal (both). Number of variables or levels of the variables are either 1 (light orange), 2/2+ (orange), or 3+ (dark orange). Between-subjects designs, meaning designs with different participants in each group, are in black whereas within-subjects designs, meaning designs with the same participants in each group, are in light grey.



First, you need to determine what level of measurement your dependent variable (DV) is. We will only be covering statistical tests that have *one* dependent variable. Therefore, you need to know whether the variable is categorical (i.e., nominal or ordinal) or it's continuous (i.e., interval or ratio).

Next, you need to determine how many independent variables (IVs) there are and then what level of measurement your IV(s) are. In the case of a single categorical IV, we also need to know how many levels there are to the IV (i.e., how many categories there are). For categorical variables, we also need to know if the participants are different (i.e., between-subjects) or the same (i.e., within-subjects) within each level of the category.

Lastly, for many of the statistical tests we need to know whether we meet the assumptions of parametric tests. If we don't meet the assumption, then there are alternative tests we can perform.

We can both *forward map* and *backwards map* with the chart above. Forward mapping involves understanding your data and your research question and then determining what statistical test to perform. Forward mapping is mostly what you need to understand how to do! Backwards mapping involves determining what kind of data is needed to perform a particular statistical test. This is more for educational and understanding purposes and generally is not how you analyze data.

Forward Mapping: Choose the Correct Test

A researcher is interested in understanding whether athletes have higher English scores than non-athletes. In other words, what is the effect of athletic status on English test scores?

1. What is the DV? What is the level of measurement? It's English test scores, which is a continuous variable.
2. How many IVs are there? We only have one IV, and it is athletic status.
3. What is the level of measurement of the IV? Athletic status is a categorical variable.
4. How many categories to the IV? Athletic status is measured as either athlete or non-athlete, so there are 2 levels.
5. Are the same or different participants used in each category? People can either be an athlete or not an athlete, but they can't be both, so this is a between-subjects variable (aka "different").
6. Do data meet the assumptions for parametric tests? We don't know. We would need to test this. For now, let's assume we meet the assumptions.
7. Statistical test? Independent *t*-test

A researcher is interested to know whether people perform better on the exam at the start, middle, or end of the semester. The researchers has all participants complete all three exams.

1. What is the DV? What is the level of measurement? In this case, the exam is our DV and it's a continuous variable.
2. How many IVs are there? We only have one IV, and it is time of the exam.
3. What is the level of measurement of the IV? The time of the exam is a categorical variable.
4. How many categories to the IV? Type of test has three categories: start, middle, or end of the semester.
5. Are the same or different participants used in each category? Although the researcher could have designed a between-subjects design, this particular study has all participants participate in all conditions, so it is a within-subjects design (aka "same").
6. Do data meet the assumptions for parametric tests? We don't know. We would need to test this. For now, let's assume we meet the assumptions.
7. Statistical test? One-way repeated measures ANOVA

In Practice: Statistics

Now that we have reviewed the core concepts underlying descriptive and inferential statistics, how do we put this into practice? Dana Wanzer (2022) outlines four practical steps to hypothesis testing:

1. Look at the data: examine the descriptive statistics and describe your hypotheses;
2. Check assumptions: ensure your data are satisfactory for performing the inferential test that you hope to use (or select an alternative statistic);
3. Perform the test: run the inferential statistic; and
4. Interpret the results: make a decision about whether you reject or fail to reject the null hypothesis and write up the results in APA format.

We shall come back to these four steps each time we learn a new statistical test. In the next chapter, we shall get started with some of the basics of using jamovi to organize and look at data.

CHAPTER 3: INTRO TO JAMOVI

This chapter is based on the work of “Statistics with jamovi” by Dana Wanzer (Creative Commons BY-SA (CC BY-SA) license version 4.0), with some minor changes and additions.

Overview of jamovi

jamovi is a free and open statistical software that helps us run our descriptive and inferential statistics. Why are we using jamovi and not another program?

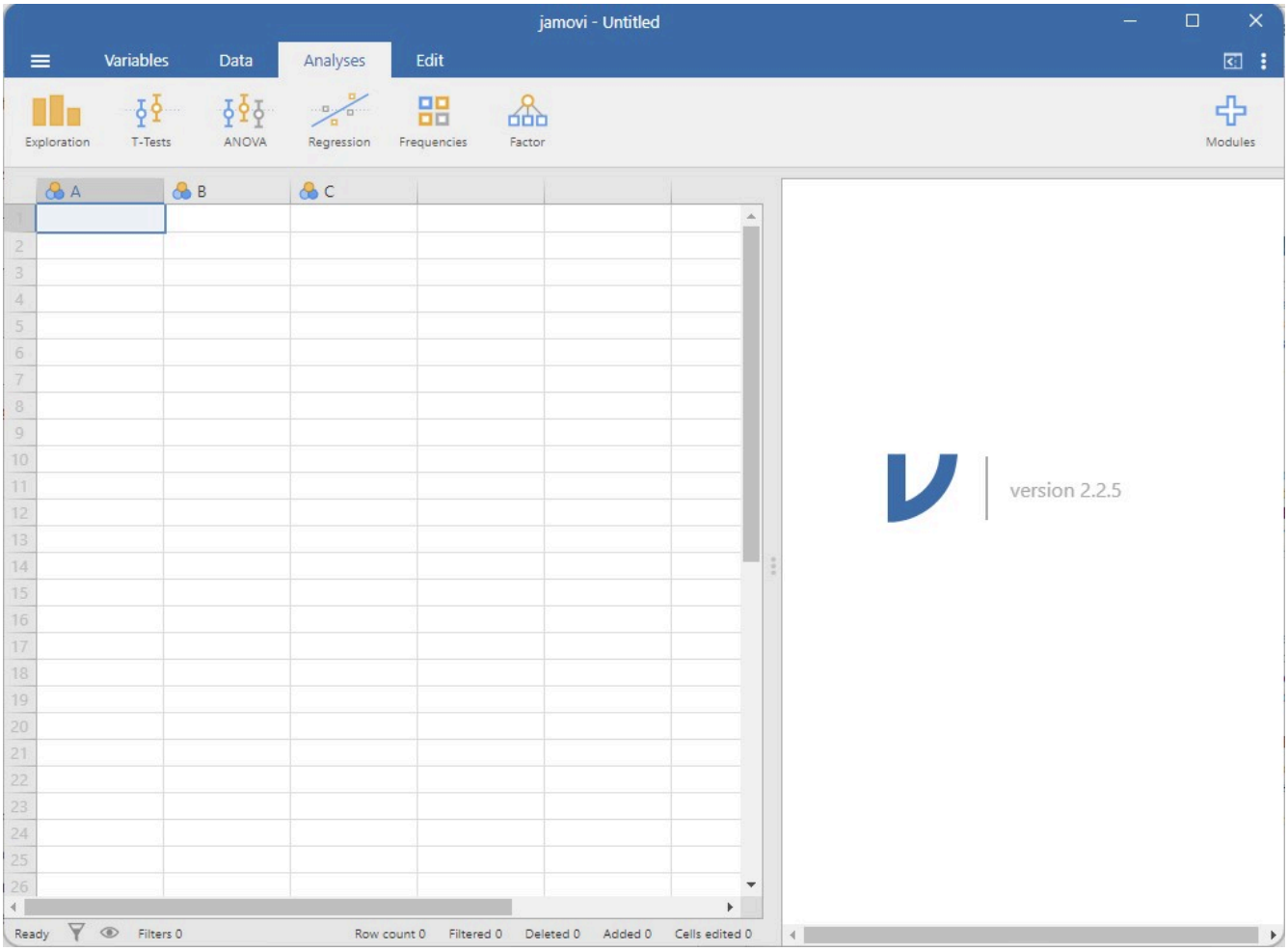
1. Did I mention it's free? You can download it for free on your home computer or laptop and you won't ever have to pay a dime to use the software in the future.
2. It's open source, meaning that the statistical community helps support and improve the program. As jamovi says, "jamovi is made by the scientific community, for the scientific community." New packages with new options for analyses, graph creation, and so on, are being added all the time.
3. It's built on top of the R statistical language, meaning you can begin learning how to code (if you want). I do a lot of my statistical analyses using R in a different program called RStudio (actually Dana Wanzer's book, which I am drawing heavily from for this book, was developed in RStudio and hosted on GitHub!). R is a very powerful tool which is also free and open source.
4. It's incredibly easy to learn and use. I have taught statistics using both SPSS and jamovi, and students greatly prefer jamovi.
5. It promotes reproducibility. jamovi will save your data, analyses, options, and results all in one file so you can easily pick up where you left off. This will make your homework and future data analyses a breeze.

There are some videos that walk you through working with jamovi, available at datalab.cc/jamovi, if you like videos for walking you through how to do things. I have found in the past that students in this class prefer step-by-step instructions "on paper" because these are easier to refer back to in future, so I shall provide step-by-step instructions for jamovi throughout this book.

Getting Started with jamovi

Install and Open jamovi

First, you need to install jamovi, if it's not already on the computer you are working on. To do this, go to https://www.jamovi.org/ and download the correct version for your computer (make sure you get the "solid" version, which is recommended for most users). Once it is installed, open the program to see what it looks like. You will see something like this:



To the left is the spreadsheet view, and to the right is where the results of statistical tests will appear. Down the middle is a bar separating these two regions and this can be dragged to the left or the right to change their sizes.

In jamovi data are represented in a spreadsheet with each column representing a "variable" and each row representing a "case" or "participant."

It is possible to simply begin typing values into the jamovi spreadsheet as you would in any other spreadsheet software. Alternatively, existing data sets in Excel or the CSV (.csv) file format can be opened in jamovi. Simply click on the "hamburger" ( ) in the top left corner of the jamovi window, click on "Import," and find the file that you want to import. You can also easily import SPSS, SAS, Stata and JASP files directly into jamovi. To open

an existing jamovi file (with the .omv file extension) select the hamburger, select “Open” and then choose the appropriate file.

Entering and Cleaning Your Data

There are four important steps to entering and cleaning your data: checking your data are set up correctly; computing new variables; transforming variables; and using filters.

1. Checking your data are set up correctly

When you enter your data in jamovi, you will typically (for now at least) enter it with one row for each participant. Always create a variable to identify each participant by. For example, each participant in the study might be given a number. The same number would appear on all parts of the study they complete, as well as in jamovi. You can call this variable “Participant” in jamovi, and set it to be an ID variable (see below). If you are collecting data online using a platform like SurveyMonkey or gorilla, those platforms might automatically assign participant IDs to your participants. In those cases, you may as well continue to use the IDs supplied by those platforms, for consistency and so that you can keep track of who is who.

Variables in jamovi

It’s important to understand the different types of variables in jamovi and how they map onto our levels of measurement.

Variables in jamovi can be one of three data types:

1. Integer, meaning the values are discrete whole numbers
2. Decimal, meaning the values are numbers with decimals
3. Text, meaning the values are alphanumeric, not just numeric

Furthermore, variables in jamovi can be one of four measure types:

1.  Nominal
2.  Ordinal
3.  Continuous (meaning jamovi combines interval and ratio and doesn’t distinguish between the two)
4.  ID (used for any identifying variable you likely wouldn’t ever analyze, like participant ID number or name)

There are a few great things about jamovi when it comes to these data variables. First, jamovi will try to

automatically determine what the data and measure types are when you type in data or when you open a dataset; this is fabulous, until it goes wrong. It's important that you always double check your data and measure types first!

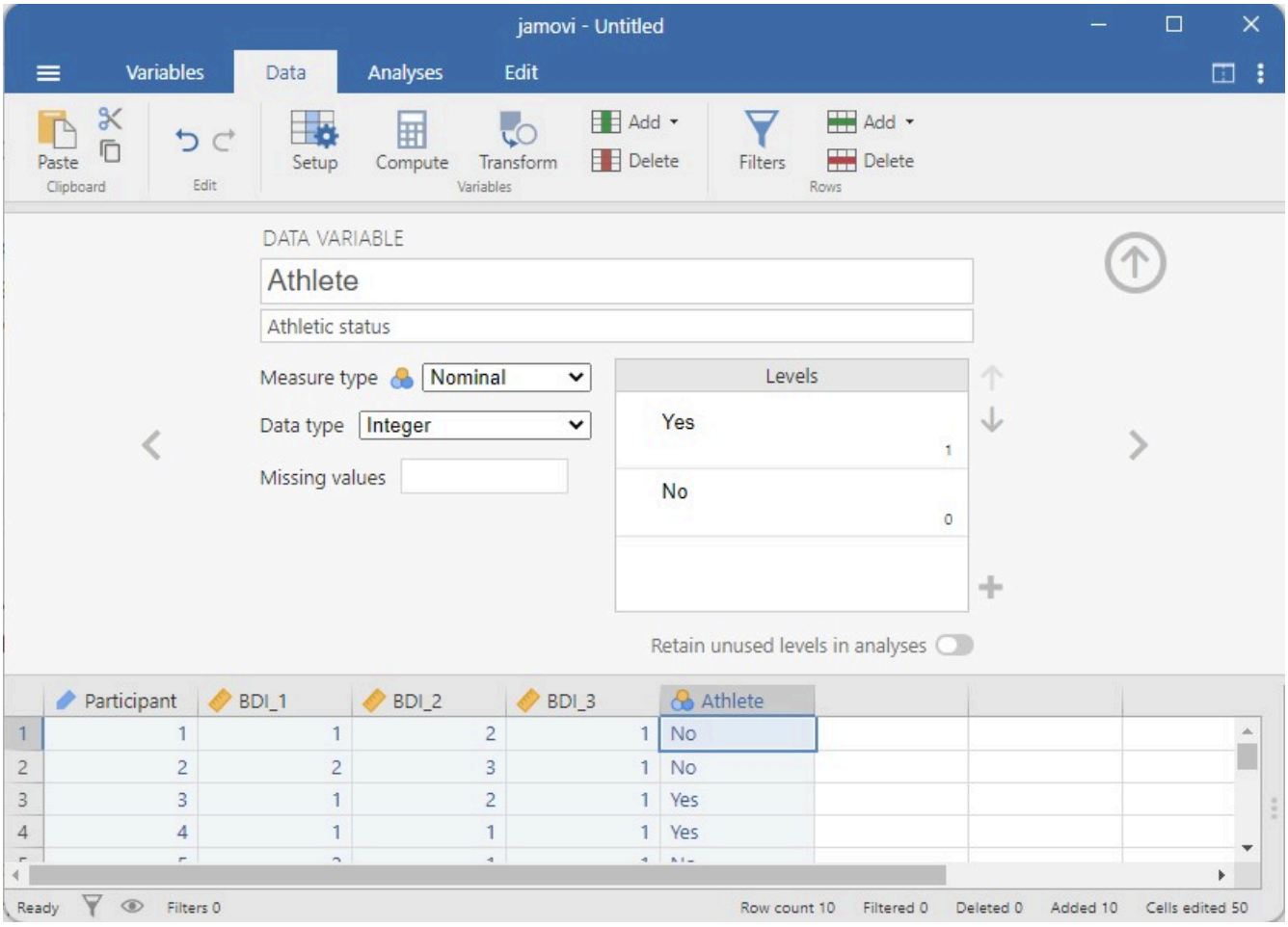
Second, those little icons will be really helpful to let you know what variables can go in which boxes. For example, we would never analyze a nominal variable as our dependent variable for a *t*-test, and jamovi will help remind you of that. When performing an independent samples *t*-test, the dependent variables box will have a little ruler icon indicating you should be putting continuous variables in that box. Similarly, it will tell you to put nominal or ordinal variables in the grouping variable (independent variable) box.\

It's really important to check that the data types and measurement types of your variables are correct. You should open the Setup () option under the Data tab to check.

When you're in Setup, here are the things you should be doing for all variables:

1. Make sure the variable name is meaningful to you and will appear nicely in your data visualizations or tables (e.g., don't write `Q1` but rather `BDI_1` for the variable that is participants' responses to the first item on the Beck Depression Inventory).
2. Add a description to your variable so you have more context. Maybe you compute a mean of all the BDI items and in the description you write `Mean of all BDI items` for the description of your `BDI_Mean` variable.
3. Check your measure and data types are correct.
4. Specify if there is a code for missing values. Make sure the code *does not* match the code you use for actual variables! For example, if I have a variable that ranges from 0-10, then I wouldn't use 9 as a code for missing values; instead, I might use 99 or -9.
5. Add labels to levels. For example, the variable `Athlete` is 0 for non-athlete and 1 for athlete. Rather than keeping just the 0 and 1, you can specify under Levels that 0 is non-athlete.

So, you might have something that looks like this:



You will see that for setting up the "Athlete" variable, I simply typed "Athlete" into the first box, "Athletic status" into the second box (where you can write a longer description). I selected "Nominal" for Measure type. Initially, the values that I imported into jamovi were 0's and 1's, so I translated this information for jamovi by typing "Yes" and "No" respectively into the boxes under "Levels."

2. Computing Variables

Sometimes you need to create new variables from your raw (meaning uncleaned) data. Perhaps you collected data on a scale that has five items. Normally, we create an average score of all the five items and that new *computed* average score is what we use in our analyses.

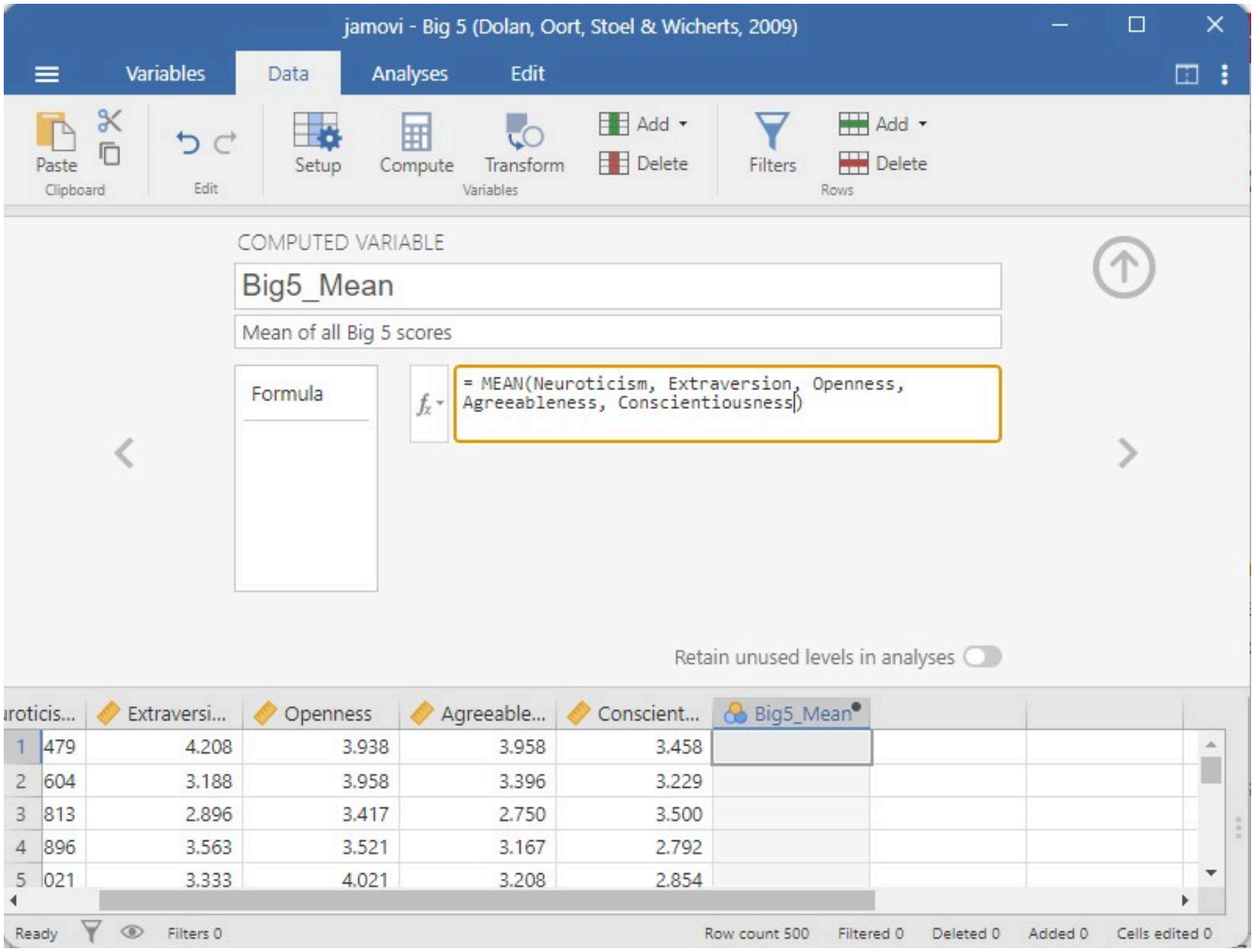
Let's open the Big 5 dataset built into jamovi. You can open this dataset by clicking the three horizontal lines on the top left of jamovi (the menu), choose Open, then select Data Library. In the main Data Library folder is a dataset called Big 5.

This dataset has the scores on all five subscales of the Big Five personality test. Let's imagine we want the average score of the entire Big Five test. To do this, first put your cursor in the column where you would like to create the new variable. Then, click on the Data tab and choose `Compute`. You may need to double-click on the column to show the computation window above your data. Rename the computed variable (e.g., `Big5_Mean`), add in a description, and then create the formula.

In this case, we need to select the function `MEAN`. To do so, click on the f_x and scroll down below Functions until you see `MEAN`. When you click on `MEAN`, it provides a template of what the formula should look like. We need to

specify the function `MEAN()`, add all the variables we want to calculate in the mean (i.e., the five subscales of the Big 5), and there are two alternative options: `ignore_missing` is defaulted to 0 (meaning DON'T ignore missing, or rather include missing) and `min_valid` is defaulted to 0 (meaning it's ignoring this; perhaps you only want to include people that have at least three valid cases).

The basic formula, then is to do MEAN(var1, var2, ... varn). You can see what we need to do with this dataset below. There's actually no missing data, so the two additional arguments aren't necessary for us to worry about.



Note that you must type the variable names *exactly* as they appear in your data area – if you mistype them, you will get an error and your mean will not be computed!

If you'd like to learn more about computed variables in jamovi, check out this jamovi blog post on the topic.

3. Transforming Data

Sometimes we want to take an existing variable and transform it in some way or we want to do a computation across multiple variables (e.g., reverse-score multiple items in a dataset). If you want to learn more about transforming variables, the jamovi blog has a great blog post on the topic.

Reverse-scoring

Sometimes items need to be reverse-scored because they are in the opposite direction of the entire scale.

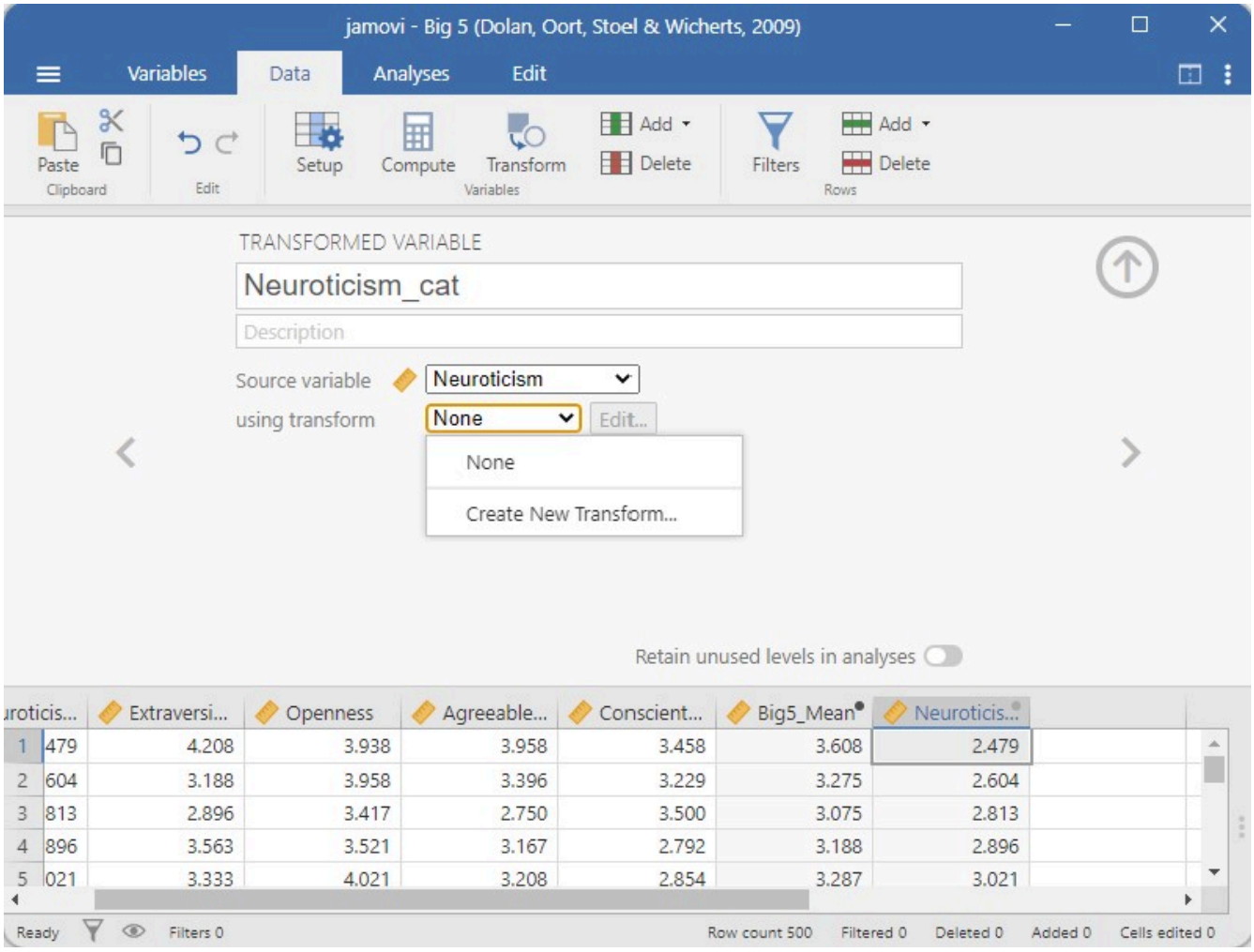
Let's imagine we have a Happiness Scale with the following four items:

1. I am happy.
2. I am content.
3. Life is overall positive.
4. I am unhappy.

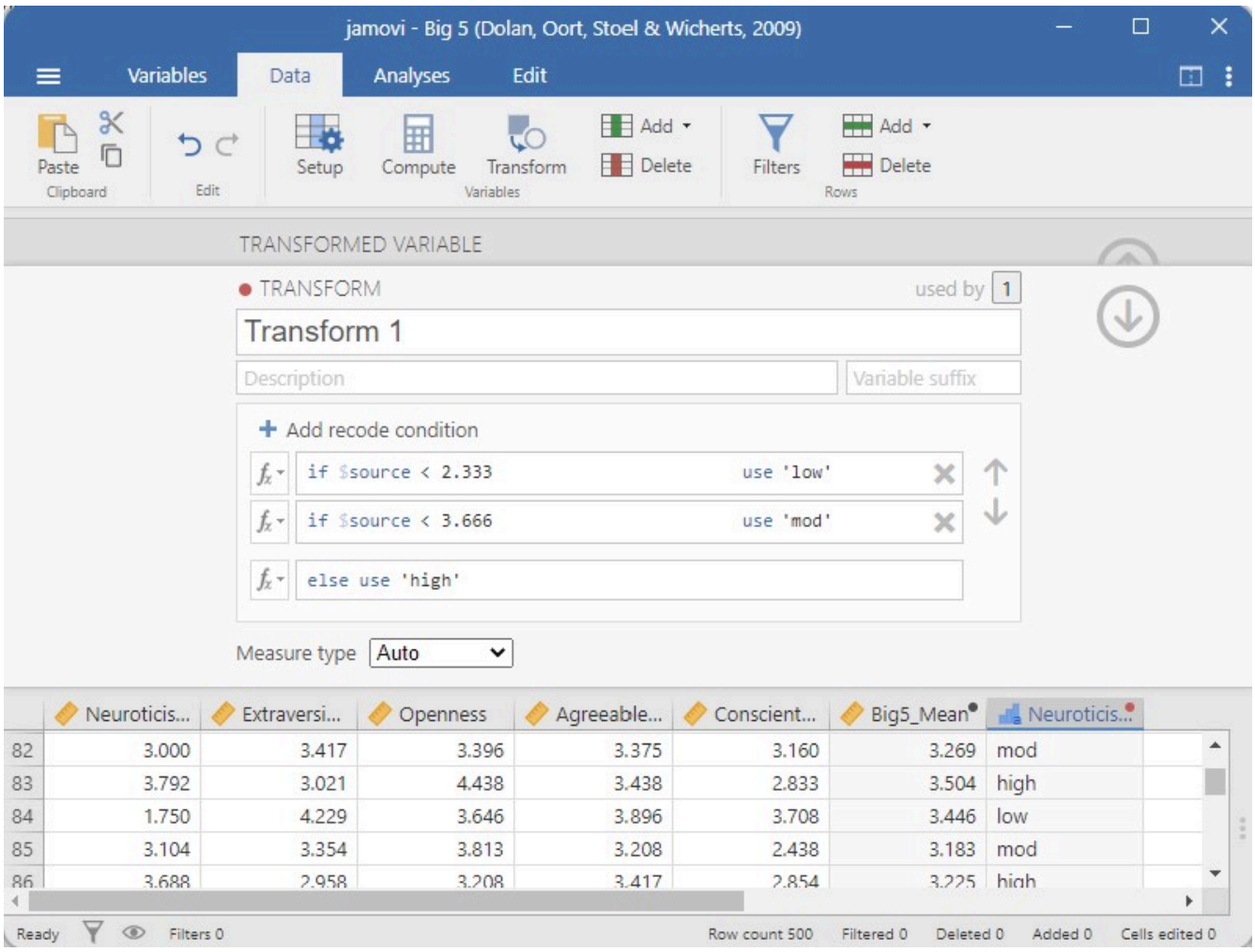
The happiness scale suggests that higher scores is higher happiness, but the fourth item is opposite. Higher scores on that item actually indicate lower happiness. Therefore, we would need to use "Transform" to recode the items so the highest score is the lowest score and so on. For example, if it were rated on a 5-point scale then you would need to recode so a 1 = 5, 2 = 4, 3 = 3, 4 = 2, and 5 = 1.

Recoding

Maybe we want to recode variables. Perhaps we want to recode the Neuroticism scale into low, moderate, and high extraversion. The scale ranges from 1-5, so I'm going to say that scores between 1-2.333 are low, 2.334 to 3.666 is moderate, and 3.667 to 5 is high. First, I create a new Transform variable. Then I need to specify the transformation. Click Edit to do so (or, when creating a new transformation, click the transformation and select Create New Transform).



We need to specify the recode conditions. Click `Add recode condition` twice. For the first formula, we want to specify that if the $source (meaning the score for the variable we're using for the transformation, in this case Neuroticism) is less than or equal to 2.333, then it will be recoded as `low`. Notice the use of apostrophes around the text! We do the same for moderate. Then we can end with an else statement: all other values (else) are recoded as `high`. We can either let it auto determine the measure type, but I like to be in control of my data and therefore specify it is an ordinal variable.



Multiple transformations

Maybe we instead want to do a computation across multiple variables. Perhaps we have multiple items that need to be reverse-scored, or in our case we want to use our previous `Low_mod_high` transformation to perform on *all* the subscales of the Big 5.

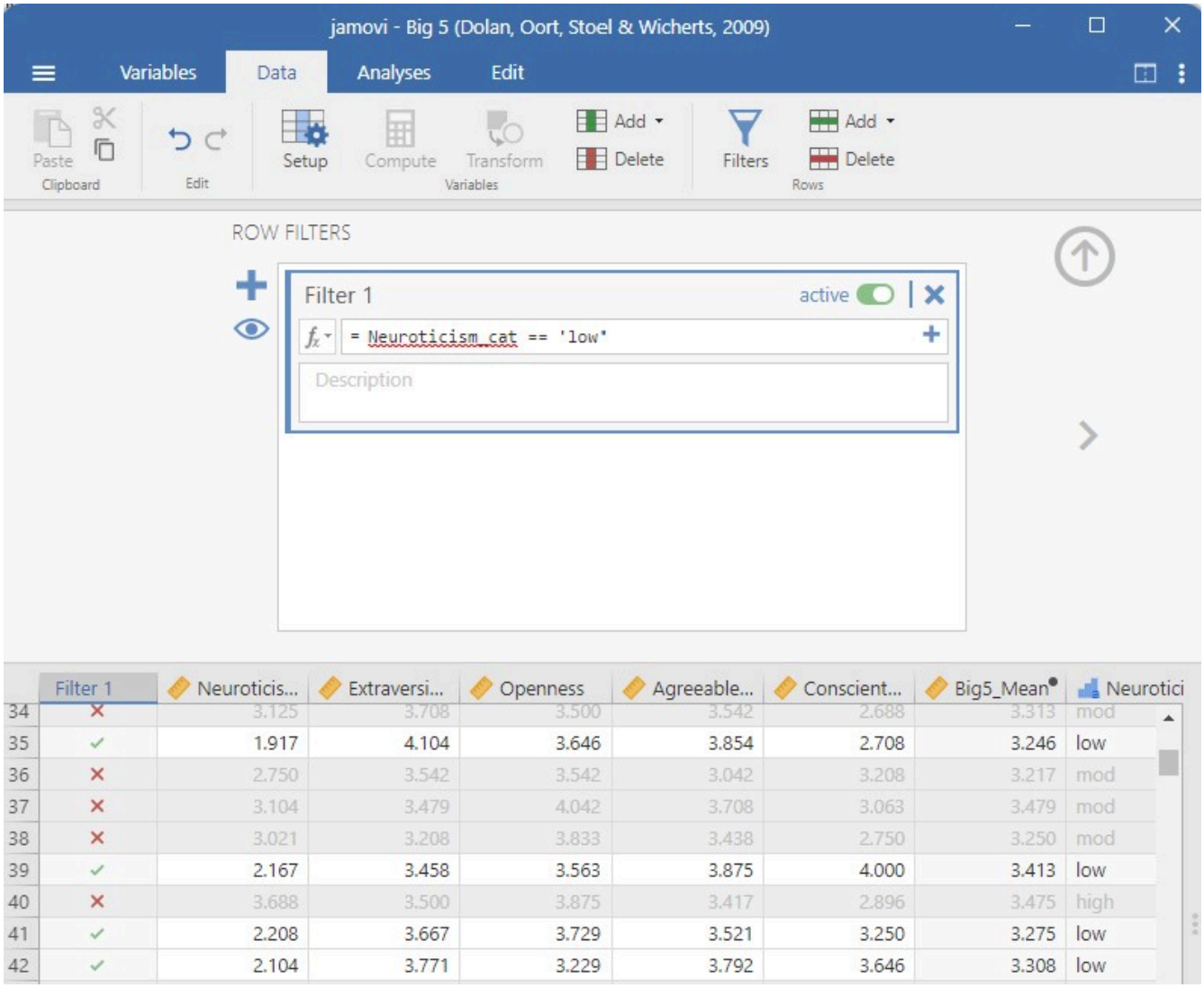
We can click a new variable (e.g., Openness), select Transform, rename the variable, and select the `Low_mod_high` transformation we already used. Voila! The work we did previously can easily be used again in this analysis.

4. Using Filters

Sometimes we only want to analyze certain pieces of our data. We can filter by rows and by columns. Check out this blog post by jamovi on more details of filters.

Row filters

Maybe we only want to analyze data from people who are low in neuroticism. We would create the following filter, by selecting Filters  . Type `Neuroticism_cat == 'low'` in the box, as shown below.



You'll notice at the very left of the dataset there is a new column named `Filter 1` (the name of the filter) and there will either be an X or a green check mark indicating whether it's removed (X) or kept (check) in the analyses.

If you want to take off the filter, but keep it available, click on the filter column and toggle the green button on the top right from `active` to `inactive`. It will then grey out the column.

A couple things to note:

- Notice that to say it equals to `low` you have to use a double equal sign: `==`
- Another common thing you may want to specify is that the variable is *not* equal to something. You would use the following: `!=`
- Otherwise you should be familiar with the other operations: `<`, `>`, `<=`, `>=`

Column Filters

Column filters are useful when you want to use a filter for *some* but not all of your analyses. Rather than creating a filter, we need to compute a new variable using the `FILTER()` function. For example, we can compute a new variable that is `FILTER(Neutoricism_cat, Neuroticism_cat == 'low'`. Then we could use that new variable in an analysis (in this case it's not very useful because there is no *variability* in this variable, but there are useful times for using column filters for analyses).

Describing Data

Descriptive statistics are used to summarize, organize, and overall *describe* our sample data.

We explore our data partly to describe our data and partly to check our data before performing inferential statistics. jamovi puts all our descriptive statistics into one useful analysis under the Exploration button (within the Analyses tab) called `Descriptives`. Whenever you want to understand the measure of central tendency or dispersion of a single variable (e.g., what is the mean score on a particular scale?) then you'll go to the `Descriptives` analysis under Exploration.

In the `Descriptives` analysis, these are under the `Statistics` drop-down menu. There are a ton of possible options!

1. **Sample size**: you can ask for the sample size (`N`) and number of missing values (`Missing`).
2. **Percentile values**: these are useful for creating quartiles (`Cut points for 4 equal groups`) or `Percentiles` of various sizes.
3. **Dispersion**: you should already be familiar with most of the measures of dispersion, particularly the `Minimum` and `Maximum` and the `Std. deviation` (SD) and `Variance` (which is just SD^2). We'll learn about the `S. E. Mean` later.
4. **Central Tendency**: similarly, you should also be familiar with all of the measures of central tendency: `Mean`, `Median`, `Mode`, and `Sum`.
5. **Distribution**: you should also be familiar with both `Skewness` and `Kurtosis` and later we will learn what those values mean and how that helps us test for normality.
6. **Normality**: lastly, there is a statistical test for normality called the `Shapiro-Wilk` test that we will learn about later.

You might be wondering how we report descriptive statistics in a Results section. Sometimes we use tables and figures to do this, or sometimes means and standard deviations are written in the text. We shall look at examples of figures in the next section, and you will see examples of tables and in-text descriptive statistics later in the book.

Visualizing Data

Why Visualize Data

"A picture is worth a thousand words," and in a world in which journal articles have word count limits, figures and graphs are priceless. They are also an incredibly powerful way to examine your data because it can often illuminate patterns you may not be able to see through a table. Critically, descriptive statistics (means, standard deviations, and even correlation coefficients) can be deceptive!

As an example, let's take a look at some summary statistics for four different datasets, each with an X variable and a Y variable:

Dataset 1

Descriptives

	X1	Y1
N	11	11
Mean	9.00	7.50
Standard deviation	3.32	2.03

Dataset 2

Descriptives

	X2	Y2
N	11	11
Mean	9.00	7.50
Standard deviation	3.32	2.03

Dataset 3

Descriptives

	X3	Y3
N	11	11
Mean	9.00	7.50
Standard deviation	3.32	2.03

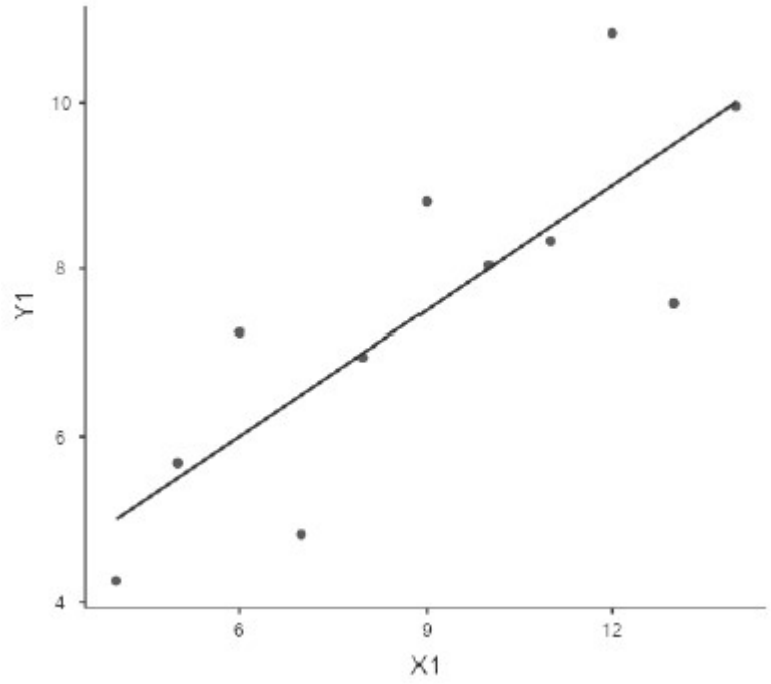
Dataset 4

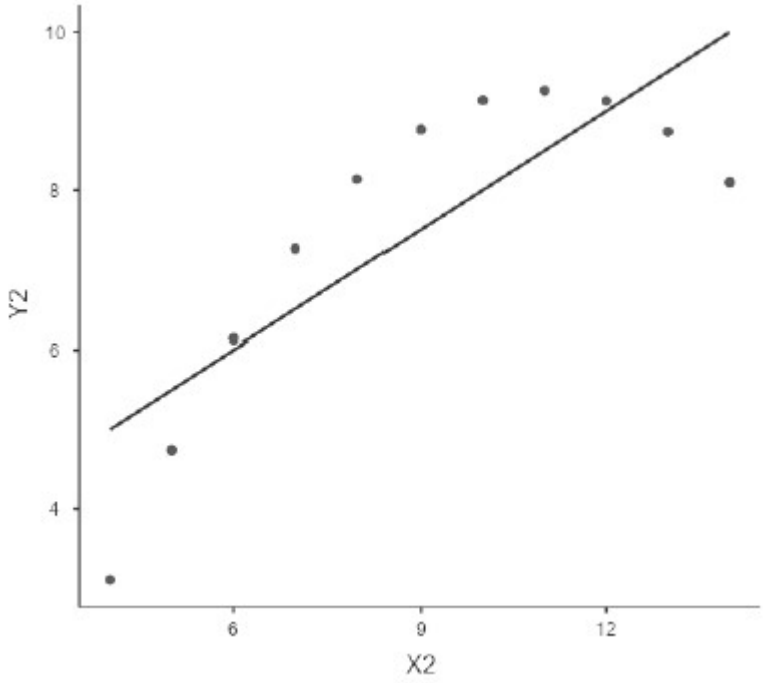
Descriptives

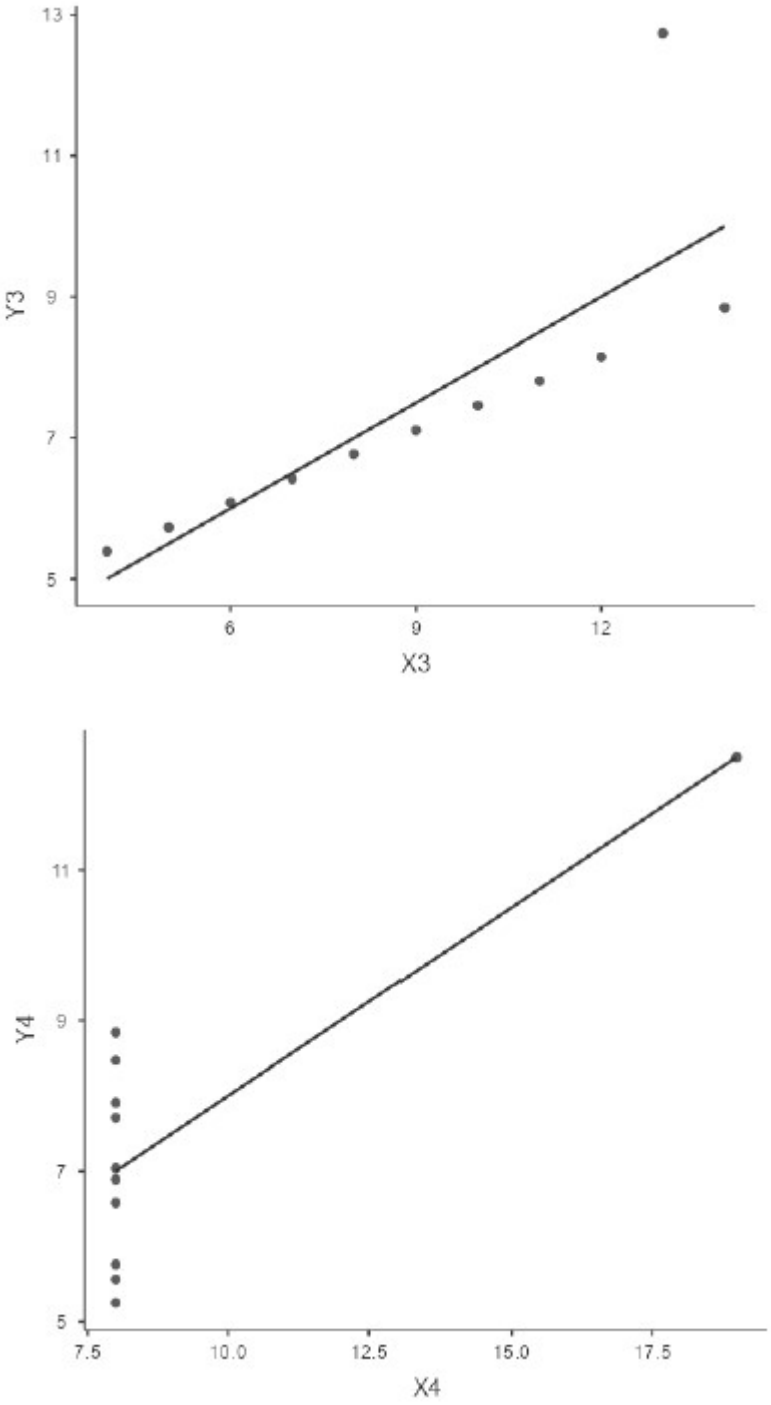
	X4	Y4
N	11	11
Mean	9.00	7.50
Standard deviation	3.32	2.03

Hopefully you notice that all four datasets have the same means and standard deviations for the X variable and the Y variable. And, no, that's not an error!

Let's take a look at a visualization (aka graph). Specifically, below, you will see a scatterplot for X and Y for each dataset. The straight line on each plot represents the line of best fit (the regression line – more on that in a later chapter).


10
8
6
4
Y1
6
9
12
X1


10
8
6
4
Y2
6
9
12
X2



Now you can see how vastly different these datasets truly are. In addition, you might be surprised to learn that for all of these datasets, the Pearson's correlation coefficient r = .82! (You might have some ideas about why this might not be a meaningful statistic in three out of the four cases, but we shall come back to that when we look at correlation coefficients in a later chapter.) So, the moral of the story is: **visualize your data**!

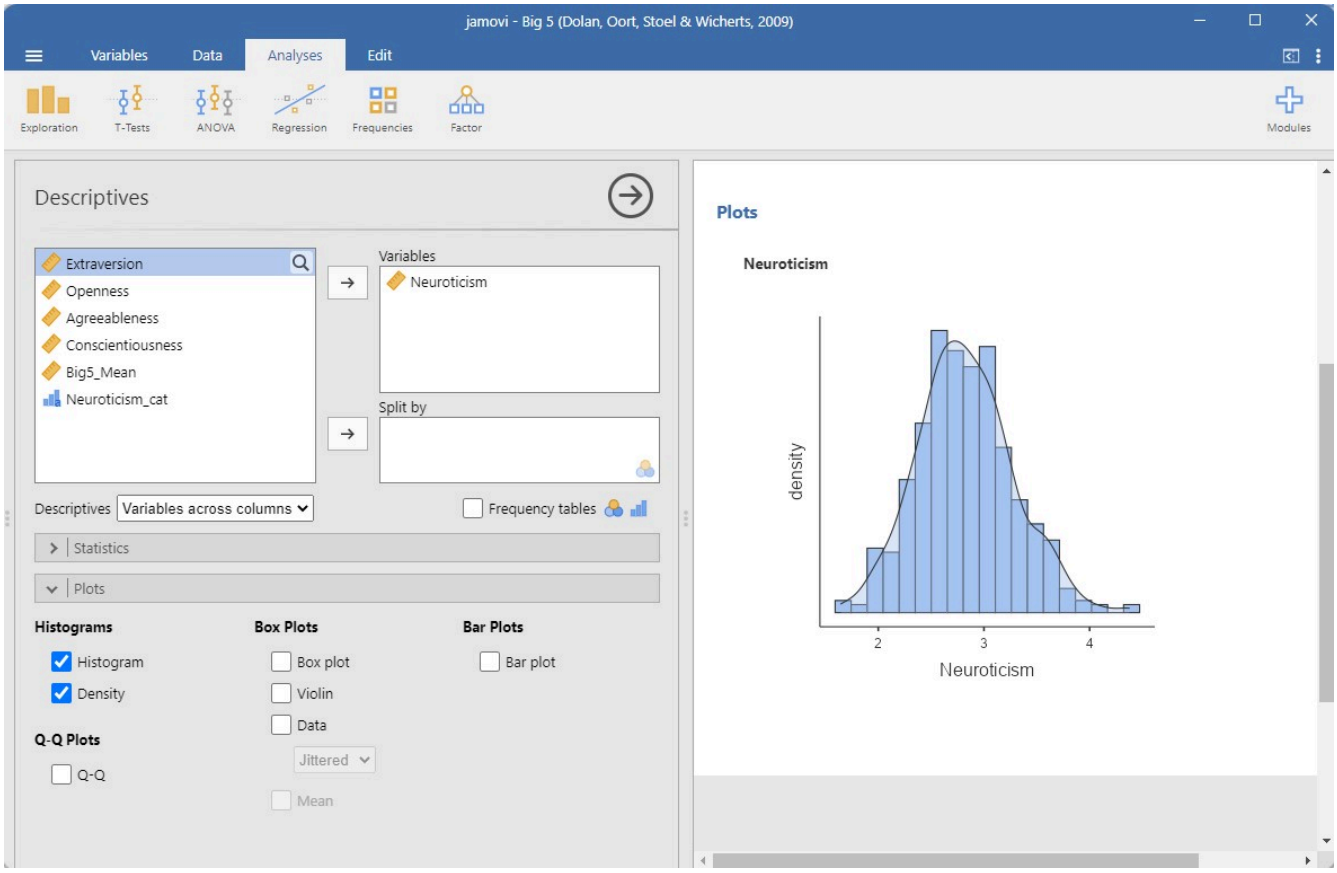
How to Visualize Data

jamovi has some plots built into its platform, both under the `Plots` drop-down menu in the `Descriptives` analysis and as options for many of the inferential statistical analyses.

We'll learn more about how to choose and conduct better data visualizations later, but for now here are some recommended visualizations depending on what you are trying to do.

When you want to visualize the distribution of a continuous variable

First, there are two **Histogram** options: `Histogram` and `Density`. These are useful for seeing the overall distribution of your data and to help check for normality. Which should you use? I think they're both pretty great, and in fact you can combine the two to have a histogram plot with a density overlay. I like this option best. If we go back to our Big5 data from the previous chapter, select the Analyses tab and then the Exploration button, and `Descriptives`, we can produce the following by selecting both `Histogram` and `Density` after selecting the variable Neuroticism:



At a glance, you can see that the Neuroticism variable is approximately normally distributed!

Other options in jamovi include the boxplot. It indicates the median, interquartile range, and range of the data. Try it for yourself using the Neuroticism variable in the Big5 dataset. You'll see a thick line in the middle, representing the median; the box itself spans from the 25th percentile to the 75th percentile; and the "whiskers" go out to the most extreme data point that does not exceed 1.5 times the interquartile range. Any observation whose value falls outside this range is plotted as a circle/dot and is commonly referred to as an **outlier**. How

many outliers are there for the Neuroticism variable? (To be able to see this clearly, I recommend also checking the `Data` box, below the `Box plot` option, because this will show you each datapoint separately on the graph.)[1]

When you want to visualize the distribution of a continuous variable split by a categorical variable

There are three options under **Box Plots**: `Box plot`, `Violin` (which is really a density plot with its mirror image!), `Data` (which can be Jittered or Stacked; I prefer Jittered so you can see the density of data points really well), and `Mean`. Personally, I love checking all four boxes! This gives you the best of all them: the distribution of your data with the `Violin` option, the quartiles and mean with the `Box plot` option, a visualization of all your data points using the `Data` option, which is really useful because the other two options can be *hiding* weird things in your data, and what the `Mean` is.

When you want to visualize the frequencies of a categorical variable

For this you would choose the single option under **Bar Plots**: `Bar plot`. It will simply show the frequencies of a categorical variable.

Once you've drawn all these pretty graphs, you might be wondering how you can get them out of jamovi to show your friends or create a new fabric for a dress. Well, you can do that quite easily: right click on the plot image, select Image, and then Export, and you can export it as a PDF, SVG, PNG, or EPS. You can also select All and export all your analyses as a PDF.

1. There are four outliers!

CHAPTER 4: POWER, EFFECT SIZE, AND REPRODUCIBLE RESEARCH

The sections on effect sizes and In practice are based on “Statistics with jamovi” by Dana Wanzer, with some minor changes .

Power and Effect Size

Power

What is power in reference to statistics? It is the **probability that your statistical test will detect an effect, given that there is an effect in the population**. It can be reported as a number from 0 to 1 or as a percentage, from 0 to 100% As you might have guessed, we want high power! So, how do we achieve high power?

Go to this app and follow the instructions to find out:

https://rpsychologist.com/d3/nhst/

First of all, adjust the settings as follows:

- Solve for? Power
- Significance level $\alpha = .05$
- Sample size n = 30
- Effect size d = 0.50 (this is sometimes called a medium effect size – more on this later)
- One-tailed test

How much power did you have? (Write it down—the number is a % below the graph)

Now, try the following and observe what happens to the value for power (i.e., does it go up or down?). In each case, write down your answer.

1. Increase the significance level to .10 (i.e., a less stringent α). Then return the significance level to .05.
2. Decrease your sample size to 20.
3. Increase your sample size to 50. Then return the sample size to 30.
4. Decrease the effect size to 0.20 (this is sometimes called a small effect size).

Feel free to play around with these variables some more until you get a sense of how each of them influences power. Now you should be able to fill in the blanks in the following paragraphs:

A very simple way to influence power is to tweak your α. Remember that trade-off between type I and type II error? Well, it turns out that power is related to type II error. In fact, power = 1- β (where β is the probability of making a type II error). So, if you increase your α (i.e., use a less stringent test) then you _____ power. So using $\alpha = .10$ instead of .05 will _____ the power of the test. Of course that comes at a cost of increasing your chance of a type I error, so this is not recommended.

What are some practical ways to influence power?

1. Increasing your sample size will _____ power.
2. Increasing the size of the effect will _____ power. Remember how we talked about aiming for a strong manipulation? This will increase the size of the effect and _____ power.
3. Given what you know about the roles of systematic and unsystematic variation in test statistics, and that "test statistic = signal/noise," or "test statistic = systematic variation/unsystematic variation" what do you think will happen to power if you reduce that unsystematic variation? Reducing variability will

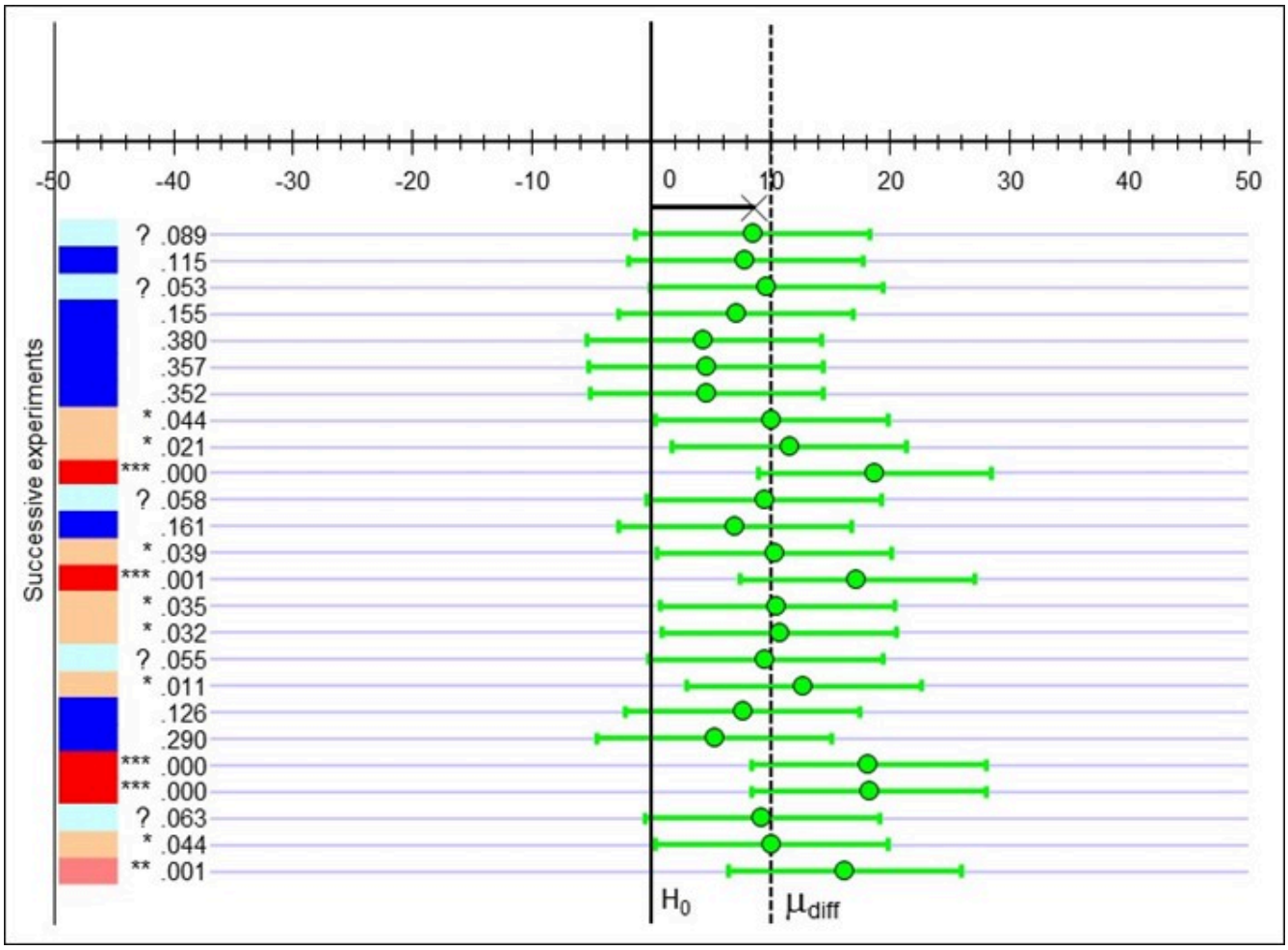
_____ power.

You might be wondering how much power we should aim for? Well, just as there is no perfect answer to what α level we should use, there is no ideal value for power. However, many researchers agree that 80% is a basic minimum that we should be aiming for. More on this later.

The key takeaway here is that effect size, sample size, alpha, and power are all related. If you know three out of the four of them, you can determine the fourth. Later in this chapter, you'll learn how to determine the appropriate sample size, given power, alpha, and the effect size of interest to you.

Effect Size

After running an inferential test, it might be tempting to look at the *p*-value and if $p < .05$, to call it a day! However, there are some big limitations to looking at *p*-values alone. Imagine you run a study to test the effects of completing practice test questions on students' stats exam scores. You are interested in the difference between the mean scores when students either do, or do not, complete the practice test questions. Under the null hypothesis, you expect no difference. Let's say the true difference in scores is a full 10 percentage points. You run a study with 32 participants but get no significant results. Should you give up your career as a researcher? Not so fast! If you go back and run that same study, with a different sample each time, 25 times, you might get something like this:



Here is a series of replications (i.e., the same study repeated again with different participants from the same

population). Each row in the figure represents a replication. In this hypothetical example, the true effect in the population is a medium effect size (*d* = 0.50). The figure gives us the difference between means under the null hypothesis—the black bar (= no difference between groups)—and the true difference between the mean test scores—the dashed line. The small green circles represent the mean difference for each study, and the green bars extending out from those circles represent the confidence intervals (CIs). If those CIs do not cross the solid black line, then you have a significant result. The *p*-values are on the very left (remember, they are significant if they drop below .05. Those that are close are marked with a ?).

Note that only 11 of the 25 studies (fewer than half of them) turned out to be significant here! All the means (the circles), are above the value for the null hypothesis, so the effect was in the expected direction for all the studies, but fewer than half were significant. The problem here is that power is only .52, or 52%. And your sample was relatively small at only 32 participants. The overall effect size across these many studies would average out to be about .50 – which is considered a medium effect size. The take-away point here is that just looking at the p-value from one study is not really very meaningful! And this really illustrates why it is important to think about power and effect size. If we repeated the exercise with a larger sample size for each replication (i.e., producing higher power), more of these studies' results would have been significant. Or, if the effect size in the population had been larger, more of these results would have been significant.

And... if you want to look at another cool visualization that will show you how confidence intervals vary across multiple experiments, go here:

https://rpsychologist.com/d3/ci/

You can adjust the sample size, too, and see how with larger samples, the confidence interval gets smaller and is more likely to include the true population mean (in this example, the true population mean is 0 and is represented by the dashed orange line going up through the middle of the graph).

Hopefully by now you have seen that *p*-values alone are not very useful. Remember, too, that they lead you to make these binary decisions—effect significant or not significant—but they do not tell us how big or important the effect is. And, by the way, the *p*-value is *not* linearly related to the size of the effect.

Therefore, the American Psychological Association Publication Manual now requires researchers to report effect sizes. An effect size is simply a standardized measure of the size of an effect. You might think that you can just report the difference between your two means (for example) as the effect size. That could be somewhat helpful, but the number obtained would not be easy to compare with results obtained in other studies. Therefore, we compute a standardized measure, which is comparable across studies: it is less reliant on sample size than *p*-values are, and it takes into account variability in the sample. Below are some of those standardized measures. Effect sizes can be measures of association (like Pearson's *r*) or the size of the difference (in standardized units) (like Cohen's *d*).

Standardized mean differences

- Cohen's *d* is one of the most popular standardized mean difference effect size measures
- Hedges' *g* is a less biased version of Cohen's *d*. Cohen's *d* is problematic for small sample sizes, so Hedge's *g* is preferred. Unfortunately, you will not always see it provided in statistical software.

Measures of association:

- Pearson's *r* indicates strength of association and R^2 indicates proportion of the variance explained.
- η^2 (eta-squared) measures the proportion of variance in the dependent variable associated with the different groups of the independent variable. This is considered a biased estimate, especially when trying to compare values across

studies, so there are two more preferred effect sizes.

- η^2_p (partial eta-squared) is calculated slightly differently and is considered a less biased estimate. This can allow for better comparisons of effect sizes across studies. It's still not perfect, though. You might also come across η^2_G (generalized eta-squared) and this is preferred by some authors.
- ω^2 (omega-squared) is calculated even more differently and is considered the least biased estimate. There is also ω^2_p and ω^2_G (generalized omega-squared) but we won't get into that.

(If you nerded out over this information and want to learn more, check out this great and very practical journal article by Daniel Lakens.)

For example, this is how to calculate Cohen's *d*. You can use it when you have two means:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s}$$

How would you interpret the value for *r* or Cohen's *d*? Well, there are some guidelines for *r* and *d*:

- *r* = .1, *d* = .2 (small effect):
  - the effect explains 1% of the total variance
- *r* = .3, *d* = .5 (medium effect):
  - the effect accounts for 9% of the total variance
- *r* = .5, *d* = .8 (large effect):
  - the effect accounts for 25% of the variance

We'll discuss % of total variance more when we get into regression, but essentially it tells us how much variability in the outcome is explained by the predictor. Beware of these 'canned' effect sizes though: the size of effect should be placed within the research context. A small effect size might be really important if you are talking about life and death situations. E.g., aspirin reduces the incidence of heart attacks. It's a small effect – *r* = .034, R^2 = .0011 (that is saying that 0.1% of the variance in getting a heart attack is explained by whether or not you take aspirin!). That's very small, but pretty important in this context, given that aspirin is cheap and with minimal side effects in the adult population.

In Practice: Sample Size Determination

When people talk about doing power analysis, really what they should be saying, usually, is sample size determination.

Here are a few key points to get you started.

1. When designing your research, aim for a large effect, small variability, and as large a sample as is necessary to achieve desired power (80% or more).
2. *Before* you start to collect data, you should decide what sample size is necessary to achieve desired power (i.e., an ***a priori* power analysis** or **sample size determination**). To determine what sample size is necessary, based on an expected given effect size and variability, use software like G*Power (it is free, and moderately complicated to use, depending on what you are trying to achieve). Fortunately, jamovi will also allow you to do some simple sample size estimation in the jpower module.
3. Always report confidence intervals and effect sizes, not just *p*-values! We shall look at appropriate measures of effect size for each of the tests we use in this course. Fortunately, jamovi will typically compute these for you.

When it comes to sample size determination, sometimes, for various reasons, you cannot obtain the desired sample size. For example, you might only have two weeks to collect data for a class project or your population of interest is difficult to recruit or test (e.g., babies or people with a specific medical condition). Therefore, when we think about power and sample size determination, there are three things you might be interested in figuring out:

1. What sample size do I need given the effect size of interest, alpha level, and power level?
2. What power do I have to detect the effect size of interest given my alpha level and sample size?
3. What effect size can I reasonably detect given my alpha level, power level, and sample size?

Sample Size Determination

Let's say we want to address the first of the three above: for a specified level of power, alpha, and given effect size of interest, we are going to determine what sample size we need.

Note that the jpower module in jamovi is fairly limited in the analyses you can do. (You can install jpower by clicking on the  at the top right hand corner of the screen in jamovi. Then select jamovi library, scroll down to find jpower, and install it.) If you need to determine sample sizes for other types of analyses, then go to G*Power or, better yet, ask someone with experience in R to do it for you in R.

jpower will allow you to compute the required sample size for *t*-tests, only. To do this, open a blank jamovi datafile (click on the hamburger in the top left corner, and select New). Click on the jpower module and choose the kind of test you plan to run in your study (as I mentioned, this is limited to *t*-tests).

Let's say we are going to have a two-group between-subjects design, for which we shall run an independent

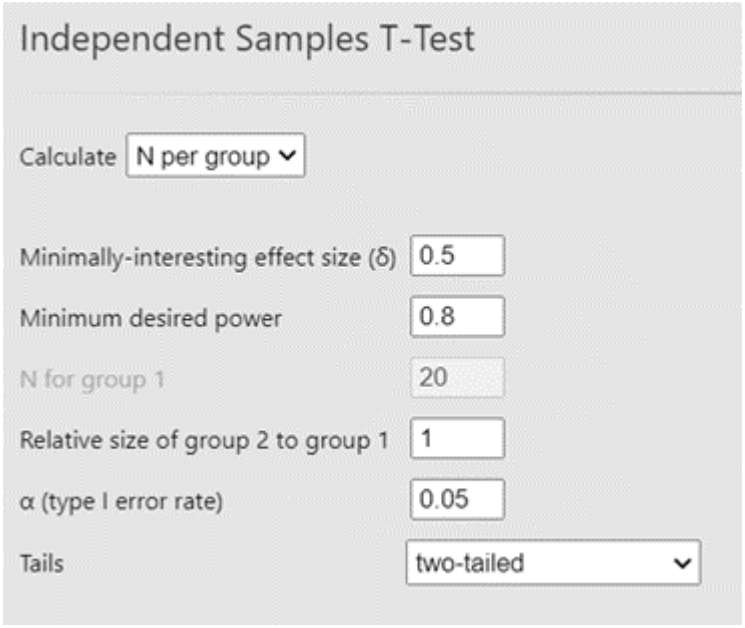
samples *t*-test. We want to know how many participants we shall need in each group in our study. You will need to decide a few things. A minimum level of acceptable power is typically set at 0.8 (80%), but you can also run the analysis for different levels of power. I would plan to have the same number of participants in each group (though you can specify a different ratio if you need to, by changing the Relative size of group 2 to group 1) and assume for this demonstration that you are confident about being able to run a one-tailed test (you have a clear, directional hypothesis, based on prior research and/or theory: e.g., you expect group A to score higher than group B, and not the other way around). You can use $\alpha = .05$.

Finally, you need to choose the minimally-interesting effect size. According to Daniel Lakens (https://psyarxiv.com/9d3yf/), there are a few ways to decide on what effect size interests you:

1. "Smallest effect size of interest: what is the smallest effect size that is theoretically and practically interesting?
2. Minimally statistically detectable effect: given the test and sample size, what is the critical effect size that can be statistically significant?
3. Expected effect size: which effect size is expected based on theoretical predictions or previous research?
4. Width of confidence interval: which effect sizes are excluded based on the expected width of the confidence interval around the effect size?
5. Sensitivity power analysis: across a range of possible effect sizes, which effects does a design have sufficient power to detect when performing a hypothesis test?
6. Distribution of effect sizes in a research area: what is the empirical range of effect sizes in a specific research are, and which effects are a priori unlikely to be observed?" (p. 3)

Basically, these are saying: what does past research have to say about what effect size you can expect (#3 and #6)? What is the smallest effect size you care about (#1)? Consider how important it is to spend time/money/other resources chasing small effect sizes. If your research could have profound consequences for society even with really small effect sizes, then it might be worth it. In other situations, you might decide that it is only worth looking for a medium effect size. What is the smallest effect size you can reasonably obtain (e.g., due to sample size limitations; #2, #3, and #4)? This is the justification you use to determine what effect size you are looking for. Note that in jamovi, the δ (Greek delta symbol) denotes Cohen's *d*.

Let's say you decide that it is reasonable to expect/look for a medium effect size of $d = 0.5$, aiming for power = 80%. This is what you should see in jamovi:



To the right, jamovi automatically populates the "Results" pane. You will see that you will need 64 participants in each of your two groups to achieve 80% power with effect size *d* = 0.50

You can also select different plots. Check all the boxes and also ensure you check the box next to Explanatory text, under Additional Options. jamovi provides a very detailed explanation of what each plot shows you.

The Power Contour graph shows you the relation between power, effect size and sample size. From this you can get a sense of how much power you will have if your effect size turns out to be smaller or larger or if you do not reach your intended sample size.

The Power Curve by N graph gives you a sense of how much you would need to increase your sample size if you wanted to achieve 90% power (somewhere closer to 100 per group). You might want re-run the test with power set to 0.9 and see what happens.... Now you will need 86 per group.

What to aim for? Well, of course 90% power is better than 80% power, but if you have limitations on your data collection (e.g., the pool of participants is not that large), you might just aim for 80% power.

CHAPTER 5: TWO-SAMPLE EXPERIMENTS AND T-TESTS

Portions of this chapter (descriptions of three different types of *t*-tests, In practice, and Alternatives to the *t*-test) are based on "Statistics with jamovi" by Dana Wanzer, with some changes and additions (including more extensive discussion of the assumptions and robust tests and Welch's *t*-test).

Designing the Two-Sample Experiment

The most simple experiment is one in which there is just one independent variable, with two levels, and one outcome. Systematically manipulating the levels of the independent variable allows us to make causal inferences. Note: it is not the kind of statistical test that we use that allows us to make causal inferences, but the *design of the study*. It needs to be an experimental design in order for us to make causal inferences.

For example, you might ask:

1. Are people who write down their New Year's resolutions for gym attendance more likely to stick to them than people who just think about them?
2. Do parents who make their kids' Halloween costumes eat more candy than those who buy them?
3. Can you type your essay faster if you have had a cup of coffee before typing than if you have not?

Error Variance: Why You Should Care

When you design your study, you need to think carefully about error variance. Imagine you want to find out if people who write down their New Year's resolutions to go to the gym actually attend the gym more often than people who just think about them. You randomly assign people to either write down or think about their resolutions and then measure frequency of gym attendance in the month of January. Now, you are going to have variability in your dependent variable just due to random fluctuations. What kinds of things will affect gym attendance, other than whether people wrote or thought about their resolutions? Things like: pre-existing fitness level; preference for gyms vs. other forms of exercise, distance to the gym, the fact that people had to keep track of their gym attendance, and so on. These are all **extraneous variables** (see the first chapter in the book if you need a refresher). Some of these are nuisance variables, i.e., they vary randomly and cause unsystematic variation in the dependent variable.

As you will also recall from the first chapter all test statistics can be boiled down to the general equation:

$$Test\ statistic = \frac{Systematic\ variation}{Unsystematic\ variation}$$

Ideally, we want a large amount of systematic variation and a small amount of unsystematic variation, so that in the end we get a large value for our test statistic. That unsystematic variation is what is also known as **error variance**, and is caused by nuisance variables. So we need to minimize the influence of nuisance variables on our dependent variable, as much as possible. By minimizing the variability within groups, we will have a more powerful design. So, error variance (and our ability to minimize it by controlling nuisance variables) is important for **power**.

For example, compare the following two, different datasets ("Low power" is one study and "High power" is another study) for our new year's resolutions study. The scores represent the number of times each participant (n = 6 per group – yes, that is a low sample size, which is problematic for many reasons, but it works for the purposes of this particular demonstration) went to the gym in the month of January:

Low power		**High power**	
Wrote	Thought	Wrote	Thought
5	2	12	7
18	5	13	7
7	4	13	8
16	14	14	7
13	10	13	7
19	7	13	6

On the left, we have a dataset with low power and on the right we have a dataset with high power. In the low power scenario, you'll notice by glancing over the numbers that there is no obvious difference in gym attendance between folks who wrote about their new year's resolutions compared to those who simply thought about them. However, in the high power dataset, the difference between the groups jumps out at you – it is easy to see just at a glance that the people who wrote about their new year's resolutions scored, on average, higher than those who thought about them.

Let's take a look at some basic descriptive statistics for each of these scenarios:

Descriptives

	Resolution	Low power	High power
Mean	Thought	7.00	7.00
	Wrote	13.0	13.0
Standard deviation	Thought	4.38	0.632
	Wrote	5.83	0.632

You'll notice that the means for the Thought and Wrote group are actually the same for the low and high power datasets! However, the standard deviations are much larger in the low power dataset compared to the high power dataset. This is due to error variance: when we have a lot of error variance, it is like listening to the radio when there is a lot of static – it is hard to detect the signal from the noise. When error variance is low, the differences 'pop out' at you, and in those cases we have high power. Let's run the *t*-test and see what that shows us:

Independent Samples T-Test

		Statistic	df	p
Low power	Student's t	−2.01	10.0	0.072
High power	Student's t	−16.43	10.0	< .001

Notably, the *p*-value is well below .05 in the high power scenario (we would say that there is a significant difference between the groups), but not in the low power scenario. So, controlling for those nuisance variables in order to reduce error variance matters!

Of course, don't forget that we also want to reduce the presence of **confounds**. Confounds are a special form of extraneous variable that *systematically vary* along with our independent variable – so their effects on the dependent variable are actually captured in the "systematic variation" part of our general equation for the test statistic above. They can inflate the value of the test statistic, making it look like there is an effect of the independent variable on the dependent variable when really there is not!

Error Variance: What To Do About It

Here are two general things we can do, regardless of our design, to reduce error variance:

1. Use controlled conditions: in other words, keep everything constant except for the variable we are manipulating; and
2. Use reliable measures.

There are also some more specific things we can do about error variance that depend on the type of design we have.

Between- or Within-Subjects Design?

One major consideration is whether to use a between-subjects or within-subjects design.

In the between-subjects design, participants would only take part in one level of the independent variable. In the within-subjects design, all participants take part in both levels.

The table below shows some other names you might hear for these two types of designs:

Between-subjects	**Within-subjects**
Independent samples	Dependent samples
Independent groups	Dependent groups Repeated measures

You might hear these described as independent samples (between) or dependent samples (within). In the first case, scores in the two conditions are independent of one another, because different participants contribute to the scores in the different conditions. In the second case, the scores are dependent, or associated with each other, because the same participants contribute to the scores in the two conditions.

Now, depending on which design you use, there are different potential confounds to consider and different ways to control for extraneous variables.

Strategies for Between-Subjects Designs

There are three general tricks to dealing with extraneous variables in the between-subjects design:

1. Random assignment of participants to condition:
    - This does not guarantee equivalent groups but reduces the risk of confounds due to selection differences.
2. Balancing participant variables: for example, in our new year's resolution example, we might first ask how close people live to the gym, and then randomly assign them to each condition while simultaneously ensuring an even distribution of people who live close (e.g., < 1k) or far (e.g., >= 1k) from the

gym in each group.

    - This reduces the likelihood of participants with particular characteristics being more prevalent in one group or another when participants are randomly assigned – i.e., the extraneous variable is still present in the dataset, but will not become a confound.

3. Limit the population: for example, we might limit the study to people who first report that they prefer going to the gym over other forms of exercise

    - This eliminates the extraneous variable in question (but reduces external validity).

Within-Subjects Designs

If the between-subjects design is not adequately controlling extraneous variables, we can go to a within-subjects design. One approach is to use a **matched-groups design**. In this design, people are matched, in pairs, on one or more variable (e.g., age, distance to gym, current level of fitness, in our new year's resolution example) and then randomly assigned from each pair to one of the two groups. It involves different participants in each group, but because those participants are *matched* we treat it like a within-subjects design when it comes to data analysis. This approach helps control for nuisance variables that can vary across conditions and mask the effect of the independent variable.

An ideal form of matching is to go to a full within-subjects (repeated measures) design. In this case, all participants take part in both conditions. This has the advantage of fully controlling for extraneous variables, because each participant serves as their own control. Any differences between the two conditions cannot be attributed to participant variables because they are the same participants in each condition. It is therefore an ideal form of matching.

If we are going to use a within-subjects design, we need to ensure that we use **counterbalancing**, otherwise, the order of conditions is a potential confound. In other words, we vary the order of exposure to the levels of the independent variable. Let's say we have condition A and B. Half the participants would be exposed to condition A, then B, and the other half of the participants would be exposed to condition B, then A.

Caveat: of course, we cannot always use a within-subjects design. For example, a within-subjects design would not work if we wanted to test the effects of a new anxiety intervention on test anxiety, compared to a no intervention condition. The demand characteristics would likely be high, and participants would not be able to "unlearn" the effects of the anxiety intervention when they take part in the control condition. Also, within-subjects designs are more likely to result in fatigue (because the experiment will be twice as long) and have added potential confounds (e.g., practice effects) to consider.

Between or Within?

There are lots of things to consider when choosing a design. You should always weigh up the pros and cons of each design – and choosing between a between- or within-subjects design is a critical part of this process. One extra consideration is power: all other things being equal, power will be higher for the within-subjects design, because you will have the best control over participant variables in this case.

t-Tests

The *t*-test looks at difference in means between two things (e.g., groups, time, observations). There are three different types of *t*-tests:

1. The **one-sample *t*-test** tests how the sample mean relates to the population mean.
2. The **independent samples *t*-test** has two *independent* groups. The participants or things in group 1 are *not* the same as the participants or things in group 2. This is a between-subjects design in which different participants are in the two groups.
3. The **dependent samples *t*-test** has *dependent* or *paired* data. The dependent variable is measured at two different times or for two different conditions for all participants or things. This is a within-subjects design in which case the same participants are in both groups. Note also that if you have a matched groups design, you would also use the dependent samples *t*-test.

These tests are all what we call *parametric tests*. This means that they rely on a certain set of assumptions, which must be met, in order for us to be able to trust the results of the test. We shall talk about testing assumptions for the *t*-tests at the end of this chapter. In this book, we are going to focus on independent and dependent samples *t*-tests.

What Is a *t*-Test?

The test statistic we compute when we run a *t*-test is called *t*. Here is a generic formula for a *t*-test (Field, 2013):

$$t = \frac{difference\ between\ sample\ means\ -\ difference\ between\ population\ means\ if\ null\ is\ true}{estimate\ of\ standard\ error\ of\ difference\ between\ two\ sample\ means}$$

When we run a *t*-test, we compute *t* for our particular sample and then determine the probability of getting a *t* that extreme or more extreme if the null hypothesis were true. The larger our *t*-value, the lower the likelihood of it occurring when the null is true and therefore the smaller the *p*-value. At some point (usually at $p < .05$), our *p*-value is small enough that we think it's pretty unlikely to get a value that extreme (or more extreme) when the null is true, and so we reject the null hypothesis and conclude that there was a significant difference between our groups, or a significant effect of the independent variable on the dependent variable.

The Independent Samples *t*-Test

When we run a *t*-test for independent groups, we first imagine plotting a sampling distribution of all the possible differences between the means of the two groups that we could obtain if the null hypothesis were true, and with the sample size of our particular sample (because the shape of the distribution changes with our sample size, or, more specifically, the degrees of freedom; the larger the degrees of freedom, the thinner the

tails of the distribution). Statisticians have handily computed the distribution of *t* for any different sample size that you want.

For independent groups, the *t*-test looks like this (Field, 2013):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

Note that we just have the difference between two means as the numerator in the equation. What about the difference between the means when the null is true (which you saw earlier in the chapter)? Well, this is assumed to be zero, and so it drops out of the equation. We divide the difference between our two means by the estimated standard error. How do we compute this standard error when we have two groups? Essentially, we use a weighted average of the variance estimates for each group (weighted based on the sample sizes of each group). The formula below shows you how the pooled variance is calculated (Field, 2013):

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^1}{n_1 + n_2 - 2}$$

(I won't test you on the pooled variance formula!)

The Dependent Samples *t*-Test

When we run a dependent samples *t*-test, we are imagining going into the population an infinite number of times and plotting a sampling distribution for all the possible mean differences between the scores in our samples.

$$t = \frac{\overline{D} - \mu_D}{\frac{s_D}{\sqrt{N}}}$$

D-bar is the mean difference between our samples. μ_D is the difference we would expect when the null is true (i.e., zero). The standard error in the denominator is the standard error of the differences (computed by dividing the standard deviation by the square root of *N*, our sample size) (Field, 2013).

Assumptions of the t-Test

t-tests are what are known as parametric tests. In other words, they are based on the normal distribution and, as such, they rely on meeting certain assumptions. Before you report the results of a *t*-test, you should always check that you meet the assumptions.

t-tests are based on the assumptions that:

1. The dependent variable is interval or ratio data.
2. The sampling distribution is normally distributed. For the dependent samples *t*-test, this means that the sampling distribution of the differences between the scores should be normally distributed.

The independent samples *t*-test also assumes that:

1. Homogeneity of the variance: variances in the populations are equal.
2. Scores in the different groups are independent (because they come from different people, they should be!).

In Practice: t-Tests

Let's take a look at how to run the independent samples and dependent samples *t*-tests using jamovi.

Remember our four steps to data analysis:

1. Look at the data
2. Check assumptions
3. Perform the test
4. Interpret results

Please install the lsj-data module by clicking on the plus sign in the top right hand corner of the jamovi window. Once you have installed it, close jamovi and re-open it. When you go to Open and select Data Library, you'll see a new folder called "learning statistics with jamovi." In that folder, you'll find several example datasets that we shall use in the remainder of this book.

Independent Samples *t*-Test

For now, open "Harpo." This dataset is hypothetical data of 33 students taking Dr. Harpo's statistics lectures. We have two tutors for the class, Anastasia (n = 15) and Bernadette (n = 18). Our research question is "Which tutor results in better student grades?"

1. Look at the Data

To conduct the independent t-test, we first need to ensure our data is set-up properly in our dataset. This requires having two columns: one with our continuous dependent variable and one indicating which group the participant is in. Each row is a unique participant or unit of analysis.

Below are the first ten rows of our data from the Harpo dataset.

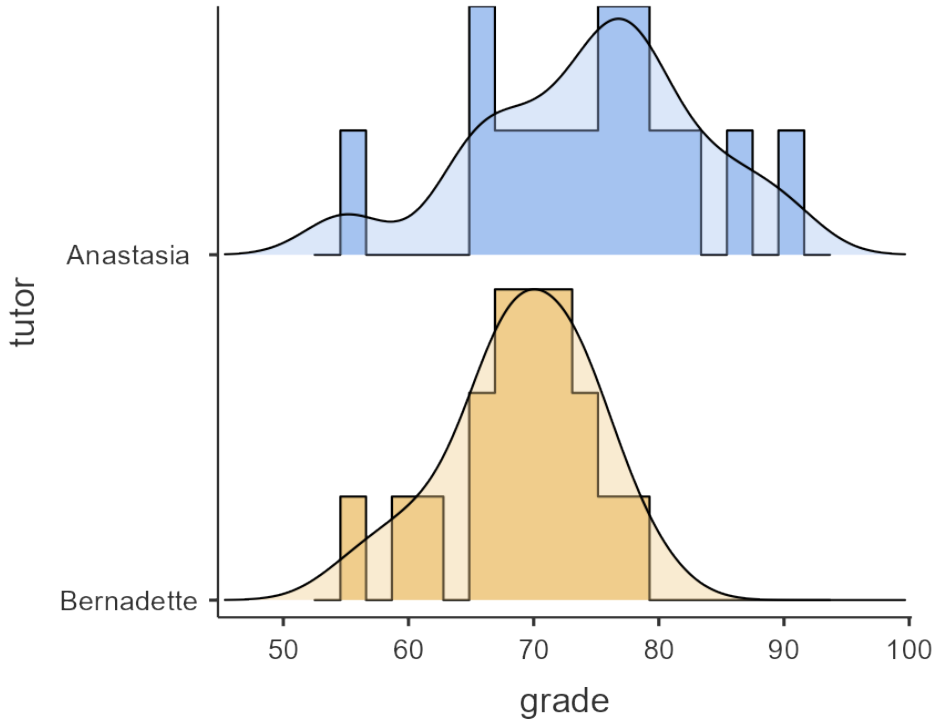
	ID	grade	tutor
1	1	65	Anastasia
2	2	72	Bernadette
3	3	66	Bernadette
4	4	74	Anastasia
5	5	73	Anastasia
6	6	71	Bernadette
7	7	66	Bernadette
8	8	76	Bernadette
9	9	69	Bernadette
10	10	79	Bernadette

In this dataset, what is the independent variable and what are the levels? What is the dependent variable?

Notice that the variable "grade" is set up as a nominal variable, but we should change this to continuous. Once we have done so, we should look at our data using descriptive statistics. As you learned before, select Exploration and Descriptives. They are shown below. You'll see that I split the data by tutor. I left the default settings for the statistics and also requested Histogram and Density under Plots.

Descriptives

	tutor	grade
N	Anastasia	15
	Bernadette	18
Missing	Anastasia	0
	Bernadette	0
Mean	Anastasia	74.5
	Bernadette	69.1
Median	Anastasia	76
	Bernadette	69.0
Standard deviation	Anastasia	9.00
	Bernadette	5.77
Minimum	Anastasia	55
	Bernadette	56
Maximum	Anastasia	90
	Bernadette	79



Our overall data consists of 33 cases. The minimum and maximum values look plausible; theoretically, student grades should range from 0-100. Lastly, the distribution of data looks roughly normally distributed. Before we proceed with our analyses, we should look a bit more closely at our assumptions.

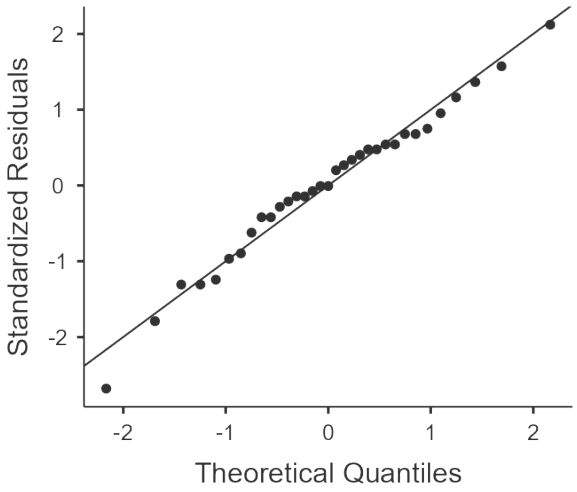
2. Check the Assumptions

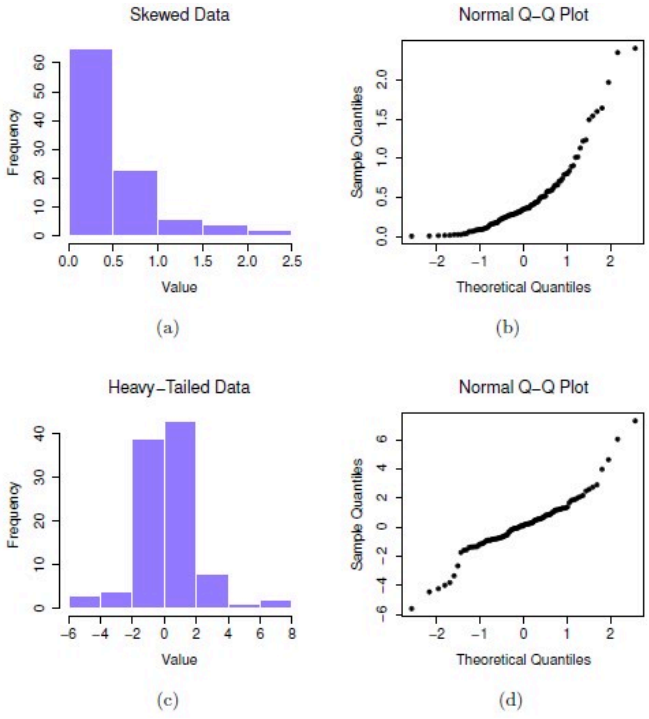
The dependent variable seems to be ratio data (grade) and let's assume that whoever collected these data is certain that scores are independent (no students attended both Anastasia's and Bernadette's tutorial sessions!). What about normality and homogeneity of the variances?

Check Normality

We can check normality by adding a request for Normality test and Q-Q plot under the Assumption Checks when we request the independent samples *t*-test: go to Analyses, T-Tests, and then Independent samples T-Test. Select the grade variables in the Variables box, and put the tutor variable in the Grouping variable box. Under Statistics select the Q-Q plot and Normality test.

The Q-Q plot, shown on the right, is created as follows: each observation from the dataset is plotted as s single dot, where the x coordinate is the theoretical quantile that the observation should fall in if the data were normally distributed (with mean and variance estimated from the sample), and the y co-ordinate is the actual quantile of the data within the sample. If the data are normal, the dots should form a straight line. Ours looks pretty good – and what we see is in alignment with what we would expect given the visual inspection of the histograms above.





On the other hand, the examples on the left show data that are not normally distributed. (a) and (b) are the histogram and Q-Q plot for positively skewed data and (c) and (d) show the histogram and Q-Q plot for data with heavy tails (i.e., high kurtosis). In each case, you can see that the histogram does not show a nice, normal shape, and dots on the Q-Q plot do not fall close to a straight line.

Another way to assess normality is by examining the results of the Shapiro-Wilk statistic. The Shapiro-Wilk statistic is produced by your having checked the Normality test box (note, you can also obtain both this test and the Q-Q plot in the Descriptives when you initially explore your data).

The Shapiro-Wilk test assesses whether the data deviate significantly from normality. We interpret it the same way we would any other inferential test. If $p < .05$ then the data deviate significantly from normality. In our case, with the grade variable from the Harpo dataset, the value for the Shapiro-Wilk statistic W = 0.98, and $p < .827$. Therefore, it is not significant and we would conclude that the data do not deviate significantly from normality.

However, in reality, it is a bit more complicated than that. First, many would argue that the assumption is not that the data themselves should be normally distributed, but that the sampling distribution should be normally distributed. You may recall, for example that the sampling distribution of the means is always normal, regardless of the shape of the distribution of the underlying raw scores, as long as the sample is large enough (at least $N = 30$; you might remember this from central limit theorem and I'll show you a demo of this in class!).

In addition, there's a caveat to interpreting the Shapiro-Wilk test. As you'll recall from earlier in this book, sometimes we do not have enough power to detect something that actually exists in the population, and we can make a type 2 error as a result. You might also remember that sample size, effect size, alpha, and power are all interrelated and that a small sample size yields lower power than a large sample size (all other things being equal). What this means is that with small sample sizes, Shapiro-Wilk may not be significant even with relatively large deviations from normality. And, it is with small sample sizes that central limit theorem does not hold. In other words, with small sample sizes, the Shapiro-Wilk test might tell you that the data are normally distributed (because you do not have enough power to detect non-normality) and you cannot safely assume that the sampling distribution will be normally distributed. Thus, there is a risk that you are unknowingly violating the assumption of normality.

Fortunately, there is a solution. If you are concerned about normality, especially with a small sample, you can use what is called a "robust *t*-test." A robust test is a version of the test that is less impacted by violations or normality. For the *t*-test, in jamovi, you can download the Walrus package and use the robust *t*-test within that package (there's a robust test for independent samples and for dependent samples; use the default trim proportion of 0.2). Another alternative is to use the Mann-Whitney U test (this one is available with the regular jamovi *t*-tests options).

Checking Homogeneity of the Variances

The assumption of homogeneity of the variances is that the variances (in the population) are homogenous (i.e.,

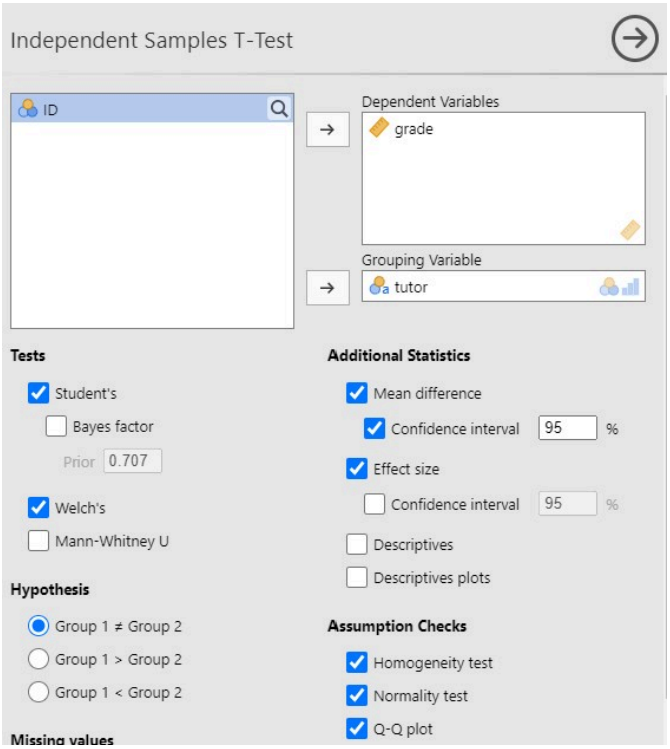
the same). If they are not close enough in our sample, then we cannot trust the result of the *t*-test. You can easily test for homogeneity of the variances in jamovi: when you select your *t*-test, you can request the Homogeneity test (again, under Assumption Checks). This will produce Levene's test for equality of variances. This test does exactly what the name suggests: it tests whether the variances in the two groups are significantly different from each other. If the *p*-value from this test is < .05, then the variances are significantly different. However, the same caveat applies here as to our interpretation of the *p*-value for the Shapiro-Wilk test: if our sample size is small, power will be low, and we will not be able to detect differences between the variances even when they are quite large. At the other end of the spectrum, with large samples, the test may be significant even when there is only a small difference in the variances. It is for this reason that many researchers now recommend using Welch's *t*-test instead of the Student's *t*-test (Student's is the standard one that you probably learned about in your lower level stats class). Welch's *t*-test does not make the same assumption of homogeneity of the variance. Fortunately, Welch's *t*-test is easily available to us in jamovi: you will see it as an option under the Student's test.

In our Harpo dataset that we are working with, Levene's test for homogeneity of the variances shows that *p* = .125 (i.e., not significant). This suggests that the variances of the grades of the two groups (i.e., students who had Anastasia as a tutor vs. those who had Bernadette as a tutor) are not significantly different from each other. However, as discussed above, with small samples, we cannot trust Levene's test. We should probably use Welch's *t*-test.

3. Perform the Test

"At last!" I hear you sigh. Let's imagine you had proposed a two-tailed hypothesis: that there will be a difference in grades for Anastasia's and Bernadette's students. Now you can do your inferential test and answer the question: did Anastasia's students and Bernadette's students have significantly different grades?

Do not forget the outcome of the assumption checks. We talked about how, given the sample size, we should be cautious about interpreting the Shapiro-Wilk test and Levene's test. You might be wondering if you should do the Robust test or Welch's *t*-test in this situation. A glance at the histogram showing the distribution of scores suggested that the raw scores themselves are somewhat normal. Thus, in this case, I would recommend we go with Welch's *t*-test. Let's also get the Student's *t*-test and we can see how the results compare.



In the Analyses tab, select T-Tests and Independent Samples T-Test. Ensure that you move the grade variable over to the Dependent Variables box, and the tutor Variable over to the Grouping Variable box.

You'll note on the right that I've selected both Student's and Welch's. In addition, I have requested the Mean difference and Confidence interval, as well as Effect size. If you wish, you can get Descriptive statistics and plots again.

Under Hypothesis, I have specified that I am interested in a two-tailed test – I am simply predicting that there is a difference between the two groups, rather than the direction of the difference.

4. Interpret Results

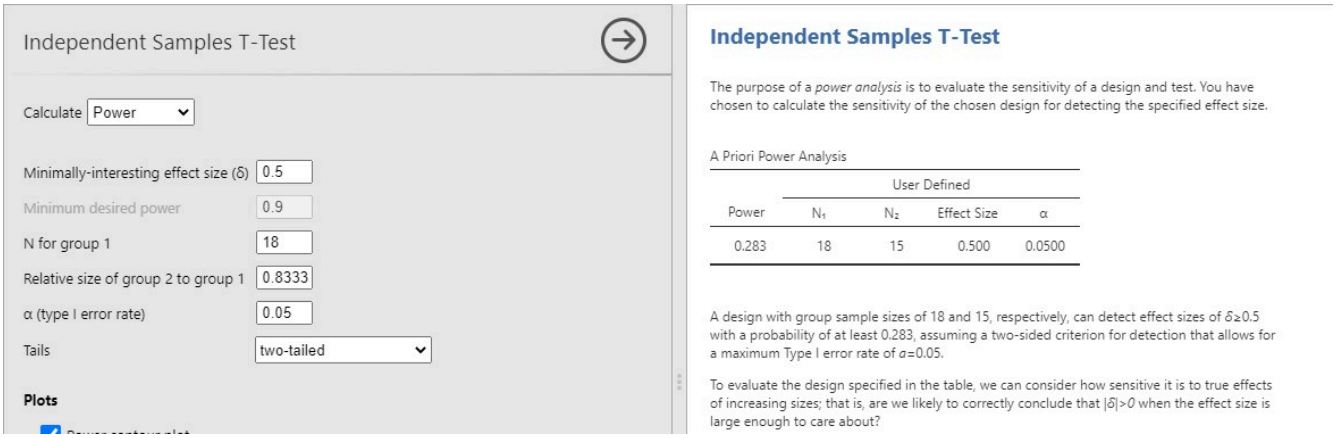
Let's look at the output, below:

Independent Samples T-Test

							95% Confidence Interval			
		Statistic	df	p	Mean difference	SE difference	Lower	Upper		Effect Size
grade	Student's t	2.12	31.0	0.043	5.48	2.59	0.197	10.8	Cohen's d	0.740
	Welch's t	2.03	23.0	0.054	5.48	2.69	−0.0925	11.0	Cohen's d	0.724

If we look at the Student's *t*-test result, our *p*-value is less than our alpha value of .05, a statistically significant result. Like most of the statistics we'll come across, the larger the *t*-statistic (or *F*-statistic, or chi-square statistic…), the smaller the *p*-value will be.

However, remember that we were a bit concerned about assuming homogeneity of the variances, given the small sample size? Therefore, let's look at the Welch's t-test result instead. Here you will see that the *p*-value is .054. This is not statistically significant. Remember the small sample size? This is rather concerning and our study would be underpowered. In fact, if I go into the jpower module in jamovi, I find that we have only 28% power to detect a medium effect size, *d* = .50 (see below) and 60% power to detect a large effect, *d* = .80!



This study really needed to have a larger sample size (and the researcher should have determined, *a priori*, what the effect size of interest was, and what sample size would be required to detect that effect.

Write Up the Results in APA Format

Finally, we can write up the results. We should always include the following information:

1. Description of your research question and/or hypotheses and the test used to assess it.
2. Description of your data – in this case means and standard deviations for each group. If you fail to meet assumptions, you should specify that and describe what test you chose to perform as a result.
3. The results of the inferential test, including what test was performed, the test value and degrees of freedom, *p*-value, and confidence intervals (in this case around the difference between the means). If the effect is significant, then also report effect size.
4. Interpretation of the results, including any other information as needed.

We can write it up something like this:

I used Welch's independent samples *t*-test to assess whether there was a difference in student grades between Anastasia's and Bernadette's classes. Anastasia's students (M = 74.5, SD = 9.00) did not have significantly higher grades than Bernadette's students (M = 69.1, SD = 5.77), $t(23) = 2.03$, $p = .054$, $M_{diff} = 5.48$, 95% CI [-0.09, 11.00].

(Note that I did not report *d*, the effect size, because the result was not significant.)

A note about positive and negative t values

Students often worry about positive or negative t-statistic values and are unsure how to interpret it. Positive or negative t-statistic values simply occur based on which group is listed first. Our t-statistic above is positive because we tested the difference between Anastasia's class (M = 74.53) versus Bernadette's class (M = 69.06) and so Anastasia – Bernadette is a mean difference of 5.48.

However, if our classes were reversed, our mean difference would -5.48 and our t-statistic would be -2.12.

All that is to say, *your positive or negative t-statistic is arbitrary and is just a function of which group is listed first, which is also arbitrary*. So do not fret!

One last note: this positive or negative t-statistic is only relevant for the t-test. You will not get negative values for the F-statistic or chi-square tests!

Dependent Samples *t*-Test

We are going to work with the chico dataset in the lsj-data Data Library. Open the dataset now.

1. Look at the Data

This dataset is is hypothetical data from Dr. Chico's class in which students took two tests: one early in the semester and one later in the semester. Dr. Chico thinks that the first test is a "wake up call" for students. When they realise how hard her class really is, they'll work harder for the second test and get a better mark. Is she right?

Data Set-Up

To conduct the dependent t-test, we first need to ensure our data is set-up properly in our dataset. This requires having two columns: one is our dependent variable score for the participant in one category and the other column is our dependent variable score for the participant in the other category. Each row is a unique participant or unit of analysis. The first few rows of your data should look like this:

	id	grade_test1	grade_test2
1	student1	42.9	44.6
2	student2	51.8	54.0
3	student3	71.7	72.3
4	student4	51.6	53.4
5	student5	63.5	63.8
6	student6	58.0	59.3

In this dataset, what is your independent variable? What is your dependent variable?

Describe the Data

Once we confirm our data is setup correctly in jamovi, we should look at our data using descriptive statistics and graphs. To get descriptive statistics, go to Exploration, then Descriptives. Move the two dependent scores (grade_test1 and grade_test2) over to the Variables box.

Our descriptive statistics are shown below.

Descriptives

	grade_test1	grade_test2
N	20	20
Missing	0	0
Mean	57.0	58.4
Median	57.7	59.7
Standard deviation	6.62	6.41
Minimum	42.9	44.6
Maximum	71.7	72.3

Our overall data consists of 20 cases (students) and the average grade is 56.98 (*SD* = 6.62) at the first test and 58.38 (*SD* = 6.41) at the second test. We have no missing cases, and our minimum and maximum values look accurate; theoretically, student grades should range from 0-100. Lastly, the distribution of data looks fairly normally distributed, although I'm personally a little worried about our small sample size. Before we can proceed with our analyses, we'll need to check our assumptions.

Let's assume that Dr. Chico thinks that the first test is a "wake up call" for students. When they realise how hard her class really is, they'll work harder for the second test and get a better mark. That suggests Dr. Chico thinks students will have better scores on the second test compared to the first test, so it is a *directional prediction*, which calls for a *one-tailed test*.

2: Check Assumptions

As a parametric test, the dependent t-test has the same assumptions as other parametric tests (minus homogeneity of variance because we are dealing with the same people across categories):

1. The *differences in scores* in the dependent variable are **normally distributed**
2. The dependent variable is **interval or ratio** (i.e., continuous)
3. Scores are **independent** across participants

We cannot *test* the second and third assumptions; rather, those are based on knowing your data and research design.

However, we can and should test for the first assumption. Fortunately, the dependent samples t-test in jamovi has two check boxes under "Assumption Checks" that lets us test normality. The same caveat applies as for checking normality for the independent samples *t*-test.

One thing to keep in mind in all statistical software is that we often check assumptions simultaneously to performing the statistical test. However, we should always check assumptions first before looking at and interpreting our results. Therefore, whereas the instructions for performing the test are below, we discuss

checking assumptions here first to help ingrain the importance of always checking assumptions for interpreting results.

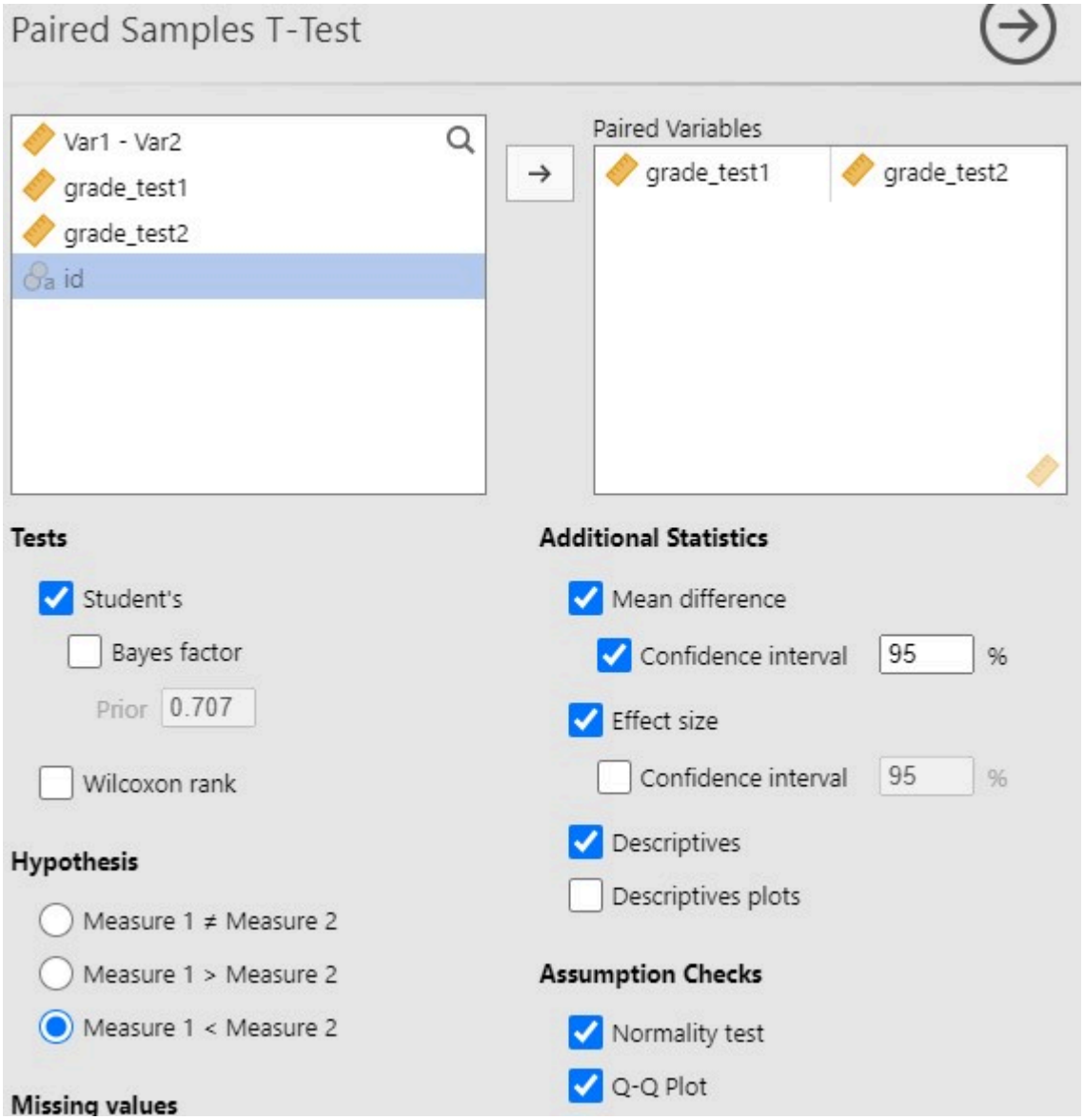
Testing Normality

Notice how our dependent variable is really the difference in scores, and therefore that is what we are testing for normality. The easiest way to do this is by selecting the Paired Samples T-Test (under T-Tests) and entering your variables as described in step 3 below. Under Assumption Checks, you can then select Normality and Q-Q Plot.

If you want to test normality via the Descriptive options, there's an extra step to do. First, you need to calculate a new variable that is the difference in scores. Go to Compute and enter into the formulate box `Var1 - Var2` (in this case, `grade_test1-grade_test2`). Rename it to something meaningful to you. Then you are going to use that variable to test for normality. Select Exploration and Descriptives and choose the new variable that you just created. Select the Shapiro-Wilk test and, under Plots, the Q-Q plot, as well as looking at the density/ histogram plot . The Shapiro-Wilk test was not statistically significant (W = .97, p = .678), suggesting the data are normally distributed – but remember the caveat about interpreting this result when we have a small sample size! Furthermore, the lines are fairly close to the diagonal line in the Q-Q plot (although it's a bit hard to tell because our sample size is small). The histogram shows a roughly normal distribution. I would be hesitant to assume normality because the sample size is so small (N = 20) and I would recommend using the Wilcoxon rank test (the non-parametric equivalent to the dependent samples t-test – more on that later), but for illustrative purposes, let's run the dependent samples t-test anyway.

3. Perform the Test

Select T-Tests and then the Paired Samples T-Test. Note that earlier we suggested that Dr. Chico had a directional prediction, specifically that scores would be higher at test 2 than at test 1. Make sure you select the correct Hypothesis (Measure 1 < Measure 2, assuming you enter the test 1 score first in your list of Paired Variables). You can ask jamovi to compute the Mean difference as well as the 95% confidence for the mean difference and the Effect size. Note that you can get assumption checks again here, if you wish.



4. Interpret Results

Our results will look like this:

Paired Samples T-Test

								95% Confidence Interval			
			statistic	df	p	Mean difference	SE difference	Lower	Upper		Effect Size
grade_test1	grade_test2	Student's t	−6.48	19.0	< .001	−1.40	0.217	−Inf	−1.03	Cohen's d	−1.45

Note. H_a $\mu_{Measure\ 1\ -\ Measure\ 2} < 0$

The *p*-value is less than .05, so the results are statistically significant, and note that Cohen's *d* shows us we have a large effect size.

Why is my CI showing as -Inf? You might be wondering why the lower bound of the confidence interval shows "-Inf." This stands for "negative Infinity." We are getting this value instead of a numerical value because when we use a one-tailed test of a < b (or the other one-tailed scenario where a > b), we no longer care about the situation in which a > b. If a > b we will get a very large p-value. In the same way, with confidence intervals, we no longer care about the lower bound of a – b (or b – a in the other one-tailed scenario), because, what we are most interested in is whether a – b is < 0 and the upper bound of the confidence interval tells us how far beyond zero a – b might be. That is why the lower bound shows as negative infinity.

Write up the Results in APA Format

Finally, we can write up the results:

I conducted a dependent samples *t*-test to assess whether students performed better on the second test compared to the first test. The students performed better on the second test (M = 58.4, SD = 6.41) than they did on the first test (M = 57.0, SD = 6.62), $t(19) = 6.48$, $p < .001$, $d = 1.45$, $M_{diff} = -1.40$, 95% CI [-Inf, -1.03].

What if the result of our inferential test is not significant? In this case, we would report that there was no significant difference between performance on the first and the second test, providing the same statistical values as above, except that we would not need to report the measure of effect size.

Alternatives to the t-Test

Sometimes we cannot run a *t*-test because our data do not meet the assumptions.

Alternative to Independent Samples *t*-Test – Mann-Whitney U

If you have a small sample and you are concerned about meeting the normality assumption, you can use the Mann-Whitney U test. This is the non-parametric equivalent to the independent samples *t*-test. I will not go into specifics, but the idea behind the Mann-Whitney U test is that you take all the values (regardless of group) and rank them. You then sum the ranks across groups and calculate your U statistic and p-value. You interpret the p-value like you normally would, but there are differences in how we report the results because this statistic is based on the *median* not the *mean*.

It is very easy to conduct this test in jamovi – when you select the independent samples *t*-test, simply check the box to run the Mann-Whitney U test. You will interpret the *p*-value in the same way, but note that we report the median, not the mean. With the harpo data, the results look like this:

Independent Samples T-Test

Independent Samples T-Test

		Statistic	p	Mean difference	SE difference		Effect Size
grade	Mann-Whitney U	79.5	0.046	6.00		Rank biserial correlation	0.411

Group Descriptives

	Group	N	Mean	Median	SD	SE
grade	Anastasia	15	74.5	76.0	9.00	2.32
	Bernadette	18	69.1	69.0	5.77	1.36

We would report the results as follows: using the Mann-Whitney U test, there was a statistically significant difference in grades between Anastasia's students (*Mdn* = 76, *n* = 15) and Bernadette's students (*Mdn* = 69, *n* = 18), *U* = 79.50, *p* = .046, r_{pb} = .41.

Alternative to Dependent Samples *t*-Test – Wilcoxon Rank

If we have dependent samples and fail to meet the assumption of normality, especially when we are concerned about small sample sizes, then we perform the Wilcoxon rank test in stead. This is one of the options available after selecting to do Paired Samples *t*-test in jamovi. If we run this test with the Chico data we get the following output:

Paired Samples T-Test

			Statistic	p	Mean difference	SE difference		Effect Size
grade_test1	grade_test2	Wilcoxon W	2.00[a]	< .001	−1.40	0.217	Rank biserial correlation	−0.979

Note. H_a $\mu_{\text{Measure 1 - Measure 2}} < 0$

[a] 1 pair(s) of values were tied

Descriptives

	N	Mean	Median	SD	SE
grade_test1	20	57.0	57.7	6.62	1.48
grade_test2	20	58.4	59.7	6.41	1.43

We could report the results as follows: using Wilcoxon rank test, students' test scores were significantly higher at the second test (*Mdn* = 59.70) than at the first test (*Mdn* = 57.70), $W = 2.00$, $p < .001$, $r_{pb} = .98$.

The note about tied values is not necessary to discuss. It is just telling us one participant had identical values for both test1 and test2 (student15). You can check this yourself in the dataset.

CHAPTER 6: THREE OR MORE MEANS: THE ONE-WAY ANOVA

Portions of this chapter (When and why do we use ANOVA, Why not use multiple *t*-tests, and Relationship between ANOVA and the *t*-test) are based on “Statistics with jamovi” by Dana Wanzer, with some minor changes .

When and Why Do We Use ANOVA?

So far we have learnt about *t*-tests as a way to compare means, but often our research design is more complex and we might have three or more levels of our independent variable, or we might have more than one independent variable. The *t*-test is useful for we have only two means and one independent variable. **Analysis of variance (ANOVA)**, on the other hand, allows us to compare three or more means at the same time, and can be used for one or more independent variables.

The one-way ANOVA is used when we have a continuous dependent variable and a categorical independent variable with three or more categories/levels, in which different participants are in each category.

Why Not Use Multiple *t*-Tests?

Imagine we have three groups to compare: fall, spring, and summer. Why not just perform three separate *t*-tests: fall vs. spring, fall vs. summer, and spring vs. summer?

However, the reason we do not perform multiple *t*-tests is because multiple *t*-tests inflates the type I error rate. If I had performed three separate *t*-tests, set my alpha (type I error rate) at 5% for each test, then each test has a type I error rate of 5%. Because we are running three tests, our alpha actually becomes $1 - (.95^3) = 1 - .857 = 14.3\%$! So now our *familywise* or *experimentwise* error rate is 14.3%, not the 5% we originally set alpha at.

With three groups, that's not so bad, but let's see what happens with more tests we perform:

- **1 test**: $1 - (.95^1) = 1 - .95 =$ **5%**
- **2 tests**: $1 - (.95^2) = 1 - .9025 =$ **9.8%**
- **3 tests**: $1 - (.95^3) = 1 - .857 =$ **14.3%**
- **4 tests**: $1 - (.95^4) = 1 - .814 =$ **18.6%**
- **5 tests**: $1 - (.95^5) = 1 - .774 =$ **22.6%**
- **10 tests**: $1 - (.95^{10}) = 1 - .598 =$ **40.1%**
- **20 tests**: $1 - (.95^{20}) = 1 - .358 =$ **64.1%**

Ouch! 10 tests would have a type I error rate of 40%! That means that if we performed 10 statistical tests (assuming the effect does not exist), then 40% of the results could be significant just by chance, even when the null is true – i.e., they would be false positives. That's not good!

Therefore, we use the one-way ANOVA as one test to see if there is a difference overall. We can also do things to control or limit our familywise error rate, which we'll look at later.

This comic by xkcd provides a great visualization and description for why we need to be super careful about making multiple comparisons.

Relationship Between ANOVA and the *t*-Test

A fun little fact is that an ANOVA with two groups is identical to the *t*-test. That means the F and t statistics are directly related, and you will get the same p-value. For example, imagine you run a *t*-test and get a *t*-statistic of $t(16) = -1.31$, $p = .210$. If you ran it as a one-way ANOVA, you would get an F-statistic of $F(1, 16) = 1.71$, $p = .210$.

(In very simple terms, this essentially comes back to the basic idea that all these tests can be boiled down to:

test statistic = systematic variation/unsystematic variation and that these tests are built on a regression model, but that's another story!)

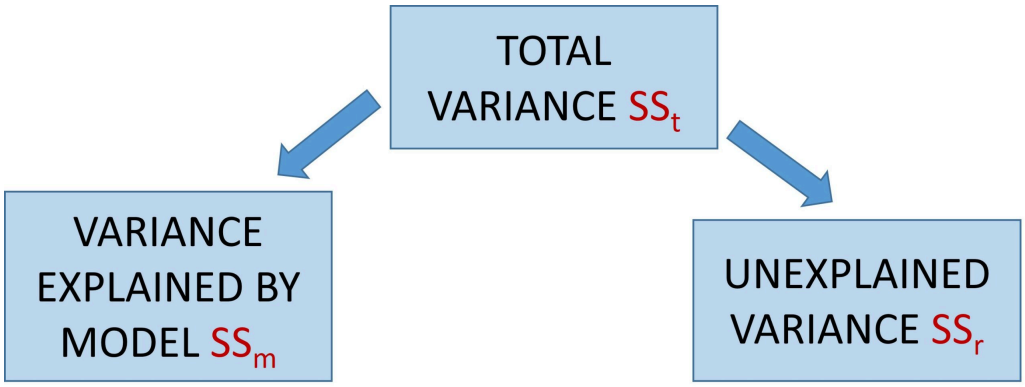
What Does ANOVA Tell Us?

With the one-way ANOVA, the null hypothesis is that all the means are the same and the alternate hypothesis is simply that the means differ. Because ANOVA is an **omnibus test** it just tests for an overall difference between group means, but does not tell us which means differ from each other. If we have three levels, A, B, and C, a significant one-way ANOVA might mean that A and B each differ from C, but there is no difference between A and B; *or* that A and B and C all differ from each of the others; *or* that A and C each differ from B, but there is no difference between A and C; and so on. You might be wondering how useful there is, but there are various kinds of follow-up tests we can use to find out where the differences lie, as you'll learn about later.

Theory of ANOVA

ANOVA is called ANOVA because it analyses/breaks down/partitions the variance in our experiment. In essence, ANOVA takes the overall variability in scores in our experiment and breaks it down into a part that can be explained by the experimental manipulation (systematic variation), or what we'll call "the model," and the part that cannot be explained, that is due to nuisance variables (unsystematic variation).

In other words, the one-way ANOVA, the variance is partitioned as follows:



The total variance, sum of squares total, SS_t, is an indication of how much all the scores in the experiment vary around the grand mean (i.e., the mean of all the scores). The model sum of squares, SS_m, sometimes reported as SS_b (between-groups sum of squares, for the one-way ANOVA) reflects how much the group means vary around the grand mean. And the residual sum of squares, SS_r, reflects how much participant scores vary around their own group means. So, SS_r is the amount of variability that is left over when we use the model (i.e., the group means) to predict scores, compared to when we just use the grand mean to predict scores. SS_r is the variability that we want to minimize. We want participants' scores within each group to be as close as possible to their group mean, to make it easier to detect a difference *between* the groups.

To compute the one-way ANOVA, having completed the SS, we then compute mean square (MS, average variation):

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

We do this for both SS_m and SS_r. Next, we can compute *F* as follows:

$$F = \frac{MS_m}{MS_r}$$

Now, one thing to keep in mind is that MS_m actually reflects variance not just due to the experiment but also it contains some error. There are going to be differences between your groups, not just due to the manipulation of the independent variable, but also due to nuisance variables. For example, we have different participants in each condition, and they are going to score slightly differently from each other even without the manipulation of the independent variable. What we are going to find out is whether the difference between the groups is larger than what we would expect if there were just nuisance variables at play. Note, on the other hand, that MS_r reflects *only* variability due to extraneous variables. So, when we compute *F*, if the null hypothesis is true (i.e., there is no effect of the independent variable on the dependent variable, i.e., no difference between the groups) then both MS_m and MS_r are just capturing variation due to nuisance variables (albeit calculated in different ways), and the *F*-value will be close to 1. On the other hand, if the null hypothesis is false, then MS_m will reflect both variability due to the manipulation of the independent variable and variability due to nuisance variables and so it should be larger than MS_r. As a result, *F* should be greater than 1 (note, this means that we only ever have a one-tailed test with ANOVA). At some point, *F* will be large enough that we will determine it is unlikely to be that large when the null hypothesis is true (when all means in the population are equal), and we claim that we have a significant effect.

Note a caveat here: if there are confounds, then MS_m will also reflect the influence of confounds on the dependent variable.

You'll sometimes see the formula for the one-way ANOVA as follows: $$F = \frac{MS_b}{MS_w}$$ where the "b" refers to "between-groups" (i.e., variation between-groups) and the "w" refers to "within-groups" (i.e., variation within groups).

Follow-Up Tests

As described earlier, ANOVA only tells you that there is a difference(s) somewhere among the level means, but it does not tell you where the differences lie. How we proceed depends on whether or not we have *a priori* predictions regarding specific differences among group means. Let's take an example to work with. The dataset "clinicaltrial" in the lsj-data Data Libary is a hypothetical dataset in which the researchers tested the effectiveness of a new anti-depressant drug, Joyzepam. In the study, participants with moderate-severe depression received either placebo, Joyzepam, or an existing drug, Anxifree. (In addition, half of the participants are also undergoing cognitive behavioural therapy (CBT) and half are not, but we shall ignore this detail for the moment.) The researchers assessed participant mood after three months of taking the medication and their mood gain is scored on a sacle from -5 to +5.

In the Joyzepam example, the researchers probably had a good idea of what they expected to find, based on prior research with Anxifree and perhaps based on some early tests with Joyzepam. They may have clearly-justified hypotheses that participants who received Anxifree will have higher mood scores than participants who received placebo, and that participants who received Joyzepam will have higher mood scores than participants who received Anxifree. In this case, they have some **planned, *a priori* predictions, which can be tested with planned comparisons**. On the other hand, perhaps the researchers really have no idea what to expect. Let's say there is no prior research on either Anxifree or Joyzepam and no reason to predict that they will be any better than placebo. In this case, the researchers' approach is more exploratory and will likely require the use of ***post hoc* tests**. Let's look at each of these situations in a bit more detail. We shall start with *post hoc* tests, because these are often easier for students to understand.

Post hoc Tests

When we conduct ***post hoc* tests, we follow a significant one-way ANOVA with comparisons of all the possible pairs of means**. In other words, continuing our example from above, we would compare placebo to Anxifree, Anxifree to Joyzepam, and placebo to Joyzepam. Note that as you have more groups in your experiment, the number of possible pairwise comparisons increases! Critically, because we are now running multiple (exploratory) tests, we must adjust our alpha level and use a stricter criterion to accept an effect as significant (otherwise we are back to the situation that we were trying to avoid by using ANOVA in the first place – we shall inflate the family-wise error rate and increase the chance of a type I error!).

Corrections for *post hoc* Comparisons

There are several different corrections you can apply to your alpha when correcting for multiple comparisons. Some are more conservative, or strict, such as the Bonferonni adjustment, which essentially divides alpha by the number of tests conducted. This means if you have four groups in your one-way ANOVA, and therefore 6 pairwise comparisons, meaning that the new alpha would be .0083 (.05/6 = .0083)! Of course this is a much more stringent criterion than the original .05. Although some might argue that we should use this criterion, in order to protect against type I error, many researchers agree that, in practice, we should use a correction that is more sensitive (powerful).

You will see, in jamovi, that there are many different types of *post hoc* tests that you can select. They vary in

how conservative or liberal they are. Also, Andy Field (2013) notes that we should consider three criteria when selecting a *post hoc* test:

1. Control for type I error – how conservative or liberal is the test?
2. Statistical power – Is the test sensitive to detecting differences between means?
3. Robustness of the test – how much does it matter if we have violated parametric assumptions?

The table below summarizes the *post hoc* tests that are available in jamovi (see Sauder & DeMars, 2019, for a full summary of 18 *post hoc* tests!).

Test	**Type 1 error control**	**Power**	**Robustness**
No correction	None	High	As for *t*-tests
Tukey	High	Low (more powerful than Bonferroni when many comparisons)	Not advised for unequal sample sizes
Scheffe	High	Low	Similar to ANOVA, different sample sizes permissible
Bonferroni	High	Low (more powerful than Tukey when few comparisons)	Low
Holm	High	Medium	Low

As you can see from table above, the tests available in jamovi are not necessarily robust to unequal sample sizes, while also maintaining high type I error control and high power. If you are limited to using jamovi and have equal sample sizes, can safely assume equal population variances, and have met the assumption of normality, then in most cases you will probably go with Holm, because it has better power than the other options, while controlling for type I error. On the other hand, if you have violated some of the assumptions, you might want to try using SPSS or R to implement one of the procedures summarized in Sauder and DeMars (2019), such as Games-Howell (which is the one that Field, 2013, also recommends).

Planned Comparisons or Contrasts

What about when we have specific predictions about how the groups will differ from each other? This is often considered to be the preferred situation, because then we can conduct fewer tests than all the possible pairwise comparisons and it is easier to achieve power. jamovi gives us several options for running what are called contrasts. **Contrasts** are sets of comparisons among means that enable you to make different kinds of comparisons, depending on your research question and the specific contrast that you choose to use. They are usually specified using something called 'dummy coding' (which is beyond the scope of this book, but can be a lot of fun!). Researchers can specify their own contrasts, or go with one of the standard contrasts provided by jamovi.

Do I Need to Correct for Multiple Comparisons?

When we get into contrasts, students are often confused about whether or not they still need to correct for

multiple comparisons. The answer is: it depends. If your contrasts are orthogonal (and this goes back to how they are specified using the dummy coding, but essentially means that they are independent) then you do not need to correct for multiple comparisons. However, if the contrasts are not orthogonal (i.e., they are related), then you need to apply a correction. The simplest way to do this is to use the Bonferroni correction. Really? I hear you ask! Now, I know that I've suggested earlier that the Bonferroni correction is not ideal because it is very conservative, but if you are only running a small number of comparisons (two or three tests) then power is still going to be moderate. It is only when you are running multiple tests (as occurs when we run *post hoc* tests that we are going to have a substantial loss of power). Note that jamovi will not automatically apply a correction for the non-orthogonal contrasts – you must do this yourself! The table below provides a summary of the contrasts available in the ANOVA menu, what they do, and whether or not they are orthogonal.

Contrast type	**What it does**	**Orthogonal? (If not, apply correction!)**
Deviation	Compares the mean of each level (except a reference category) to the mean of all of the levels (grand mean)	No
Simple	Compares the mean of each level to the mean of the first group*. This type of contrast is useful when there is a control group.	No
Difference (reverse Helmert)	Compares the mean of each level (except the first) to the mean of previous levels.	Yes
Helmert	Compares the mean of each level of the factor (except the last) to the mean of subsequent levels.	Yes
Repeated	Compares the mean of each level (except the last) to the mean of the subsequent level.	No
Polynomial	Looks for a linear trend (level means increase proportionately) and quadratic effect (curve). Useful when levels of independent variable are ordered (e.g., increasing dose of drug).	Yes

Table adapted from Learning Statistics with jamovi (Navarro & Foxtrot, 2019).

*The "first group" is the one that is at the top of the list of "Levels" in the jamovi file when you specify your data variable.

What if jamovi does not have the contrast I want?

If none of the contrasts offered in jamovi fit the hypotheses you would like to test, you can construct your own. If you want to learn about dummy coding and how to create orthogonal contrasts you'll have to go to a different resource (e.g., one of Andy Field's "Discovering Statistics..." series). For a more simple approach, you can use the following procedure:

1. Decide which the pairs of levels you wish to compare and write them down, preferably before you run your study (note, if you do not have specific predictions and find yourself wanting to look at all the possible pairwise comparisons, then you are back to doing *post hoc* tests, and you should follow the instructions in that section!);
2. Select the *post hoc* tests option in jamovi (note, we are not really wanting *post hoc* tests here, but we are using this tool to get our selected planned comparisons);
3. Select "no correction;" and
4. Apply the Bonferroni correction manually (alpha = .05/number of tests)

(assuming you are running only a small subset of the possible pairwise comparisons, this will not have a substantial effect on power) and use your new alpha to interpret the results, *ignoring all the comparisons that you did not previously specify!* This last part is absolutely critical. You must not now look at *all* the pairwise comparisons and select which ones to look at and report based on which ones were significant. This would be *p*-hacking and would inflate your chance of a type I error again!

In Practice: One-Way ANOVA

Let's take a look at how to run the one-way ANOVA using jamovi.

Remember our four steps to data analysis:

1. Look at the data
2. Check assumptions
3. Perform the test
4. Interpret results

1. Look at the Data

For this chapter, we're going to work with example data from lsj-data. Open data from your Data Library in "lsj-data." Select and open "clinicaltrial" (not Clinical Trial 2). This dataset is hypothetical data of a clinical trial in which you are testing a new antidepressant drug called *Joyzepam*. In order to construct a fair test of the drug's effectiveness, the study involves three separate drugs to be administered. One is a placebo, and the other is an existing antidepressant / anti-anxiety drug called *Anxifree*. A collection of 18 participants with moderate to severe depression are recruited for your initial testing. Because the drugs are sometimes administered in conjunction with psychological therapy, your study includes 9 people undergoing cognitive behavioral therapy (CBT) and 9 who are not. Participants are randomly assigned (doubly blinded, of course) a treatment, such that there are 3 CBT people and 3 no-therapy people assigned to each of the 3 drugs. A psychologist assesses the mood of each person after a 3 month run with each drug, and the overall improvement in each person's mood is assessed on a scale ranging from -5 to +5.

Data Set-Up

To conduct the one-way ANOVA, we first need to ensure our data is set-up properly in our dataset. This requires having two columns: one with our continuous dependent variable and one indicating which group the participant is in. Each row is a unique participant or unit of analysis.

Note that in this dataset we actually have two independent variables: `drug` and `therapy`. If we were looking at the effect of `therapy` on `mood.gain` (our DV) then we would only need to perform an independent samples t-test because there are only two groups (no.therapy and CBT). However, if we were looking at the effect of `drug` on `mood.gain`, which is our goal in this chapter, then we would perform a one-way ANOVA because there are three groups (placebo, anxifree, and joyzepam).

Describe the Data

Once we confirm our data is setup correctly in jamovi, we should look at our data using descriptive statistics and graphs. First, our descriptive statistics are shown below. We see that there are 18 cases in our dataset (a bit small, but let's ignore that for now) with no missing data. The mean mood gain was .88 (*SD* = .53) with a minimum

mood gain of .10 and maximum of 1.80. Furthermore, there are 6 people in each of our three conditions in the study so we have a *balanced* research design.

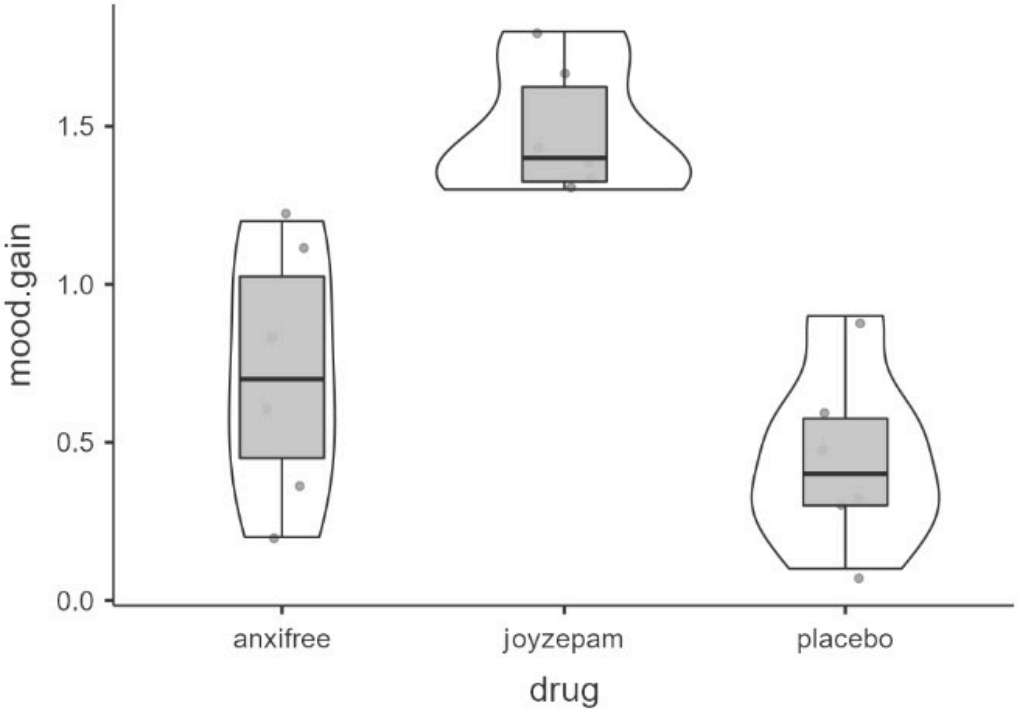
Descriptives

	drug	mood.gain
N	18	18
Missing	0	0
Mean		0.88
Median		0.85
Standard deviation		0.53
Minimum		0.10
Maximum		1.80

Frequencies of drug

Levels	Counts	% of Total	Cumulative %
anxifree	6	33 %	33 %
joyzepam	6	33 %	67 %
placebo	6	33 %	100 %

In addition, we may want to look at the distribution of mood gain across our three conditions. In the Descriptives analysis, we can choose to "split by" `drug` and then ask for a box plot with violin and data points like below. Visually, it seems like joyzepam might be leading to greater mood gain than the other two conditions, but we need to analyze it statistically to know for sure!



Specify the Hypotheses

Our basic research question for the one-way ANOVA is whether there is a difference in mood between the three drugs. Therefore, our null hypothesis would be that there is no difference in mood between the three drugs, and the alternate hypothesis would be that there is a difference in mood between the three drugs.

2. Check Assumptions

As a parametric test, the one-way ANOVA has the same assumptions as other parametric tests:

1. The dependent variable is **normally distributed**
2. Variances in the two groups are roughly equal (i.e., **homogeneity of variances**)
3. The dependent variable is **interval or ratio** (i.e., continuous)
4. Scores are **independent** between groups

We cannot *test* the third and fourth assumptions; rather, those are based on knowing your data.

However, we can and should test for the first two assumptions. Fortunately, the one-way ANOVA in jamovi has three check boxes under "Assumption Checks" that lets us test for both assumptions. Note that all the same caveats to interpreting the assumption checks (which we noted with in the chapter on *t*-tests) also apply in the case of one-way ANOVA (review that section if need be!).

ANOVA is (Somewhat) Robust to Violations

Although we should attend to the assumptions, in general the F-statistic is *robust* to violations of normality and homogeneity of variance. This means that you can still run the one-way ANOVA if you violate the assumptions, but *only when group sizes and variances are equal or nearly equal*. If you have vastly different variances (such as 2:1 ratio or greater) or vastly different group sizes (such as a 2:1 ratio or greater), and especially if one group is really small (e.g., 10 or fewer cases), then your F-statistic is likely to be very wrong. For example, if your larger group has the larger variance, then your F-statistic is likely to be non-significant or smaller than it should be; however, if your larger group has smaller variance, then your F-statistic is likely to be significant or bigger than it should be! Thus, you might either conclude there is no effect when there is indeed one in the population, or vice versa. In these situations, you can use the Walrus package in jamovi (it's an add-on so you will need to add it manually and select the Robust ANOVA), or you can use a non-parametric test (described later in this chapter).

If you select ANOVA in jamovi and then choose the ANOVA (not the One-Way ANOVA option – we shall not use that for this class), you can select the assumption checks Homogeneity test, Normality test, and Q-Q plot.

Assumption Checks

Homogeneity of Variances Test (Levene's)

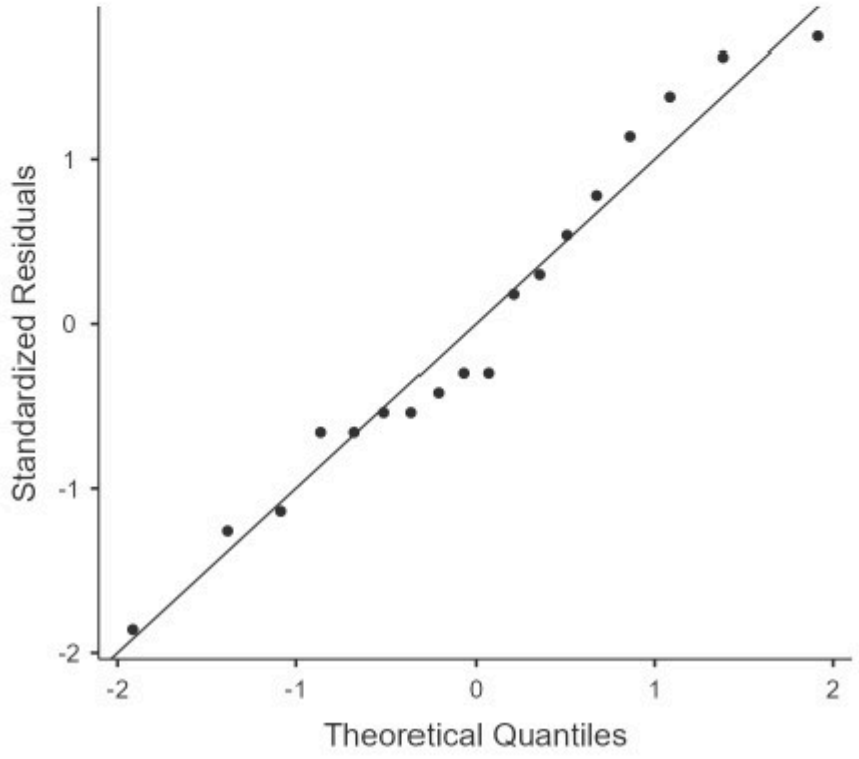
F	df1	df2	p
1.45	2	15	0.266

[3]

Normality Test (Shapiro-Wilk)

Statistic	p
0.960	0.605

Q-Q Plot



The Shapiro-Wilk test was not statistically significant (W = .96, p = .605); it might appear that we can conclude that the data are normally distributed. However, as previously discussed, this test is going to be non-significant even with significant deviations from normality when we have a small sample size, and, when we have such a small sample size, we cannot assume that central limit theorem applies. Similarly, for Levene's test of homogeneity of the variances: it is not significant, but this might be because we have such a small sample size.

The table below shows some options for alternative tests when assumptions are not satisfied (or when we cannot conclude they are satisfied because we have a very small sample size).

	Normality: satisfied	Normality: not satisfied
Homogeneity of the variance: satisfied	One-way ANOVA (using the ANOVA function)	Kruskal-Wallis test or robust ANOVA (Walrus package)
Homogeneity of the variance: not satisfied	Welch's *F*-test (using the one-way ANOVA function)	Kruskal-Wallis test or robust ANOVA (Walrus package)

With our clinical trial data, given that we have such a small sample size, I would go to a non-parametric test, but for illustrative purposes, let's first perform the ANOVA.

3. Perform the Test

I recommend using the ANOVA analysis in jamovi. Do not use the One-Way ANOVA analysis, unless you need Kruskal-Wallis or Welch's *F*-test – the options are too limited for our purposes.

In our example, the researcher did not have specific hypotheses about which groups would differ from each other so we should use *post hoc* tests to obtain all the possible pair-wise comparisons. For illustrative purposes, I am also going to run planned comparisons as well. In this case, I am going to run a simple contrast to compare joyzepam with placebo and anxifree with placebo. Note that we would only run these contrasts if we have specific predictions about which groups will differ from each other, and that we need to apply an appropriate correction if we use non-orthogonal comparisons (see previous section of this chapter!).

1. To perform a one-way ANOVA in jamovi, go to the Analyses tab, click the **ANOVA** button, and choose "ANOVA."
2. Move your dependent variable to the Dependent Variable box and your independent variable to the Fixed Factors box. In this case, move `mood.gain` to the Dependent Variable box and `drug` to the Fixed Factors box.
3. Select ω^2 (omega-squared) for your effect size.
4. Ignore the Model drop-down menu. If you are doing more complicated ANOVAs you will need this. We will ignore it.
5. In the Assumption Checks drop-down menu, select all three options: `Homogeneity test`, `Normality test`, and `Q-Q plot`.
6. To obtain *post hoc* tests, under the Post Hoc Tests option, move the independent variable over to the box on the right, and select both Holm, Bonferroni, and No correction (note, we would not normally select multiple *post hoc* tests, but this will allow you to see how they vary in their consequences for the *p*-value). In addition, check the box for Cohen's *d*, to obtain a measure of effect size. To obtain contrasts, under "contrasts," click on the drop-down menu to the right of the independent variable in question and select the contrast you would like. For this example, choose a simple contrast.
7. In the Estimated Marginal Means drop-down menu, move your IV `drug` to the Marginal Means box and select `Marginal means plots`, `Marginal means tables`, and `Observed scores`, in addition to the pre-selected `Equal cell weights`.

4. Interpret Results

Once we are satisfied we have satisfied the assumptions for the one-way ANOVA, we can interpret our results.

ANOVA

ANOVA - mood.gain

	Sum of Squares	df	Mean Square	F	p	ω^2
drug	3.45	2	1.7267	18.6	< .001	0.662
Residuals	1.39	15	0.0928			

[3]

The result shows a significant ANOVA and we would follow-up by looking at the *post hoc* tests or planned comparisons as appropriate. Below is the table showing the results of the Post Hoc Tests in jamovi.

Post Hoc Tests

Post Hoc Comparisons - drug

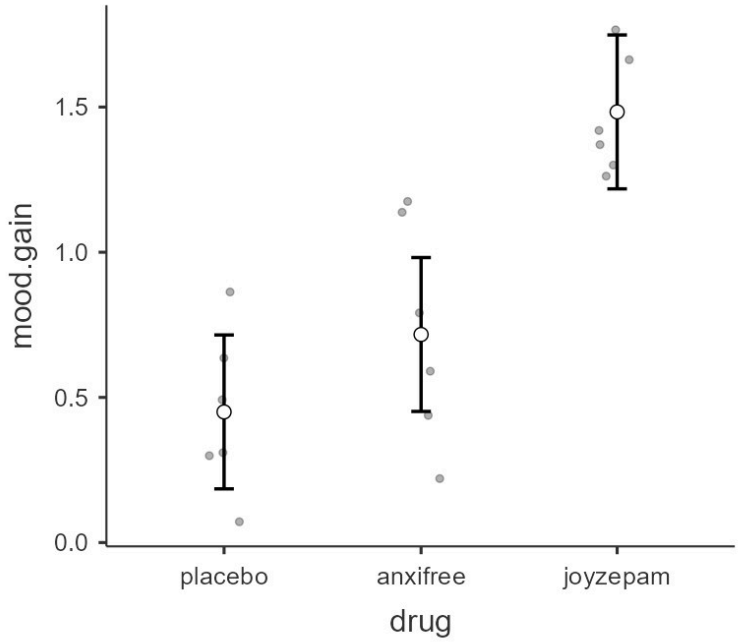
Comparison										
drug		drug	Mean Difference	SE	df	t	p	$p_{bonferroni}$	p_{holm}	Cohen's d
placebo	-	anxifree	−0.267	0.176	15.0	−1.52	0.150	0.451	0.150	−0.875
	-	joyzepam	−1.033	0.176	15.0	−5.88	< .001	< .001	< .001	−3.392
anxifree	-	joyzepam	−0.767	0.176	15.0	−4.36	< .001	0.002	0.001	−2.517

Note. Comparisons are based on estimated marginal means

Remember that, normally, we should select the appropriate *post hoc* test based on our data, rather than running three different tests. However, by running these three tests, you can see how they differ in the output they provide. You will notice that the *p*-value changes according to whether you look at the first p (which is for the comparison with no correction), the Bonferroni, or the Holm test. Bonferroni results in the largest *p*-value and is the most conservative. In our example, Holm and no correction result in the same *p*-value. In some scenarios, Holm will result in smaller *p*-values than no correction (because it involves some correction and so is less powerful than applying low correction). If we had elected to use Holm, we would note that there is a significant difference between placebo and joyzepam, and also between anxifree and joyzepam, but not between placebo and anxifree. We can look at the graph to verify the direction of the effect.

Estimated Marginal Means

drug



Estimated Marginal Means - drug

			95% Confidence Interval	
drug	Mean	SE	Lower	Upper
placebo	0.450	0.124	0.185	0.715
anxifree	0.717	0.124	0.452	0.982
joyzepam	1.483	0.124	1.218	1.748

You can see from the marginal means that it is that joyzepam show higher mood gain scores than anxifree and placebo.

Let's say, instead, that you had some specific predictions that both joyzepam and anxifree would result in higher mood gain scores than placebo. In this case, you have planned comparisons, and so you would look at the contrasts results in jamovi. The results for the simple contrasts are below.

Contrasts

Contrasts - drug

	Estimate	SE	t	p
anxifree - placebo	0.267	0.176	1.52	0.150
joyzepam - placebo	1.033	0.176	5.88	< .001

As described in the previous section of this chapter, simple contrasts are non-orthogonal comparisons. Therefore, we would need to apply a correction to the *p*-value ourselves. You can use the Bonferroni correction, dividing α by two, because we conducted two tests for the contrasts. Therefore, $\alpha = .05/2 = .025$. By this criterion, joyzepam results in significantly higher mood gain scores than placebo, but anxifree does not.

Write up the results

In APA format, and reporting on the results of the *post hoc* tests, we could say something like this:

> There was a significant difference in mood gain across the three drug conditions, $F(2, 15) = 18.61$, $p < .001$, $\omega^2 = .66$. Participants who received joyzepam had higher mood gain scores ($M = 1.48$, $SE = 0.12$, 95% CI [1.22, 1.75]) than participants who received either anxifree ($M = 0.72$, $SE = 0.12$, 95% CI [0.45, 0.98]), or placebo ($M = 0.45$, $SE = 0.12$, 95% CI [0.19, 0.72]), $t(15) = 4.36$, $p_{holm} = .001$, $d = 2.52$, and $t(15) = 5.88$, $p_{holm} < .001$, $d = 3.39$, respectively. However, there was no significant difference between participants who received anxifree compared to those who received placebo, $t = 1.52$, $p_{holm} = .15$.

If we instead wanted to report the contrasts (note, we would only do this if we had specified these as planned comparisons *a priori*), we could report as follows (the contrasts do not give me a Cohen's *d* so I can obtain that by looking at the *post hoc* test results):

> There was a significant difference in mood gain across the three drug conditions, $F(2, 15) = 18.61$, $p < .001$, $\omega^2 = .66$. I used planned comparisons to compare mean mood gain scores for participants in the anxifree and joyzepam groups to the placebo group, applying the Bonferroni correction ($\alpha = .05/2 = .025$). Participants who received joyzepam had higher mood gain scores ($M = 1.48$, $SE = 0.12$, 95% CI [1.22, 1.75]) than participants who received placebo ($M = 0.45$, $SE = 0.12$, 95% CI [0.19, 0.72]), $t = 5.88$, $p < .001$, $d = 3.39$, but participants who received anxifree ($M = 0.72$, $SE = 0.12$, 95% CI [0.45, 0.98]) did not have higher mood gain scores than participants who received placebo, $t = 1.52$, $p = .15$.

Note that in this situation, whether we use the *post hoc* tests or the contrasts, the results turn out the same. However, this will not always be the case. Therefore, you should always decide prior to conducting your analyses whether you are going to look at the *post hoc* tests or whether you have specific contrasts of interest (based on prior research and/or theory) and will use planned comparisons, i.e., contrasts.

Finally, what would we write up if the ANOVA result itself was not significant? In that case, we would *not* report follow-up tests (whether we had planned comparisons or were going to use *post hoc* tests). Nor would we describe which group scored higher than which. However, we would still report descriptive statistics. So, we might write something like this (note the numbers below are fabricated for demonstration purposes!):

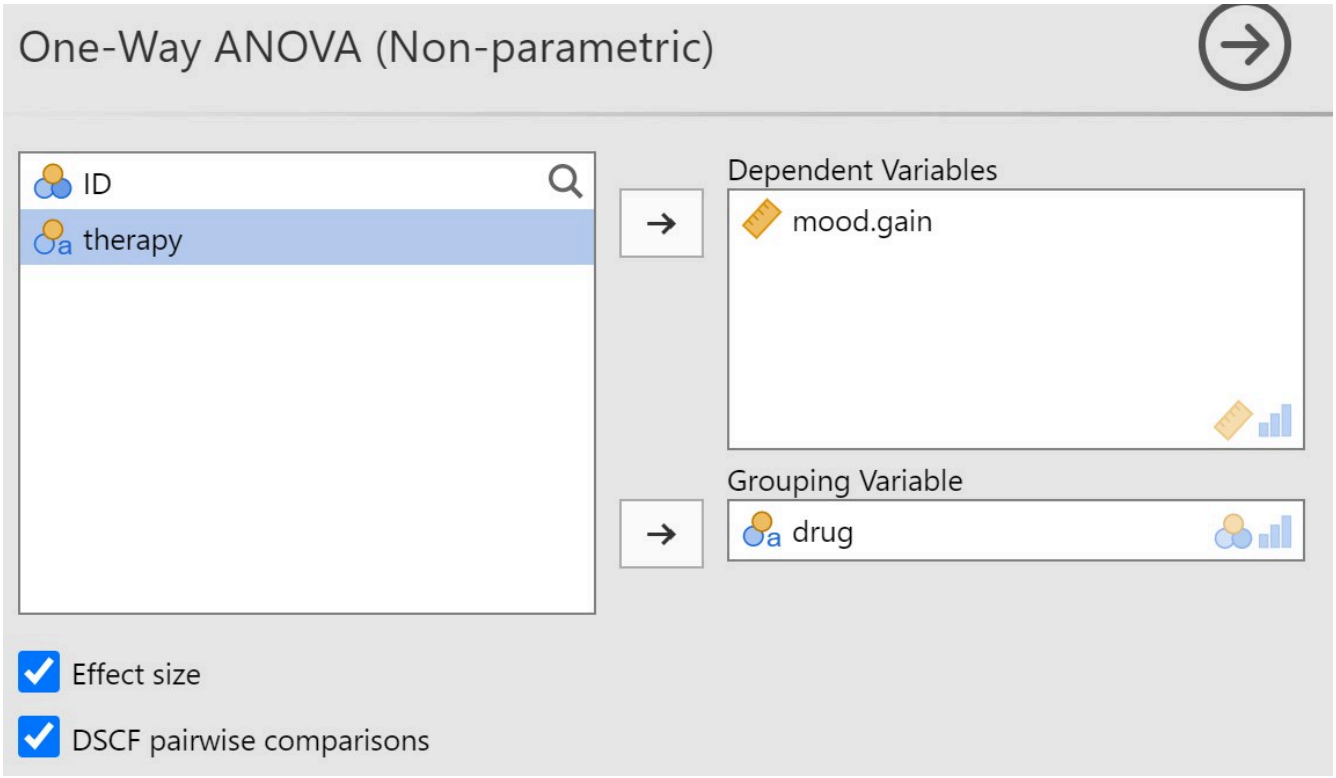
> There was no significant difference in mood gain across the three drug conditions, $F(2, 57) = 0.48$, $p = .620$ (anxifree: $M = 1.52$, $SD = 0.31$; joyzepam: $M = 1.60$, $SD = 0.29$; placebo: $M = 1.55$, $SD = 0.30$).

Alternatives to the One-Way ANOVA

In the previous section, we noted some possible alternatives to ANOVA if we violate the assumptions. Here's the table showing those alternatives, again:

	Normality: satisfied	**Normality: not satisfied**
Homogeneity of the variance: satisfied	One-way ANOVA (using the ANOVA function)	Kruskal-Wallis test or robust ANOVA (Walrus package)
Homogeneity of the variance: not satisfied	Welch's *F*-test (using the one-way ANOVA function)	Kruskal-Wallis test or robust ANOVA (Walrus package)

With our clinical trial data, given that we have such a small sample size, I would go to a non-parametric test, which in this case would be the Kruskal-Wallis test. It's easy to obtain this in jamovi by selecting ANOVA and then, under Non-Parametric, One-Way ANOVA Kruskal-Wallis. You can also ask jamovi to give you the effect size and some pairwise comparisons.



One-Way ANOVA (Non-parametric)

Kruskal-Wallis

	χ^2	df	p	ε^2
mood.gain	12.1	2	0.002	0.710

Dwass-Steel-Critchlow-Fligner pairwise comparisons

>

Pairwise comparisons - mood.gain

		W	p
placebo	anxifree	1.70	0.450
placebo	joyzepam	4.10	0.011
anxifree	joyzepam	4.09	0.011

Write this up the same way as you would write-up the regular one-way ANOVA, but reporting χ^2 (chi-squared) instead of *F*, ε^2 (epsilon-squared) for effect size for the ANOVA, and *W* instead of *t* for your *post hoc* tests.

Finally, if you need to run Welch's test, select ANOVA and then one-way ANOVA, and under Variances click the box Don't assume equal (Welch's). You will see that the result is a bit limited. You cannot get a measure of effect size for the ANOVA and you cannot run contrasts. If you need these, your best bet will be to go back to the regular ANOVA menu to obtain them. You could report the Welch's one-way ANOVA but then report the effect size and contrasts from the ANOVA results. It is not ideal but it will get you the information you need. For the *post hoc* tests, there are some other options presented that we did not see in the regular ANOVA menu. In particular, if you have unequal variances, you will want to use the Games-Howell test.

CHAPTER 7: REPEATED MEASURES ANOVA

Portions of this chapter (In practice: Repeated measures ANOVA) are experts from Dana Wanzer's "Statistics with jamovi" licensed under a Creative Commons BY-SA (CC BY-SA) license version 4.0, with minor changes.

Repeated Measures Design

A repeated measures design involves all participants completing all levels of the independent variable. This includes designs where time is the independent variable, and participants are measured on the same dependent variable over time. Why use it? There are two key benefits: one is to increase the sensitivity or power of our experiment. By having the same participants take part in each level of the independent variable, we control for participant variables (e.g., age, personality, etc. – these should be the same across levels of the independent variable). Participant variables are usually a major source of error variance or unsystematic variation. By controlling for them, it should be easier to detect the experimental effect. A second key benefit is economy: all other things being equal, we do not need as many participants for a repeated measures design as a between-subjects design (because of the control for participant variables). At the same time, there are some limitations to this design: by using the same participants in all levels of the independent variable, the experiment becomes longer, with a greater risk of fatigue and attrition. In addition, we have to consider the possibility of increased practice effects and demand characteristics. We have to balance these drawbacks with the benefits of the repeated measures design.

We can go some way to addressing the limitations by using counterbalancing.

Counterbalancing

Counterbalancing involves changing the order of the design for different participants in order to balance out the effects of practice, learning, and so on, across participants. **Complete counterbalancing** means that you create all the possible combinations of order for the design. Let's say we have three levels for our independent variable: level A, level B, and level C. All the possible orders are: ABC ACB BCA BAC CAB CBA. To implement this in our experiment, each participants would be exposed to one of these possible orders. In our statistical analyses, we might collapse across order (i.e., ignore it) and just look at the effect of the independent variable, or, we might analyze order effects. The problem with complete counterbalancing is that it creates complex designs when you have three or more levels to your independent variable.

In **partial counterbalancing** we use only a subset of the possible orders. One way to do this is by using a Latin square design. A Latin square design draws on the principles of the Latin square. The Latin square is an array of numbers where each number appears once in each row and once in each column. In the Latin square design, our square follows the rules: 1) the number of orders = the number of conditions; and 2) each condition appears only once in each row and once in each column. So, for example, a Latin square for a three-level design might look like this:

A	B	C
B	C	A
C	A	B

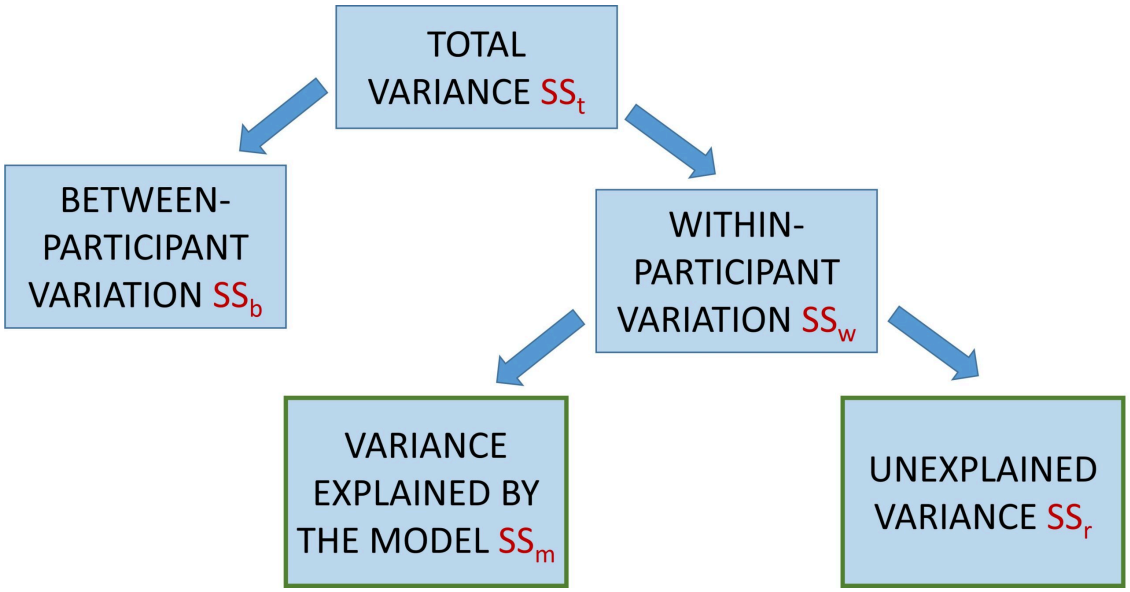
The orders we would use in our experiment would therefore be: ABC, BCA, and CAB (thus only three different orders of levels, instead of six!).

Randomizing

Sometimes, instead of counterbalancing, we might randomize the order of conditions, creating a new random order for each participant.

Theory of One-Way Repeated Measures ANOVA

MS_m and MS_r are computed slightly differently from in the one-way ANOVA (between-subjects). Specifically, the variance is partitioned as follows:



Within-participant variation is decomposed into two types of variance. Our experimental effect (our model sum of squares – variance explained by the model) is now part of the within-participant variation, because all participants complete all conditions in the experiment. So, some of the variability within participants is due to our manipulation of the independent variable (SS_m) and some of it is due to random variation within participants over the levels of the independent variable (SS_r – "error").

Without getting into the calculations for each of these types of variance, they can be described as follows:

- SS_t – variability of scores around the grand mean in the experiment
- SS_w – variation of individual participant scores
- SS_m – variation of level means around the grand mean
- SS_r – how much variation cannot be explained by the model

The one-way repeated measures of ANOVA, like the one-way ANOVA (between-subjects) that you learned about in the previous chapter, then uses the following equation for the *F*-ratio (where MS = SS/df):

$$F = \frac{MS_m}{MS_r}$$

Assumptions of Repeated Measures Designs

In between-subjects ANOVA, we assume independence (scores for different levels of the independent variable do not influence each other). In the repeated measures design, because the same participants are in all conditions, scores across levels of the independent variable are correlated and so we have violated the assumption of independence. Therefore, we make an extra assumption: the **assumption of sphericity**. This assumption specifies that the correlation of scores across conditions should not differ. It is tested by testing whether the variances in the differences between conditions are significantly different from one another. To do this, we use Mauchly's test. Let's say we have three levels in our design, A, B, and C. Mauchly's tests whether the the variance of the differences between groups A and B and the variance of the differences between groups A and C and the variance of the differences between groups C and A are significantly different from one another. We want them to be the same (in the population, and not significantly different in our sample), i.e., $\text{variance}_{A\text{-}B} = \text{variance}_{A\text{-}C} = \text{variance}_{B\text{-}C}$.

If we violate the assumption of sphericity (i.e., if the variances are significantly different from one another), we can apply one of three corrections: Greenhouse-Geisser, Huynh-Feldt, or Lower-bound (the third one is not available in jamovi but is in some other software packages). One problem with Mauchly's is the same as as arose with our tests of normality and homogeneity of the variance that we discussed in previous chapters: for small samples, the test will not be significant (due to lack of power), even when the assumption is violated; for very large samples, it will be significant even when there are only small differences among the differences in the variances. One approach is to err on the side of caution and apply the Greenhouse-Geisser correction whenever reporting the results of a repeated-measures ANOVA. We shall look at this further in the next section.

In Practice: Repeated Measures ANOVA

1. Look at the Data

Let's run an example with data from lsj-data. Open data from your Data Library in "lsj-data." Select and open "Broca's aphasia."

This dataset contains hypothetical data in which six patients suffering from Broca's Aphasia (a language deficit commonly experienced following a stroke) completed three word recognition tasks. On the first (speech production) task, patients were required to repeat single words read out aloud by the researcher. On the second (conceptual) task, designed to test word comprehension, patients were required to match a series of pictures with their correct name. On the third (syntax) task, designed to test knowledge of correct word order, patients were asked to reorder syntactically incorrect sentences. Each patient completed all three tasks. The order in which patients attempted the tasks was counterbalanced between participants. Each task consisted of a series of 10 attempts. The number of attempts successfully completed by each patient are provided in the dataset.

Data Set-Up

To conduct the repeated measures ANOVA, we first need to ensure our data is set-up properly in our dataset. This requires multiple columns, one for each condition or time measurement, with the values indicating the measurement of the DV for that condition or time. Each row is a unique participant or unit of analysis.

So for our broca dataset, we have our Participant column indicating their participant number and then one column for each of the three word recognition tasks (speech, conceptual, syntax), with their scores on the knowledge test indicating the dependent variable for each condition.

	Participant	Speech	Conceptual	Syntax
1	1	8	7	6
2	2	7	8	6
3	3	9	5	3
4	4	5	4	5
5	5	6	6	2
6	6	8	7	4

In the data above, what is your independent variable and what are the levels of the independent variable? What is your dependent variable?

Describe the Data

Once we confirm our data is setup correctly in jamovi, we should look at our data using descriptive statistics and graphs. First, our descriptive statistics are shown below (obtain these by going to Exploration, Descriptives, and

selecting all three levels of the independent variable as "Variables.") We see that there are only six cases total (oof, really small data set!) and the average test score on each of the three conditions. It appears participants did best on the speech condition, but we'll need to run our repeated measures ANOVA to know for sure.

Descriptives

	Speech	Conceptual	Syntax
N	6	6	6
Missing	0	0	0
Mean	7.17	6.17	4.33
Median	7.50	6.50	4.50
Standard deviation	1.47	1.47	1.63
Minimum	5	4	2
Maximum	9	8	6

Hypotheses

Let's say we have predicted that there is an effect of task type on word recognition scores.

2. Check Assumptions

As a parametric test, the repeated measures ANOVA has the same assumptions as other parametric tests (minus the assumption of independence, because all participants participate in all conditions):

1. The dependent variable is **normally distributed**
2. Variances in the two groups are roughly equal (i.e., **homogeneity of variances**); in repeated measures ANOVA this is called the assumption of **sphericity**
3. The dependent variable is **interval or ratio** (i.e., continuous).

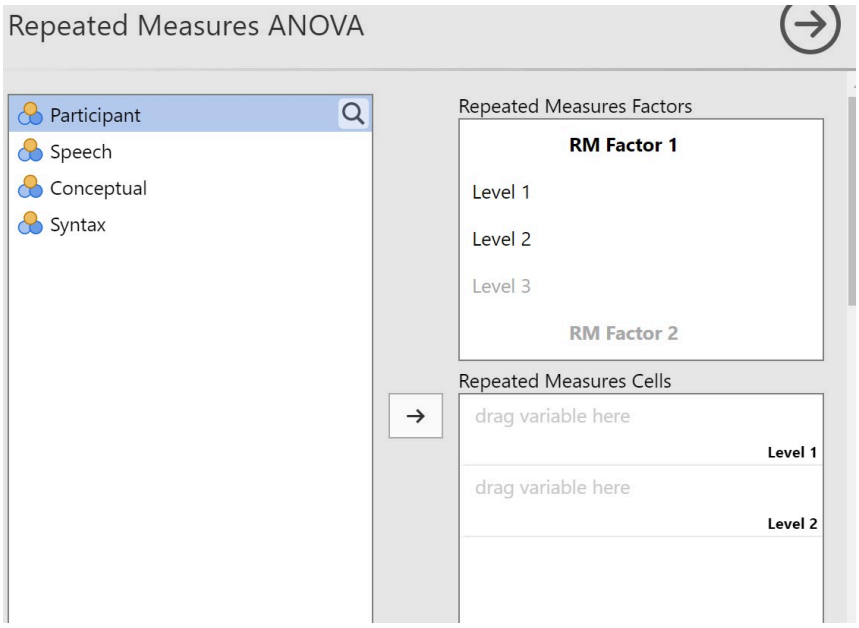
To check assumptions 1 and 2, we need to run the repeated measures ANOVA in jamovi.

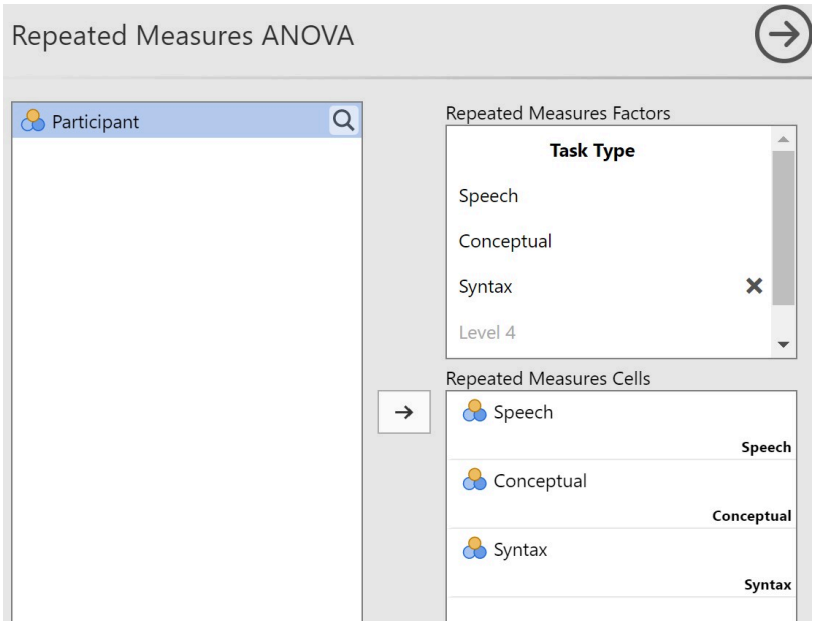
To begin the repeated measures ANOVA in jamovi

1. Select ANOVA, and then Repeated Measures ANOVA.
2. You now need to tell jamovi what your repeated measures "factor" (independent variable) is by specifying its names and the levels in the box called Repeated Measures Factors. Click where it says "RM Factor 1" and type the name of your independent variable (in this example, I am calling it "Task Type"). Next, rename hte three levels of task by typing Speech, Conceptual, and Syntax,

where it says “Level 1,” “Level 2,” and “Level 3” below.

3. Click and drag your three variables to the Repeated Measures Cells box (or select them all by clicking on each while holding down the Shift key and click on the little arrow to move them to the Repeated Measures Cells box).





Now you can select the assumptions.

Testing Normality

We cannot test normality using all four methods we previously learned about because in this case we are checking normality of the residuals. You don’t need to know what that means yet (we’ll discuss it more when we go over regression).

To check normality for the repeated measures ANOVA, we have to go run the repeated measures ANOVA in

jamovi. To test normality, under 'Assumption Checks' we select Q-Q plot. It is currently the only option we have available for testing normality in the repeated measures ANOVA. As with the Q-Q plot we looked at for the one-way ANOVA, we are looking for the dots to be close to the straight line.

There are alternatives to the repeated measures ANOVA if we are concerned about violating the assumption of normality (and/or when we have small sample sizes). The non-parametric equivalent is Friedman's test (which we shall look at briefly in the next section).

Testing Sphericity

The sphericity assumption is essentially the repeated measures ANOVA equivalent of homogeneity of variances. Sphericity means there is equality of variances of the *differences* between treatment levels. For example, if there are three groups, then the difference in all three pairs of differences (1-2, 1-3, 2-3) need to have approximately equal variances. You only need to care about sphericity when there are at least three conditions, which is why we did not talk about this with the dependent t-test.

Fortunately, like the other assumption checks, testing for sphericity is as simple as a checkbox in jamovi. Check the box `Sphericity tests` under the Assumption Checks drop-down menu. That produces the following output:

Assumptions

Tests of Sphericity

	Mauchly's W	p	Greenhouse-Geisser ε	Huynh-Feldt ε
Task	0.849	0.720	0.868	1.00

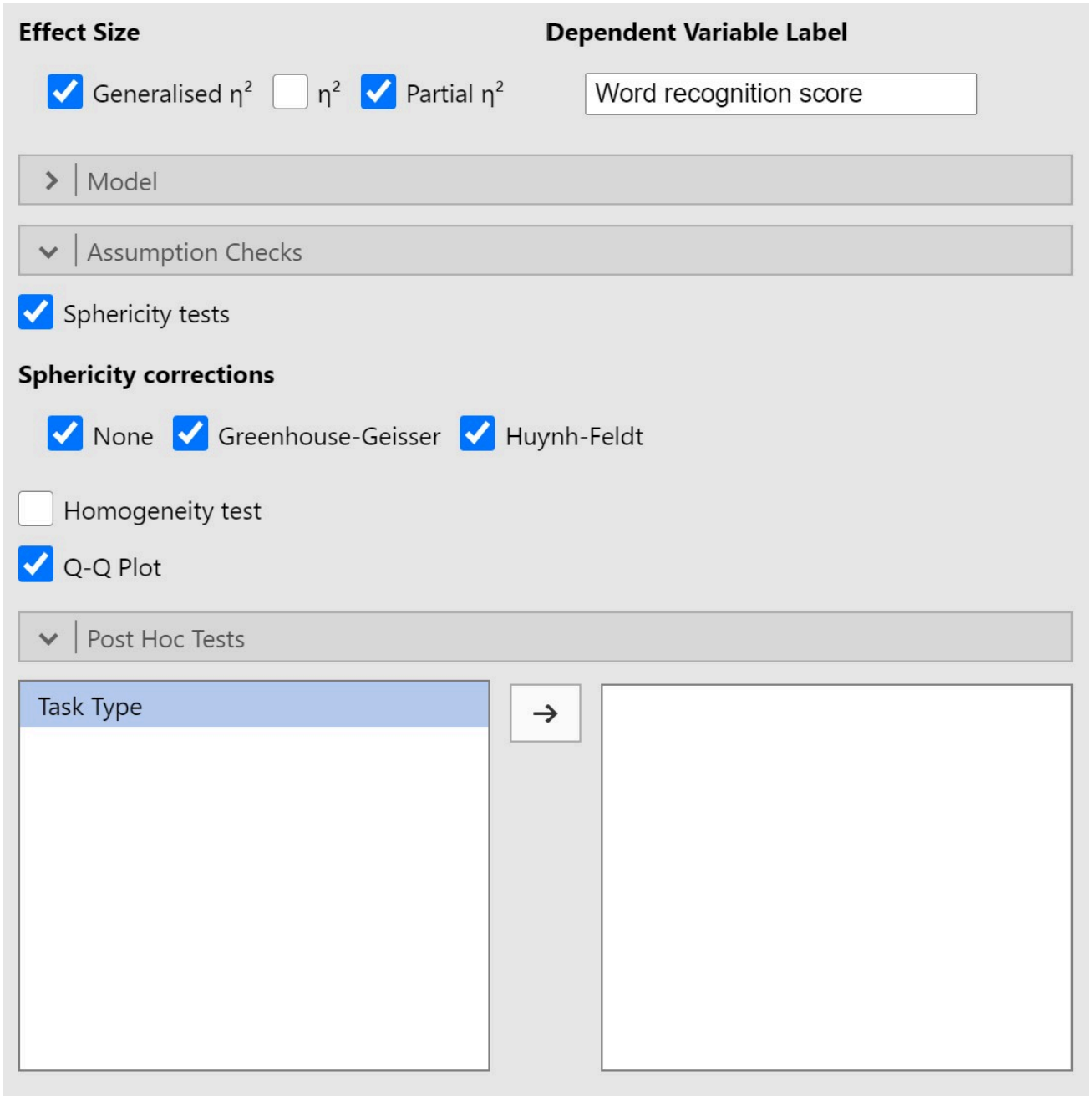
According to this output, we have not violated the sphericity assumption (because Mauchly's test was not significant, $p > .05$). However, as noted earlier, we have a small sample size and so this test will be underpowered. I recommend using the Greenhouse-Geisser correction when you run the repeated measures ANOVA (see below).

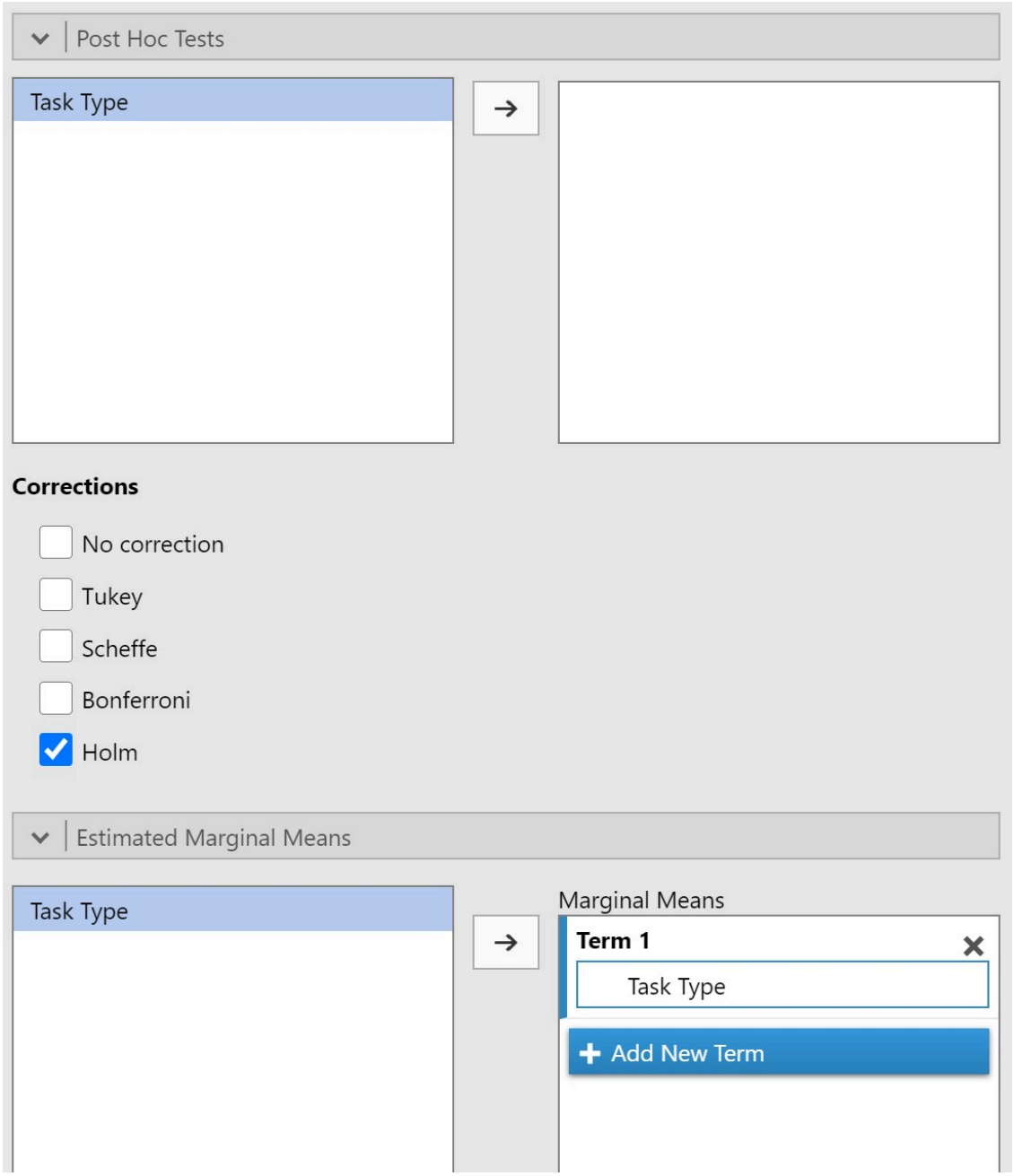
3. Perform the Test

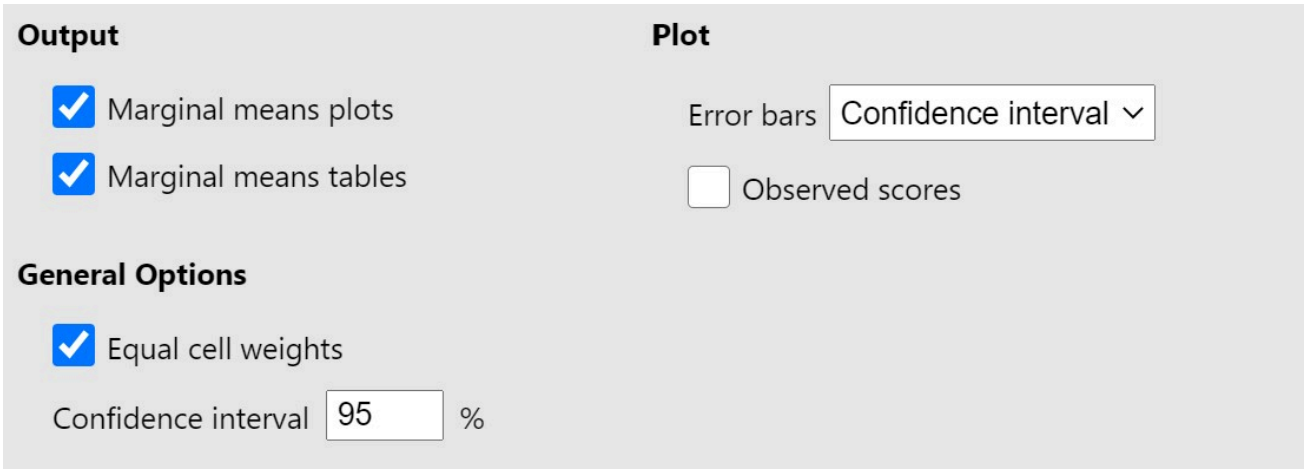
1. Select ANOVA, and then Repeated Measures ANOVA.
2. You now need to tell jamovi what your repeated measures "factor" (independent variable) is by specifying its names and the levels in the box called Repeated Measures Factors. Click where it says "RM Factor 1" and type the name of your independent variable (in this example, I am calling it "Task Type"). Next, rename the three levels of task by typing Speech, Conceptual, and Syntax, where it says "Level 1," "Level 2," and "Level 3" below.
3. Click and drag your three variables to the Repeated Measures Cells box (or select them all by clicking on each while holding down the Shift key and click on the little arrow to move them to the Repeated Measures Cells box).
4. Select Generalised η^2 (eta-squared) as your measure of effect size.

5. Rename your Dependent Variable Label to match what is being measured across all conditions or time points. In this case, in the three conditions we are measuring their scores on the word recognition task. I will just call this “Word recognition score.”
6. Under Assumption Checks, select Sphericity tests, and, for illustrative purposes, select all three boxes under the Sphericity corrections.
7. Under Post Hoc Tests, move Task Type to the box on the right and select the Holm correction.
8. Under the Estimated Marginal Means drop-down menu, move Task Type to the Marginal Means box and select both Marginal means plots and tables.

Steps 4-8 should appear as follows:




Post Hoc Tests
Task Type
Corrections
No correction
Tukey
Scheffe
Bonferroni
Holm
Estimated Marginal Means
Task Type
Marginal Means
Term 1
Task Type
Add New Term


Output
Marginal means plots
Marginal means tables
General Options
Equal cell weights
Confidence interval 95 %
Plot
Error bars Confidence interval
Observed scores

Note: you will notice that jamovi does not give us an option to run contrasts for repeated measures ANOVAs. Therefore, if you would like to run planned comparisons (based on *a priori* predictions about which specific level means will differ), then I recommend you follow the instructions under "What if jamovi does not have the contrast I want?" in the Follow-up tests section of Chapter 6.

4. Interpret Results

Let's take a look at the results. The first part of the output will look like this:

Repeated Measures ANOVA

Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η^2_G	η^2_p
Task Type	None	24.8	2	12.39	6.93	0.013	0.414	0.581
	Greenhouse-Geisser	24.8	1.74	14.26	6.93	0.018	0.414	0.581
	Huynh-Feldt	24.8	2.00	12.39	6.93	0.013	0.414	0.581
Residual	None	17.9	10	1.79				
	Greenhouse-Geisser	17.9	8.68	2.06				
	Huynh-Feldt	17.9	10.00	1.79				

Note. Type 3 Sums of Squares

[3]

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G	η^2_p
Residual	17.1	5	3.42				

Note. Type 3 Sums of Squares

You'll notice that jamovi provides you both a Within Subjects Effects table and Between Subjects Effects table. However, we only have a within-subjects effect (Task Type). With the repeated-measures ANOVA (which only has within-subjects IVs), we do not have any between-subjects independent variables so there is nothing to compute. We can ignore it in this case. It will be useful when we conduct a mixed factorial ANOVA, with both between-subjects and within-subjects effects (see later chapter mixed ANOVA).

Therefore, the Within Subjects Effects table is what we look at. Note that there are three rows for the effect of Task Type. This is because we requested no sphericity correction as well as both the Greenhouse-Geisser and Huynh-Feldt corrections. We can see that the df, MS, and *p*-value changes according to which correction we apply. The *p*-value is largest for the Greenhouse-Geisser correction, because this is the most conservative test. As mentioned in the previous section, I recommend using the Greenhouse-Geisser by default. We can see that the overall effect of Task Type is statistically significant ($p = .018$). Therefore we can look at our *post hoc* tests.

Another thing you may notice is that I asked for η^2_p (partial eta-squared) as well as η^2_G (generalised eta-squared). The η^2_G value is slightly smaller and is actually the preferred measure of effects size for repeated measures designs (Bakeman, 2005).

Let's look at the *post hoc* test results.

Post Hoc Tests

Post Hoc Comparisons - Task Type

Comparison			Mean Difference	SE	df	t	p_{holm}
Task Type		Task Type					
Speech	-	Conceptual	1.00	0.683	5.00	1.46	0.203
	-	Syntax	2.83	0.910	5.00	3.11	0.079
Conceptual	-	Syntax	1.83	0.703	5.00	2.61	0.096

The Holm *post hoc* tests show that there were no significant differences among any of the means. How can this be, when the ANOVA was significant? Did we make an error in our analysis? We did not make an error in the analysis. This can happen, because ANOVA and *post hoc* tests are different tests. ANOVA is looking at how, overall, the level means vary around the grand mean in the experiment, whereas *post hoc* tests are making comparisons between pairs of means. It might be, for example, that if we collapse across two levels and compare them to the third level, there would be a significant difference, but we cannot conduct this analysis with *post hoc* tests. Remember as well that Holm corrects for the inflation of the family-wise error rate and so is somewhat conservative (if you run the *post hoc* tests again with no correction, you'll see that two of the three comparisons are significant). Unfortunately, in this scenario, we may not have sufficient power (remember small sample size!) to detect where the differences between conditions lie.

Last, we can look at the Estimated Marginal Means – Task Type table to see the group means for reporting purposes.

Write Up the Results in APA Style

We can write this up in APA style similar to the one-way ANOVA.

> A repeated measures ANOVA was performed examining how three tasks affected word recognition in patients suffering from Broca's Aphasia. Task type significantly affected word recall (applying the Greenhouse-Geisser correction), F (1.74, 8.68) = 6.93, p = .018, η^2_G = .41. *Post hoc* tests, applying the Holm correction, indicated no significant difference between any of the pairs of level means (speech task: M = 7.17, SE = .60; syntax task: M = 4.33, SE = .67; and conceptual task: M = 6.17, SE = .60), all $p_{holm} > .05$.

This is an interesting situation where the ANOVA gives a significant effect, but the *post hoc* tests are not significant. This can happen and is not too surprising in this case because the sample size is very small. Given the small sample size , it would have been advisable to conduct a non-parametric test. Let's do that next.

Alternatives to the Repeated Measures ANOVA

If we have a repeated measures with a single independent variable, but a small sample size and/or have violated the assumption of normality, it is advisable to use the non-parametric test, Friedman's test.

We can select this in jamovi by selecting ANOVA, then Repeated Measures ANOVA – Friedman, under Non-Parametric. Move the levels of the independent variable to the Measures box, select Pairwise comparisons (Durbin-Conover) if you would like *post hoc* tests, and get descriptive statistics as well.

With the Broca's aphasia example, the output will look like this:

Repeated Measures ANOVA (Non-parametric)

Friedman

χ^2	df	p
6.64	2	0.036

Pairwise Comparisons (Durbin-Conover)

			Statistic	p
Speech	-	Conceptual	1.44	0.180
Speech	-	Syntax	3.50	0.006
Conceptual	-	Syntax	2.06	0.067

[5]

Note that this is an interesting situation where our non-parametric pairwise comparisons turned out to be *more powerful* than the parametric version of the test. Under many circumstances, the parametric test will be more powerful, but sometimes the non-parametric equivalent ends up being more powerful.

When we go to write up the results, we should report the median as well as the mean, because this is the measure of central tendency used in the Friedman test. The medians are obtained by selecting Descriptives when we run the test.

> Friedman's test was performed examining how three tasks affected word recognition in patients suffering from Broca's Aphasia. Task type significantly affected word recall, $\chi^2(2) = 6.64$, $p = .036$. Pairwise comparisons using Durbin-Conover indicated that participants recognized significantly more words in the speech task ($M = 7.17$, $Mdn = 7.50$) than participants in the syntax task ($M = 4.33$, $Mdn = 6.50$; $p = .006$). There were no differences between the conceptual task ($M = 6.17$, $Mdn = 6.50$) and both the speech and syntax tasks.

Now that we have run both these tests, which should we write up? In this situation, given the small sample size, it would be wise to report the non-parametric test.

CHAPTER 8: FACTORIAL DESIGNS AND TWO-WAY ANOVA

Portions of this chapter (In practice: Two-way ANOVA) are adapted from Dana Wanzer's "Statistics with jamovi" licensed under a Creative Commons BY-SA (CC BY-SA) license version 4.0, with changes, including an expanded discussion of hypothesizing main effects and interactions and further discussion of contrasts and *post hoc* tests.

What is a Factorial Design?

Have you ever fallen asleep during an afternoon lecture, or at least found it really hard to pay attention and then later remember what you were taught? Imagine we are interested in studying the effect of time on day on memory for lecture material. We might also expect that caffeine intake influences the extent to which people can stay alert and attend during the lectures. So, we design an experiment with two independent variables, each with two levels: time of day (morning or afternoon); caffeine intake (placebo or one dose of caffeine). This would be a **factorial design**. A factorial design simply means that we have two or more independent variables (each with at least two levels) in our experiment. Factorial designs can be useful for adding an extra independent variable with relatively little effort, but also for studying **interactions**. In other words, we can see how the effect of one independent variable changes according to the levels of the other independent variable. This can often be more interesting than just studying **main effects** (i.e., the effect of a single independent variable).

Specifying a Factorial Design

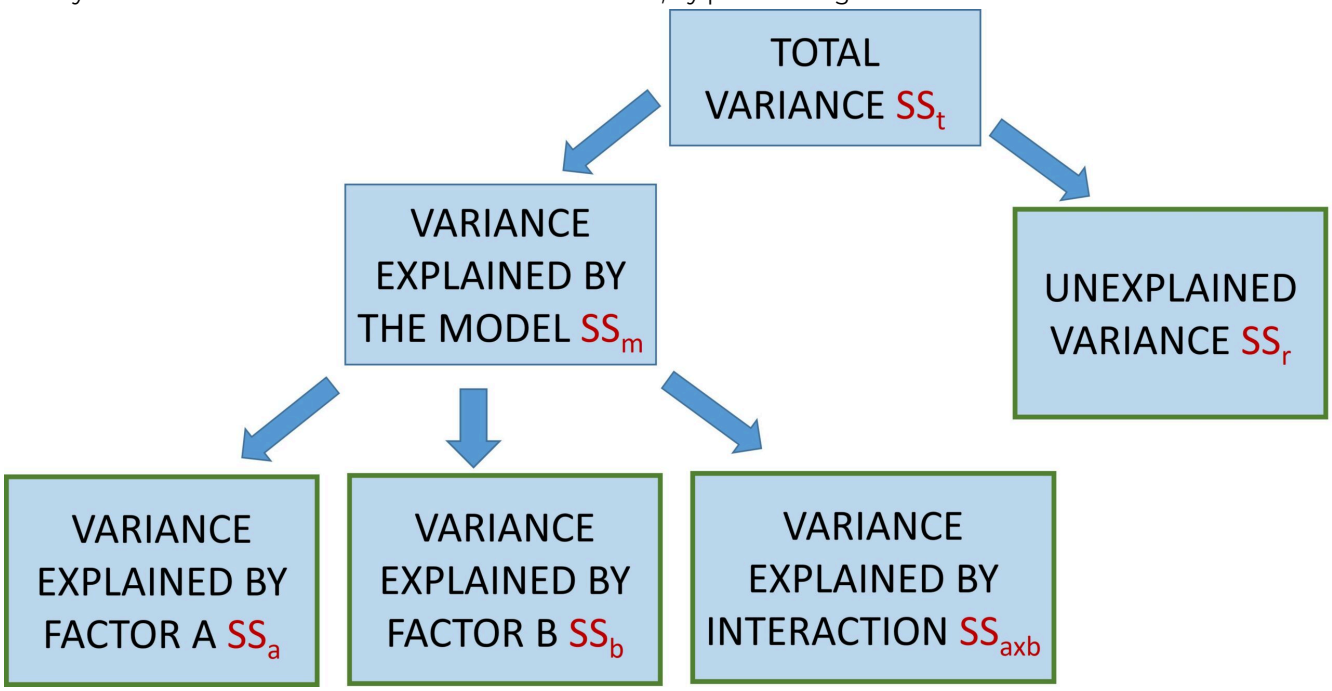
You may have seen journal articles use a particular style to describe a factorial design. For example, you might see something like:

The experiment was a 3 x 2 between-subjects design (time of day: morning, afternoon, or evening; caffeine intake: placebo or one dose of caffeine).

The "3 x 2" indicates that there were three levels of the first factor/independent variable – morning, afternoon, and evening – and two levels of the second factor/independent variable – placebo and one dose of caffeine. A 3 x 2 design like this will result in six "cells" (six means). Note that we should always specify if the design is between-subjects (each cell represents different participants), within-subjects (each cell represents the same participants), or mixed (both between- and within-subjects – more on that in a later chapter!).

Two-Way ANOVA

Two-way ANOVA examines main effects and interactions, by partitioning the variance as follows:



With a two-way ANOVA, we will therefore be computing three *F*'s, one for each main effect (main effect of factory A, main effect of factor B) and one for the interaction (A x B).

Thus, a factorial ANOVA allows us to examine both main effects and interactions.

Main Effects

Main effects compare marginal means. The **marginal means** represent the means of the levels of each variable while collapsing across the levels of the other variable.

In some designs, the marginal means will be the same as the raw score means for each cell in your design, but this is not always the case, because the marginal means are based on the statistical model, rather than just based on computing a statistical mean. That is, they are adjusted according to other factors in the model. If you have a one-way design, then the marginal means will be the same as the original means you get via computing descriptive statistics. However, if you have a two-way design or other kinds of more complex designs, the marginal means may be different from the original means. This will happen when you have an unbalanced design (e.g., different numbers of participants in each cell of a between-subjects design) or if you have a covariate (more on that when we get into analysis of covariance, ANCOVA, in a later chapter).

In a two-way design (i.e., with two factors/independent variables), there are two possible main effects – one for each factor. In a three-way design (i.e., with three factors) there are three possible main effects).

Interactions

Interactions occur when the effect of one factor depends on or changes according to the levels of the other factor. For example, perhaps the effect of time of day on students' memory for class material depends on caffeine intake. When students consume placebo, we might see a strong effect of time of day, whereby students who sit in the morning lecture later remember much more than students who sit in an afternoon lecture. On the other hand, when students consume caffeine before class, we might see a weak effect or even no effect of time of day: the caffeinated students might perform equally well in the morning and the afternoon class. Therefore, we would say that there is an interaction between time of day and caffeine intake: how time of day impacted students' performance depended on whether or not they were caffeinated!

The most helpful way to understand what a significant interaction means (or to write about a hypothesized interaction) is to graph the data (or sketch a graph of what you expect). We usually do this by putting one independent variable on the x-axis and then indicating the other independent variable by using different coloured lines or bars. If you have a significant interaction, your lines will not be parallel (note that you could have a non-significant interaction even with non-parallel lines!). In class, we shall break this down in more detail and you practice describing an interaction when you have a significant interaction in your data.

Types of Interactions

An interaction is called an **ordinal interaction** when the lines do not cross. In this case, the effect of factor A may be present for both levels of factor B, but the effect may be lessened for one of the levels. Or, the effect of factor A may be present for just one level of factor B. In the first instance, the main effect of factor A may still be meaningful. In the second case, it may not be useful to talk about a main effect of factor A, because that main effect is only present for one level of factor B. On the other hand, when we have a **disordinal interaction**, then the lines cross and it may be that the effect of factor A is opposite for the two levels of factor B. We would still report the results of the main effects, but it would only be meaningful to interpret the interaction, since this *qualifies* the main effects. We shall look at examples of these in class.

Try sketching what an ordinal interaction and a disordinal interaction might look like for our 2 x 2 time of day (morning or afternoon) and caffeine intake (placebo or one dose of caffeine) example.

Interpreting Interactions: Simple Main Effects

We can use a further statistical test, called simple main effects, to break down interactions. **Simple main effects** test the effect of one independent variable at each level of the other independent variable. For example, let's say we were looking at the effect of time of day (morning, afternoon, or evening) and caffeine intake (placebo vs. one dose of caffeine) on memory for lecture material. If we obtained a significant interaction between time of day and caffeine intake, we could run simple main effects to find out if there is a significant effect of time of day for the placebo dose, and for the one dose of caffeine, separately. Simple main effects are not available in the regular ANOVA package in jamovi, but they are available in the gamlj (linear models) add-on. These are beyond the scope of this class, but if you need them in future, you will know where to find them!

Interpreting Interactions: Planned Comparisons or *post hoc* Tests

Just as we did with our one-way ANOVA, we can also explore interactions further using planned comparisons (when we have specific *a priori* predictions about which pairs of cell means will differ from each other) or *post hoc* tests (if we had no *a priori* predictions).

It can be a little clunky to do this in jamovi. For planned comparisons, you can select to do Contrasts, but these will just break down the main effects and not a significant interaction. You can select an interaction term and run *post hoc* tests on the interaction term. This will compare all possible pairs of cell means. However, note that we should not compare cell means that are confounded (i.e., that differ on more than one factor). If we get a difference when we look at a confounded comparison, we won't know if the difference is due to factor A or factor B! Therefore, we only look at unconfounded comparisons, where the cell means differ only along one factor. As in chapter 6, we can also use the method of looking at a subset of the *post hoc* tests (and applying our own correction for multiple comparisons) as a work around to get planned comparisons (assuming we actually had *a priori* predictions!).

This starts to get a bit mind-boggling, so we shall look at some specific examples in class.

Interpreting Interactions: What If the Interaction Was Not Significant?

If you predicted an interaction and were hoping to run some planned comparisons to describe it, but the interaction is not significant when you run the ANOVA you should *not* then run the planned comparisons. We shall look at an example like this in chapter 9.

In Practice: Two-Way ANOVA

Let's see how to conduct the two-way ANOVA in jamovi. We'll start with a fully between-subjects design.

1. Look at the Data

Open data from your Data Library in "lsj-data." Select and open "rtfm." This data has two independent variables: attend (whether or not the student turned up to lectures) and reading (whether or not the student actually read the textbook). In both cases, 1 = they did and 0 = they did not. The dependent variable is their grade. There were eight participants in total. Note that the grade variable is currently set to be a nominal measure, but we should probably treat this as continuous, so go to Data, click on Setup, select the grade variable, and change the Measure type to Continuous.

Then you can obtain descriptive statistics (through Analyses, Exploration, Descriptives). When you obtain your descriptive statistics, note that you can either get them for the whole sample, or you can use Split by to get descriptives for each cell in your design. To use Split by, simply click and drag the attend variable and the reading variable to the Split by box. Note that we have a balanced design: there are 2 participants in each cell in our design.

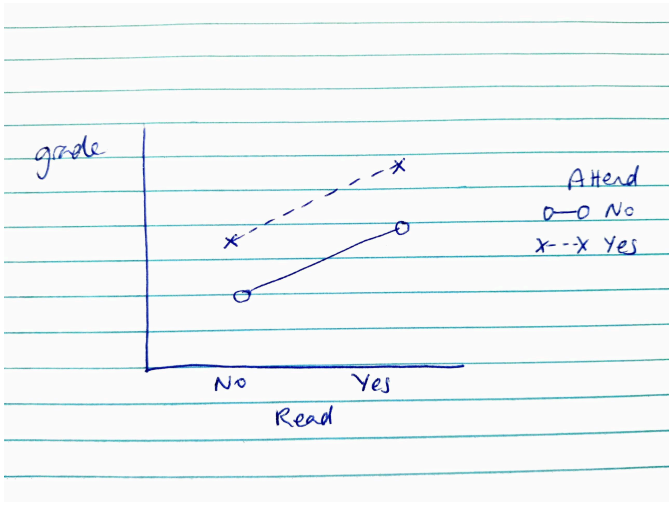
Hypothesizing Main Effects and/or Interactions

We might hypothesize that both lecture attendance and reading will have a significant effect on students' grades (i.e., two main effects), with improved grades for students who do versus do not attend/read. Would you predict an interaction? Before you read ahead, take a moment to think about whether you would predict an interaction and if you would, write it down.

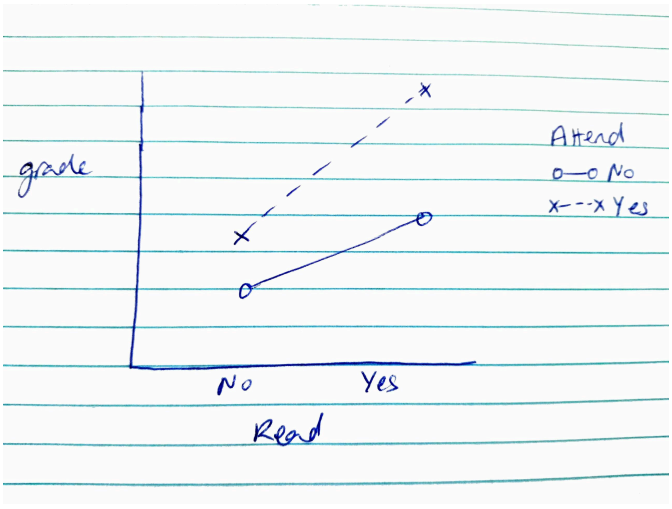
Often, when learning about interactions, students might say that they would expect an interaction and describe it as follows: "students who both read and attend lectures will have the best grades and students who neither read nor attend lectures will have the worst grades." Let's come back to this statement in a moment. First, I would like you to sketch on a piece of paper what you would expect if there were just two main effects (improved grades for students who do versus do not attend class and improved grades for students who do versus do not read) and no interaction.

Now let's look a bit more closely at this statement from above. There are three problems with this statement. The first is a general problem in that it describes how two particular cells in the design differ from the others, but it does not describe how the effect of one factor depends on or changes according to the level of the other factor. If we want to describe an interaction for this study, we would need to describe how the effect of attendance on grades is different according to whether or not students read. The second problem, which is related to the first, is that it's not clear from that description what comparisons I am going to make when I want to break down the interaction – which groups will I compare with which? Will I just compare the attend+read group with the did not attend+did not read group? That would be a confounded comparison (which we have already discussed is not appropriate!).

Now let's take a look at the third problem. Go back to the sketch of two main effects that you drew a few minutes ago. Your sketch should look something like this (note, you could have Attend on the x-axis and Read as separate lines, and that would be fine, too!). Now, let's look again at the statement from above: "students who both read and attend lectures will have the best grades and students who neither read nor attend lectures will have the worst grades." Does the graph you drew capture the information provided in the statement? If you have a graph looks anything like mine, then it does! The point here is that the statement provided there does *not* describe an interaction. Rather, it describes the two main effects and shows that the consequences of reading and attending lectures are *additive*, not that the two independent variables interact.



So, if we did want to propose an interaction for these two variables, what might it look like and how would we describe it? Perhaps you expect that the effect of attendance on grades is greater for students who read than those who do not. Again, take a moment to sketch what this would look like. You should get something like this:



In this case, we might still have two significant main effects – it is the case that grades are higher for students who attend vs. do not attend, and for students who read vs. do not read. The interaction is an ordinal interaction (the lines do not cross), whereby the effect of reading on grades is greater for students who do attend vs. those who do not attend.

The main point of our digression here is to note that we should be very careful when we want to propose an interaction. *Always* sketch what you expect to find, and then describe what you see. When describing an interaction, remember to talk about how the effect of one factor changes according to which level(s) of the other factor you are looking at.

As we get ready to proceed with our example, let's assume we are just predicting two main effects.

2. Check Assumptions

The assumptions for the two-way between-subjects ANOVA are the same as for the one-way ANOVA. Under the Analyses tab, select ANOVA, and then ANOVA. Move grade to the Dependent Variable box and attend and reading to the Fixed Factors box. You would then proceed to check for homogeneity of the variances and normality as we did for the one-way ANOVA, by selecting Homogeneity test, Normality test, and Q-Q Plot under the Assumption Checks drop-down menu. Remember the caveats about interpreting these tests: with small sample sizes they will not be significant, even if you have violated the assumption; with large samples they may be significant even with small deviations from normality/homogeneity of the variance. With this example dataset, we have a particularly small sample and so we would not usually proceed with this analyses because

we cannot have any confidence that we have met the assumptions; however, for illustrative purposes, we are going to go ahead!

Surprisingly, Levene's test for homogeneity of the variances is significant, in spite of the small sample size! So we have failed to meet that assumption. Unfortunately, we cannot perform Welch's *F*-test when we have more than one factor, so we would note the failed assumption and move on.

Shapiro-Wilk's test of normality was not significant and the Q-Q plot looks good, so we do not need to be concerned about deviations from normality (though again, remembering the caveat about small sample size in this particular instance).

Assumption Checks

Homogeneity of Variances Test (Levene's)

F	df1	df2	p
6.80e+32	3	4	< .001

[3]

3. Perform the Test

Next, we would perform the test. Make sure you select ω^2 as your measure of effect size. Under Post Hoc Tests, there is no need to select *post hoc* tests for each factor, unless you have more than two levels of a factor. If you only have two levels, the *post hoc* tests will simply repeat your tests of the main effects. However, we can select the interaction term, attend * reading, and request a *post hoc* test for that. It will give us all the possible pairwise comparisons amongst the cells in our design. In our example, we did not predict an interaction (in cases where you do, see the box below about contrasts in factorial ANOVA), but we might wish to examine differences among cell means using the *post hoc* tests, in an exploratory fashion.

Contrasts in factorial ANOVA

Note that for the contrasts, jamovi will only provide comparisons within your main effects, not the interaction. Therefore, when you only have two levels for each factor, the contrast just gives you the same information as the main effect. If we had three levels of one of one of our factors, then we could select a contrast for that independent variable if we had some planned comparisons. For example, perhaps our attendance factor has three levels: attend live lecture, watch recorded lecture, do not attend or watch recorded lecture. In this case we might expect that grades are better for those who attend the live lecture than watch the recorded lecture and better for those who watch the recorded lecture than do not attend or watch the recorded lecture. In that case, we could request a Repeated contrast (see the section on Follow-up tests in chapter 6 for other types of contrasts). If you have planned comparisons and need contrasts to break down your interaction, see the sections on Interpreting interactions, earlier in this chapter.

We can use the Holm correction and request the effect size, Cohen's *d*, for our *post hoc* tests. Under the Estimated Marginal Means menu, request plots and tables. Move attend to under Term 1, then click Add New Term and add reading under Term 2. Finally, click Add New Term again and move both attend *and* reading to under Term 3.

4. Interpret Results

Let's look at the main results next.

ANOVA

ANOVA - grade

	Sum of Squares	df	Mean Square	F	p	ω^2
attend	648.00	1	648.00	18.254	0.013	0.255
reading	1568.00	1	1568.00	44.169	0.003	0.638
attend ✻ reading	8.00	1	8.00	0.225	0.660	−0.011
Residuals	142.00	4	35.50			

We got three lines of results in addition to the typical residuals (error). The first two rows are our main effects of attend and reading on grades. The *p*-values for both are statistically significant indicating attend affects grades and reading affects grades. However, it also added an interaction term, attend * reading, which is not statistically significant. This means we do not have an interaction between attend and reading on grades.

The results of the *post hoc* tests are as follows:

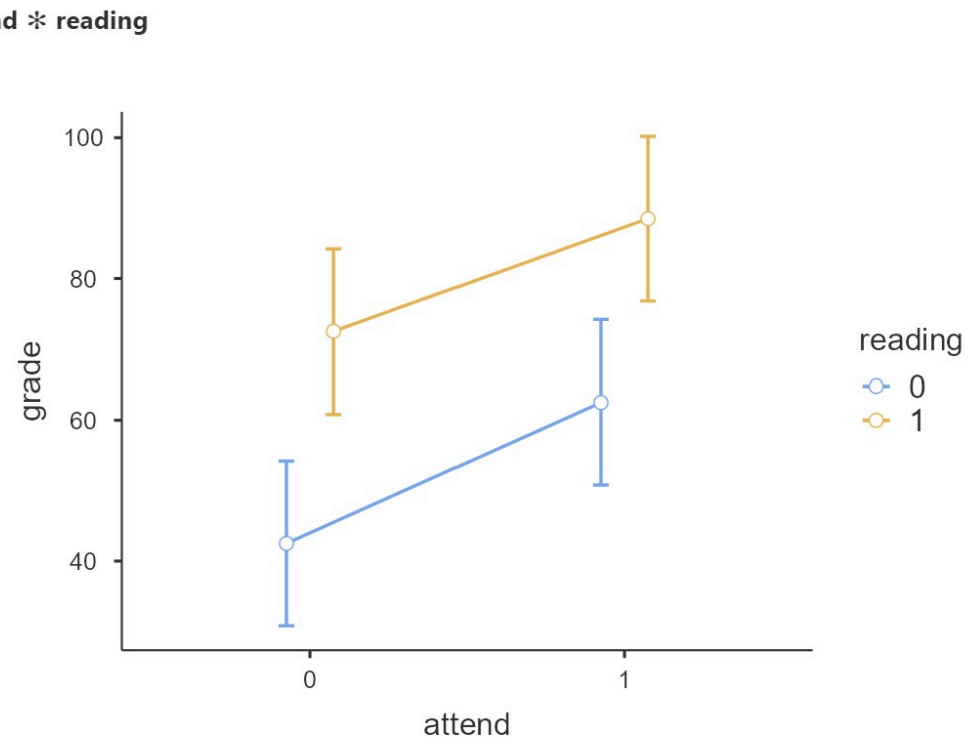
Post Hoc Tests

Post Hoc Comparisons - attend ✻ reading

Comparison										
attend	reading		attend	reading	Mean Difference	SE	df	t	p_{holm}	Cohen's d
0	0	-	0	1	−30.00	5.96	4.00	−5.04	0.037	−5.04
		-	1	0	−20.00	5.96	4.00	−3.36	0.085	−3.36
		-	1	1	−46.00	5.96	4.00	−7.72	0.009	−7.72
	1	-	1	0	10.00	5.96	4.00	1.68	0.169	−1.68
		-	1	1	−16.00	5.96	4.00	−2.69	0.110	−2.69
1	0	-	1	1	−26.00	5.96	4.00	−4.36	0.048	−4.36

Note. Comparisons are based on estimated marginal means

Because we do not have a significant interaction, the *post hoc* tests do not tell us anything more than the significant main effects. We already know from the ANOVA table that there is a main effect of both reading and attending, and so we would not look further at the *post hoc* tests. On the other hand, if the interaction had been significant, we could look at the *post hoc* tests to see where the differences in cell means lie. It can be pretty overwhelming to look at this table of *post hoc* results, so I recommend looking first at the plot of the marginal means.



In class, we shall spend some time looking at one of these plots in an example where there *is* an interaction, so that we can explore how best to describe what is going on, using the results from the *post hoc* tests.

Write Up the Results in APA Style

We can then write up our results section. Note that we should describe the direction of any significant effects by stating which group scored higher than which, and reporting the relevant means from the Estimated Marginal Means tables.

> I conducted a two-way between-subjects ANOVA to examine the effect of attendance and reading on student grades. Both attendance, $F(1, 4) = 18.25$, $p = .013$, $\omega^2 = .26$, and reading, $F(1, 4) = 44.17$, $p = .003$, $\omega^2 = .64$, affected student grades; there was no significant interaction between attendance and reading, $F(1, 4) = 8.00$, $p = .660$. Students who attended lectures ($M = 75.50$, SE = 2.98) had higher grades than students who did not ($M = 57.50$, $SE = 2.98$). Students who read ($M = 80.50$, $SE = 2.98$) had higher grades than students who did not ($M = 52.50$, $SE = 2.98$).

In this case, we did not have a significant interaction. But what if the interaction term was significant? Then, we could report the *post hoc* tests that best help us support the description of the interaction we obtained when following the steps to describing an interaction (to be outlined and practised in class). Of course, if we had planned comparisons, we would report only the comparisons that we had planned *a priori*. Therefore, when planning an experiment where you will use factorial ANOVA, it is a good idea to think very carefully about what kind of interaction you expect to see – sketch a graph of the pattern of results you expect to see in order to decide what planned comparisons you wish to make.

CHAPTER 9: OTHER ANOVAS - MIXED, 3-WAY, ANCOVA

Portions of this chapter (Analysis of covariance) are adapted from Dana Wanzer's "Statistics with jamovi" licensed under a Creative Commons BY-SA (CC BY-SA) license version 4.0, with some changes.

Mixed ANOVAs

What is Mixed ANOVA?

When we have both between-subjects and within-subjects independent variables (at least one of each) we have what is called a **mixed design**. Imagine, for example, that students in Dr. Chico's class (whose data we used in chapter 5 on *t*-tests), were given two tests (test 1 and test 2). Between test 1 and test 2, half the students received extra tutoring, and half did not. We are interested to learn if the grades of those students who received extra tutoring increased more than for those students who did not receive extra tutoring.

An Example

The dataset used in this example is adapted from the chico dataset in the lsj-data Dat Library. If you want to adapt it yourself, you can simply create a fourth variable in the dataset, called Tutoring. Students 1 through 10 should be given a score of 0 (no tutoring) and Students 11 through 20 should be given a score of 1 (tutoring). The first few rows of your datafile will then look like this:

	id	grade_test1	grade_test2	Tutoring
1	student1	42.9	44.6	0
2	student2	51.8	54.0	0
3	student3	71.7	72.3	0
4	student4	51.6	53.4	0
5	student5	63.5	63.8	0
6	student6	58.0	59.3	0
7	student7	59.8	60.8	0
8	student8	50.8	51.6	0
9	student9	62.5	64.3	0
10	student10	61.9	63.2	0
11	student11	50.4	51.8	1
12	student12	52.6	52.2	1
13	student13	63.0	63.0	1
14	student14	58.3	60.5	1

In this dataset, we know have one within-subjects factor (testing occasion: test one or test two) and one between-subjects factor (tutoring: no tutor or tutor). Therefore, it is a mixed design. Let's suppose that we hypothesize a main effect of testing occasion (scores should increase from test one to test two) and an interaction (at time 1, students would show no difference in test scores, but at time 2, students who had a tutor would show higher test scores than students who did not have a tutor).

To run a mixed ANOVA in jamovi, after selecting ANOVA, select Repeated Measures ANOVA – use Repeated Measures ANOVA for any situation where you have at least one within-subjects factor. Enter your levels of the within-subjects variable as you did for the repeated measures ANOVA in chapter 7. Tutoring will then go in

the Between Subjects Factors box. Choose Generalised η^2 as the measure of effect size and type Grade in the Dependent Variable Label box. Select all the assumption checks (you need to check assumptions for both the between- and the within-subjects factor!).

Why did I get "NaN" for Mauchly's test of sphericity? Sometimes you will get a value of "NaN" for the *p*-value for the test of sphericity. This will occur when we only have two levels of our within-subjects factor (as in the example above: test 1 and test 2). The test of sphericity works by looking at the differences among the differences in the variances, but when we only have two levels, there is only one difference to be examined, so we cannot get a value for Mauchly's. In fact, in those situations, there is no point running this test, but I wanted you to do it here so that you could see what happens.

Select Post Hoc Tests for the interaction term, applying No correction. We had predicted an interaction, with specific differences between pairs of means (at time 1, students would show no difference in test scores, but at time 2, students who had a tutor would show higher test scores than students who did not have a tutor) and so we will treat these *post hoc* tests as planned comparisons and apply the Bonferroni correction ourselves, assuming the interaction term in the ANOVA is significant. You can also obtain the estimated marginal means for both main effects and the interaction term.

Let's look at the results. Because we have both a within- and a between-subjects factor, we need to look for both of those results in the output. This is a two-way ANOVA, so we should be looking for two main effects and one interaction. Even though we did not hypothesize a main effect of tutoring, we are still going to report both main effects and the interaction when we write up our results (whether significant or not).

Repeated Measures ANOVA

Within Subjects Effects

	Sphericity Correction	Sum of Squares	df	Mean Square	F	p	η^2_G
Test	None	19.740	1	19.740	40.431	< .001	0.012
	Greenhouse-Geisser	19.740	1.00	19.740	40.431	< .001	0.012
Test ✻ Tutoring	None	0.156	1	0.156	0.320	0.579	0.000
	Greenhouse-Geisser	0.156	1.00	0.156	0.320	0.579	0.000
Residual	None	8.789	18	0.488			
	Greenhouse-Geisser	8.789	18.00	0.488			

Note. Type 3 Sums of Squares

[3]

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G
Tutoring	6.64	1	6.64	0.0749	0.787	0.004
Residual	1595.71	18	88.65			

Note. Type 3 Sums of Squares

You'll see that the Within Subjects Effects table shows you the statistics for any test that included at least one within-subjects factor (i.e., the main effect of test, and the interaction between test and tutoring). The Between Subjects Effects table shows you the statistics for any test that includes *only* between-subjects factor(s) (i.e., the main effect of tutoring). Therefore, if we had another between-subjects factor (e.g., student's major: Psychology or History), the main effect of major would also appear in the Between Subjects Effects table, as would the interaction between major and tutoring. On the other hand the interaction between major and test, and the three-way interaction between tutoring, major and test, would appear in the Within-Subjects Effects box (because that contains any effects that include at least one within-subjects factor!).

We can see that there was a significant main effect of test, but no significant effect of tutoring and no significant interaction. Because the interaction was not significant, we do *not* follow up with our planned comparisons. Our results section would look something like this (remember to report all main effects and interactions, whether significant or not).

> I used mixed ANOVA to examine the effects of test and tutoring on students' grades. Students grades improved significantly from test 1 (*M* = 57.0, *SE* = 1.52, 95% CI [53.8, 60.2]) to test 2 (*M* = 58.4, *SE* = 1.47, 95%

CI [55.3, 61.5]), $F(1, 18) = 40.43$, $p < .001$, $\eta^2_G = .012$. However, there was no main effect of tutoring (no tutor: $M = 58.1$, $SE = 2.11$, 95% CI [53.7, 62.5]; tutor: $M = 57.3$, $SE = 2.11$, 95% CI [52.9, 61.7]), $F(1, 18) = 0.07$, $p = .79$, and no significant interaction between tutoring and testing occasion, $F(1, 18) = 0.32$, $p = .58$.

Note

When using repeated measures ANOVA, there is no option to get Cohen's *d* for our *post hoc* tests, should we need them for breaking down significant main effects or an interaction. There are several different ways to get Cohen's *d* via jamovi, or, you can use the following approach. For between-subjects comparisons, use the *t* and *n* for each group as follows:

$$d_s = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

For within-subjects comparisons, use the mean difference and the standard deviation for each level of interest (obtained by running descriptive statistics through Exploration in jamovi), as follows:

$$\text{Cohen's } d_{av} = \frac{M_{\text{diff}}}{\frac{SD_1 + SD_2}{2}}$$

(see Lakens, 2013, for more discussion of appropriate calculations for Cohen's *d*)

Three-Way ANOVA

A three-way ANOVA has three factors or independent variables. These can be all between-subjects, all within-subjects, or a combination of between- and within-subjects factors.

When we have a three-way ANOVA, there are three possible main effects (one for each factor, A, B, and C), three possible two-way interactions (A x B, B x C, and A x C), and one possible three-way interaction (A x B x C). In class, we shall go through an example of how to interpret a three-way interaction!

Analysis of Covariance

When Do We Use ANCOVA?

We use analysis of covariance (ANCOVA) when we want to test for differences among group means, but we suspect that a measured extraneous variable also affects the dependent variable, in an unsystematic way. ANCOVA allows us to test the effect of the independent variable on the dependent variable, while *controlling for* the relationship between the nuisance variable and the dependent variable. The benefit of using this test is that it controls for unsystematic variation and so reduces error variance.

Let's look at an example from the lsj-data Data Library. Open the dataset called ancova. This data is fictional data from a health psychologist who was interested in the effect of transportation used to commute (1 = driving, 2 = cycling) and stress (1 = high, 2 = low) on happiness levels, with age as a covariate. (This is actually a 2 x 2 independent factorial design with a covariate!) For the purposes of introducing ANCOVA, let's keep it simple and just focus on the commute independent variable, with age as the covariate.

1. Look at the Data

I'll assume by now that you know how to look at your descriptive statistics! I recommend first changing the age variable to continuous (it is not nominal!). You should also obtain the descriptive statistics for the covariate. I have copied the descriptive statistics below, split by commute – it will be interesting to compare these to the marginal means, later!

Descriptives

	commute	happiness	age
N	1	10	10
	2	10	10
Missing	1	0	0
	2	0	0
Mean	1	43.3	37.5
	2	65.0	35.9
Median	1	42.5	39.0
	2	65.5	28.5
Standard deviation	1	11.9	13.7
	2	18.0	16.5
Minimum	1	28.0	22
	2	42.0	21
Maximum	1	60.0	65
	2	90.0	64

Our hypothesis is that there is a significant effect of transportation used to commute on happiness, *when controlling for* age. Specifically, we might predict that people who ride are happier than people who drive.

2. Check Assumptions

ANCOVA has the same assumptions as the one-way ANOVA (normal distribution, homogeneity of the variances, interval or ratio data, scores are independent between groups) *and* there are two additional assumptions we need to check. Let's focus on those here.

1. No Significant Effect of IV on Covariate

Check to see if there is a significant effect of the independent variable (commute) on the covariate (age). Run an ANOVA with commute as the Fixed Factor and age as the Dependent Variable in jamovi. There should be no significant effect! We can see that there is no significant effect of commute on age, so we are safe to proceed.

ANOVA

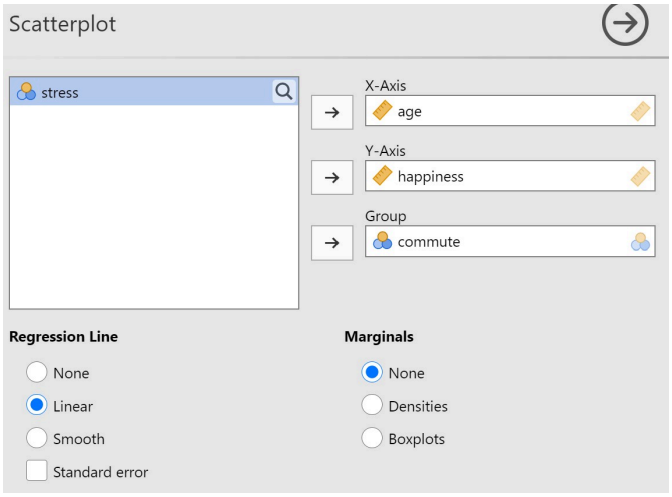
ANOVA - age

	Sum of Squares	df	Mean Square	F	p
commute	12.8	1	12.8	0.0558	0.816
Residuals	4131.4	18	229.5		

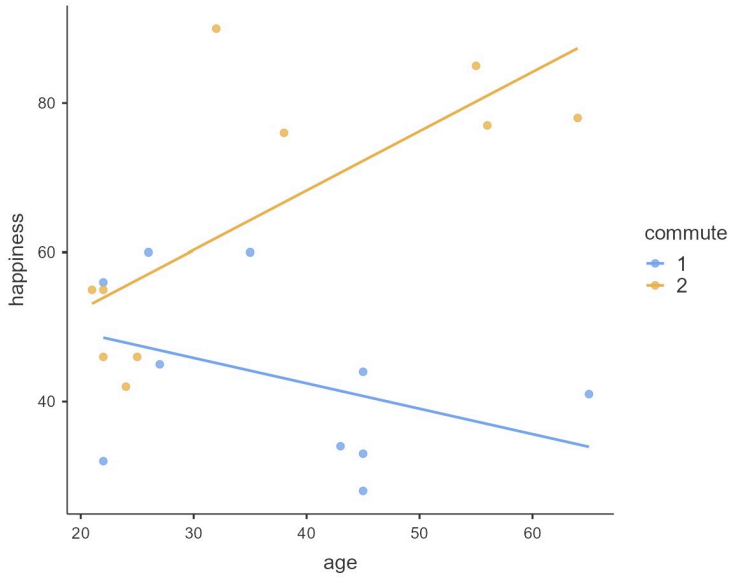
2. Homogeneity of Regression Slopes

The second additional assumption is that the relationship between the covariate and the dependent variable is similar for all levels of the independent variable (homogeneity of regression slopes). In other words, the relationship between age and happiness levels should be similar for the drivers and the cyclists. There are two ways to investigate this question. First, we can produce a scatterplot showing the relationship between age and happiness for drivers and for cyclists – i.e., let's plot the regression slopes to see if they look the same (homogeneous) or not.

In jamovi, select Exploration and Scatterplot (from the scatr module – if you did not already install this, you will need to do so). Move age to the X-Axis, happiness to the Y-Axis, and commute to the Group. Select Linear for the regression line. The regression line will show the best fit line through the points on the scatterplot (more on regression lines in a later chapter).

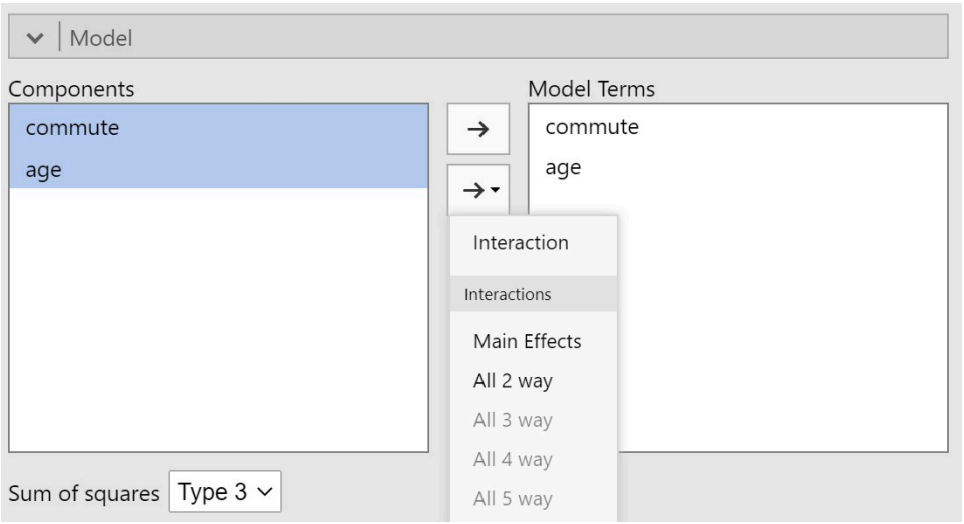


Scatterplot



In the scatterplot above, the blue dots and line represent the drivers and the orange dots and line represent the cyclists. It looks like for the drivers, happiness decreases with age, and for the cyclists, happiness increases with age. We may have violated the assumption, but to check whether this difference in the regression slopes is significant, we need to go further.

In jamovi, select the ANOVA button under Analyses, and then select ANCOVA. Move happiness to the Dependent Variable box, commute to the Fixed Factors box, and age to the Covariates box. This is how we would normally set up our ANCOVA. To check homogeneity of the regression slopes, now go into the Model drop-down menu to add an interaction term between the covariate and the independent variable. To do this, hold down the Shift key on your keyboard and click on both commute and age. Then click on the little arrow with the drop-down menu and select Interaction. Commute * age will then show in the box under Model Terms.



We can now examine the ANCOVA table to see if there was a significant interaction between commute and age in the effects on happiness.

ANCOVA

ANCOVA - happiness

	Sum of Squares	df	Mean Square	F	p
commute	247	1	247	1.62	0.222
age	206	1	206	1.35	0.263
commute ✻ age	1289	1	1289	8.43	0.010
Residuals	2447	16	153		

When we look at third row of the table above (commute * age), p = .010, which means there is a significant interaction between commute and age and the assumption of homoegeneity of the regression slopes has been violated (i.e., the regression slopes are *not* homogenous, they are heterogenous).

Note that sometimes heterogeneity of the regression slopes may be interesting or expected – perhaps we even expect that cyclists, but not drivers, should get happier with age! If so, then we need to go to another kind of statistical model, called the multilevel model, but that is beyond the scope of this book. Having violated the assumption of homogeneity of the regression slopes, we should also go to a multilevel model, but for illustrative purposes, let's proceed with our ANCOVA.

3. Perform the Test

To perform the ANCOVA, we can start a fresh analysis in jamovi. Select ANCOVA under the ANOVA analysis menu and again, Move happiness to the Dependent Variable box, commute to the Fixed Factors box, and age to the Covariates box. Select ω^2 for effect size and this time do not make any edits to the model.

There is no need to run contrasts or *post hoc* tests because we only have two levels of our factor (but if we had three or more levels, we would likely choose to run one of them according to whether or not we had specific predictions about differences among levels of the independent variable.

Under Estimated Marginal Means, move the commute variable to the Term 1 box, select both plots and tables and unselect Equal cell weights.

4. Interpret Results

The results are shown below.

ANCOVA

ANCOVA - happiness

	Sum of Squares	df	Mean Square	F	p	ω^2
commute	2464	1	2464	11.21	0.004	0.326
age	456	1	456	2.07	0.168	0.034
Residuals	3736	17	220			

The ANCOVA table shows us that the effect of commute on happiness is statistically significant. The relationship between age and happiness is not significant overall, but this does not matter – what we are interested in is whether commute affects happiness *while controlling for the relationship between age and happiness*. Even if age and happiness are not significantly associated, controlling for age will still help to reduce error variance with the groups in our experiment and increase power in our design.

In some situations, the effect of the independent variable on the dependent variable is no longer significant when the covariate is removed from the analysis. Let's check if that is the case. We can re-run the analysis but just remove the covariate. When we do that, we get the following result:

ANCOVA - happiness

	Sum of Squares	df	Mean Square	F	p	ω^2
commute	2354	1	2354	10.1	0.005	0.313
Residuals	4192	18	233			

Comparing this with the table above we can see that the *F*-value is slightly smaller and the *p*-value therefore slightly larger. Thus, including the covariate in the analysis did reduce error variance!

Let's look at the table of marginal means:

Estimated Marginal Means - commute

			95% Confidence Interval	
commute	Mean	SE	Lower	Upper
1	43.0	4.69	33.1	52.9
2	65.3	4.69	55.4	75.2

We can compare these values to the descriptive statistics we obtained earlier. For the descriptive statistics the means for driving and cycling were 43.3 and 65.0, respectively. The marginal means are the values computed by the model – the expected values for happiness *when controlling for age*. That is why the values in the marginal means table are slightly different from the values for the descriptive statistics computed on the raw data.

Finally, let's write up our results in APA format.

> We conducted a study examining how commute affects happiness levels. Furthermore, we collected data on age as a covariate of our study. We satisfied all assumptions of the ANCOVA except that we violated the assumption of homogeneity of regression slopes. Despite failing to meet this assumption, we proceeded with the ANCOVA analysis.
>
> There was a significant effect of commute on happiness, such that people who commuted via cycling

(*M* = 65.3, *SE* = 4.69, 95% CI [33.1, 52.9]) reported higher happiness than people who commuted via driving (*M* = 43.0, *SE* = 4.69, 95% CI [33.1, 52.9]), while controlling for age, $F(1, 17) = 11.21$, $p = .004$, $\omega^2 = .33$. The relationship between age and happiness was not significant, $F(1, 17) = 2.07$, $p = .17$.

CHAPTER 10: CORRELATION - ASSOCIATIONS BETWEEN PAIRS OF VARIABLES

Portions of this chapter (In practice) are adapted from “Statistics with jamovi” by Dana Wanzer, with some minor changes.

Correlational Design

Correlational Design vs. Correlational Analysis

It is commonly assumed that if you use a non-experimental (i.e., correlational design, where no independent variable is manipulated) then you will use a correlational analysis. However this is not necessarily the case: correlational design and analysis are not the same thing.

Correlational designs involve simply measuring, but not manipulating, the variables. In some situations they are more practical or ethical, or they may allow us to explore new relationships. When we have a correlational design, we may use a correlational analysis (for example, if both variables are continuous) or we may end up using something like a *t*-test (if one variable is nominal, with two categories, and the other variable is continuous).

Power in Correlational Designs

Just as we should consider power in an experimental design (how to maximize power and minimize the effect of nuisance variables), we should do the same in a correlational design. Therefore, we should try to minimize nuisance variables as much as possible (remember that nuisance variables are any variables that influence the dependent variable in a random, unsystematic way, but that are not measured or manipulated in our study). Let's say we are interested in the connection between hours of sleep (sleep quantity) and grumpiness upon waking. Other variables that could influence grumpiness upon waking up in the morning might be things like age, hours worked the day before, employment status, and so on. Therefore, we might aim to keep our sample as homogeneous as possible, by only selecting people of a certain age range and with similar employment to minimize variability in happiness that is due to those variables versus sleep quantity.

Another way to maximize power in a correlational design is to use reliable and valid measures. For example, if we measure sleep quantity by asking people at the end of the day how many minutes of sleep they had the night before, they might be fairly inaccurate in their response, resulting in an unreliable measure (compared to if we use EEG to measure time spent sleeping based on brain activity). This causes more error variance in our design and, as a result, we shall have a less consistent relationship between the two variables of interest.

Finally, we should ensure to avoid **restriction of the range**. Restriction of the range occurs when the difference between the lowest and the highest scores on one or both of our variables of interest is small. For example, in our study of sleep quantity and grumpiness, if we select people for our study who are almost always short on sleep (e.g., junior doctors), we will have restricted the range of sleep quantity scores and will likely also have little variability in morning grumpiness, making it difficult to detect if there is a relationship between sleep quantity and grumpiness.

What is r?

When we compute a Pearson correlation coefficient (the statistical test we shall learn more about shortly), we compute *r*. The value for *r* can range from -1 (perfect negative correlation) through 0 (no correlation) to +1 (perfect positive correlation). The larger the absolute value of the correlation coefficient, the stronger the relationship. When *r* is negative, as x (one variable) increases, y (the other variable) decreases. When *r* is positive, as x increases, y also increases. One thing to watch out for is that Pearson's *r* will only tell us about the strength of a *linear* relationship. That is, if we have a curvilinear relationship (e.g., U-shaped, inverted-U, or S-shaped), we might have a small value for *r* even though the relationship between x and y is actually strong! We shall have a look at some examples of this in class.

Let's take a brief look at how to compute Pearson's *r* and how to interpret it.

Computing *r*

Covariance

Correlation is computed by first calculating something called **covariance**, which indicates how much two variables vary together (i.e., covary). You may recall that variance tells us how much scores deviate from the mean for a single variable. Covariance is similar: it tells us by how much pairs of scores on two variables differ from their respective means. Covariance is calculated by the following formula:

$$Cov(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

What this means is that to calculate the covariance of x and y, we compute the product of the difference between each x score (x_i) and the mean of x and each y score (y_i) and the mean of y, for all the pairs of x scores and y scores in our sample. We sum those products and divide by the degrees of freedom (*N*-1).

Therefore, if high scores on x are consistently paired with high scores on y, and low scores on x are consistently paired with low scores on y, the value for covariance will be large and positive. If high scores on x are consistently paired with low scores on y, and low scores on x are consistently paired with high scores on y, then the value for covariance will be large and negative. In contrast, if there is no relationship between x and y – sometimes high scores on x are paired with low scores on y and sometimes with high scores, and so on, then the products of the deviations will sometimes be positive, sometimes negative, and sometimes large, and sometimes small, and so when they are summed, the result will be close to zero, i.e., small covariance.

Pearson's *r*

However, we cannot use covariance as a measure of the strength relationship between x and y, because the size of the standard deviations of x and y affect the value of covariance. Therefore, we scale covariance by the size of the standard deviation. Pearson's correlation coefficient is computed as follows:

$$r = \frac{cov_{xy}}{s_x s_y}$$

In the next chapter, we shall look at how to obtain the Pearson's correlation coefficient in jamovi.

Interpretation of *r*

A rough guide to the interpretation of the value of *r*, when used to describe the strength of the association between two variables, is as follows:

- +/- .1 = weak relationship
- +/- .3 = moderate relationship
- +/- .5 = strong relationship

Note that these are just general guidelines and that we should also think about the strength of the relationship in practical terms and in relation to our area of study.

We can also compute r^2, the **coefficient of determination**, which indicates the proportion of the variance in one variable shared by the other variable. We shall learn more about this in the chapter on regression.

Using *r*

Pearson's *r* can be used for a number of different purposes. Earlier in this chapter we have used the example of testing whether there is a relation between sleep quantity and morning grumpiness. In this case, we would use *r* to describe the strength and direction of the relation between our two variables – just to describe the results of the study.

In other situations, we may use *r* to look at a measure we have developed. In that case, we can use *r* to assess the reliability of the measure, including inter-rater reliability (the extent to which two different raters assign the consistent scores on the measure), test-retest reliability (the extent to which participants obtain the same score on a measure when re-tested), and split-half reliability (the extent to which participants' scores on half of the items on the measure correlate with the scores on the other half of the items). When using *r* to assess the reliability of a measure, then researchers are typically looking for higher values than when using *r* to describe the results of the study. Reliability of *r* = 1.0 would be rarely (never) obtained, but coefficient of *r* = .80 or above are often considered to be very good. Further discussion of this issue is beyond the scope of this class.

We can also use *r* to assess the validity of a measure, including convergent validity (the extent to which scores obtained on one measure are correlated with scores on another similar, previously-validated, measure); discriminant validity (the extent to which scores obtained on one measure are *not* correlated with scores from a measure of a different construct); and criterion validity (the extent to which a measure correlates with either a current behaviour – concurrent validity, or a future behaviour – predictive validity).

Playtime: Getting a Feel For *r*

Want to learn a bit more about what different values of *r* look like in a dataset? Try going to

www.guessthecorrelation.com to see how well you do in guessing the value of *r* for different datasets! Another great website to play around with data to see how it affects the value of *r* is here: https://rpsychologist.com/correlation/

In Practice: Pearson's Correlation Coefficient

Let's work through an example in jamovi. Open the "parenthood" datafile in the lsj-data Data Library.

1. Look at the Data

This dataset measures a new mother's daily grumpiness very precisely, on a scale from 0 (not at all grumpy) to 100 (extremely grumpy). In addition, the dataset tracks her sleeping patterns (hours of sleep) and her son's sleeping patterns across 100 days. Let's suppose we are interested in the association between the mother's sleep (dan.sleep) and her grumpiness (dan.grump) across the 100 days.

Data Set-Up

To conduct the correlation we first need to ensure our data is set-up properly in our dataset. This requires having two columns, one for each of our continuous variables. Each row is a unique participant or unit of analysis. Note that jamovi might have incorrectly imported `dan.grump` as a nominal variable but that is incorrect! This shows the importance of looking at your data and checking your measure types. Change dan.grump to a continuous variable.

Describe the Data

Once we confirm our data is setup correctly in jamovi, we should look at our data using descriptive statistics and graphs. The descriptive statistics will show you that we have 100 cases and no missing data.

2. Check Assumptions

The Pearson correlation has the three following assumptions:

1. Both variables are **normally distributed**;
2. Both variables are measured at the **interval or ratio** (i.e., continuous) level (however, we will see what we can do if we violate this); and
3. The relationship between the two variables is **linear**.

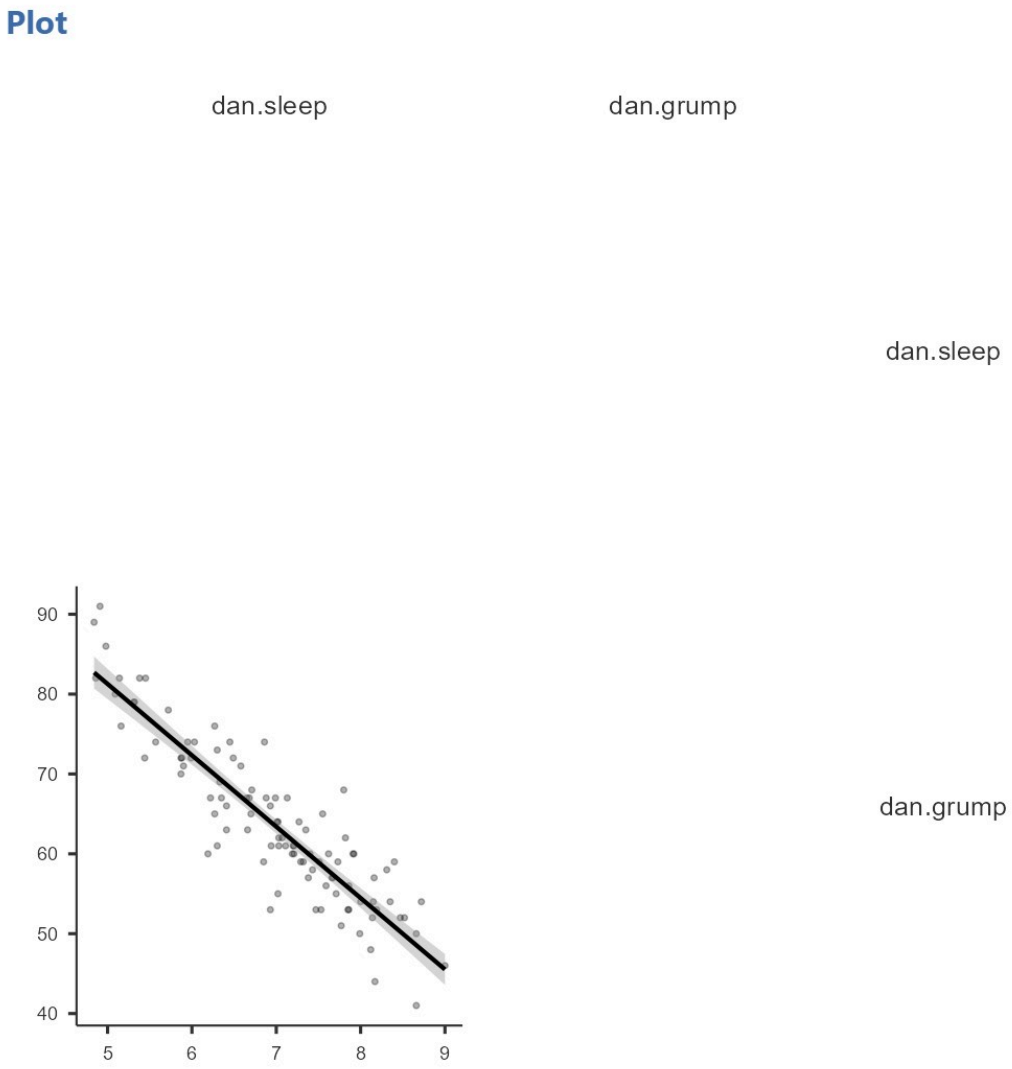
The third assumption requires looking at a scatterplot of one variable on the x-axis and the other variable on the y-axis.

To test normality, within the descriptive options in jamovi, select the Q-Q Plots. For both dan.sleep and dan.grump, the dots fall pretty close to the straight line. We also have a fairly large sample (100 datapoints) so

we are well above the 30 required for central limit theorum to kick in. Therefore, let's assume we have met the assumption of normality.

Both variables are interval or ratio data. To check the linearity of the relationship, we should go to the correlation analysis. Under the Analyses button in jamovi, select Regression and Correlation matrix. Move dan.sleep and dan.grump to the box on the right. Under Plot, check the box for Correlation matrix.

The plot will look like this:



We can see quite clearly from looking at the dots on the scatterplot that this is a linear relationship (and not a curvilinear relationship) and so it is appropriate to use Pearson's *r*. Note that we need to look at the data points themselves; the correlation matrix will always produce lines even if the underlying data looks curvilinear.

Perform the Test

In addition to the options already selected in jamovi, make sure that the Report significance box is checked, and, if you have missing data, check the box for N. We can also obtain 95% confidence intervals – this will give us a confidence interval around *r* itself (not around the means).

We also need to decide if we will use a two-tailed test (select Correlated under Hypothesis) or one-tailed test

(select Correlated positively or Correlated negatively, according to the expected direction of the relationship). Let's imagine we had no expectation about the direction of the relationship between sleep and grumpiness.

Our results will look like this:

Correlation Matrix

Correlation Matrix

		dan.sleep	dan.grump
dan.sleep	Pearson's r	—	
	p-value	—	
	95% CI Upper	—	
	95% CI Lower	—	
dan.grump	Pearson's r	−0.903	—
	p-value	< .001	—
	95% CI Upper	−0.859	—
	95% CI Lower	−0.934	—

It looks like there is a strong, negative correlation between sleep and grumpiness. We can write this up in APA format as follows:

Pearson's correlation indicated that the mother's sleep duration was significantly associated with her grumpiness, $r(98) = -.90$, 95% CI [-.93, -.86], $p < .001$. As the mother got more sleep, her grumpiness decreased.

If the result was not significant, we would report all the statistics, but we would not describe the direction of the (non-significant) association. E.g.: Pearson's correlation indicated no association between a mother's sleep duration and her grumpiness, $r(98) = .12$, 95% CI [-.08, .31], $p = .240$.

Note that if we wanted to simultaneously assess the associations among other variables in the dataset, we could enter other variables in jamovi as well, and we would get a correlation matrix showing the correlation between each pair of variables.

Alternatives to Pearson's *r*

If we have ordinal data for one or both of our variables (instead of interval or ratio data), we can, instead use Spearman's rho. This is a non-parametric statistic based on rank order. To perform Spearman's correlation, both variables need to be in rank order. In jamovi, change the check mark in jamovi from Pearson to Spearman. The interpretation is the same, but you should use r_s when reporting your result, instead of *r*.

CHAPTER 11: REGRESSION

Portions of this chapter (portions of Introduction to regression, In practice) are adapted from “Statistics with jamovi” by Dana Wanzer, with some minor changes.

Introduction to Regression

What is Regression?

Think back to the dataset we worked with in the previous chapter on correlation, where we looked at the association between a mum's hours of sleep per night and her grumpiness, across 100 days. Imagine we would like to *predict* her grumpiness on any given day based on how many hours sleep she had. We can do this using regression. Regression goes beyond correlation in that it tells us not just the strength of the association between variables but also allows us to predict the value of one variable, **the outcome, Y**, from the value of another variable, **the predictor, X**. We do this by using the equation of a straight line (flash back to high school math, anyone?!). When we conduct regression, we describe the best fitting line that summarizes a linear relationship between our variables, and then we can use the equation for this line to predict scores.

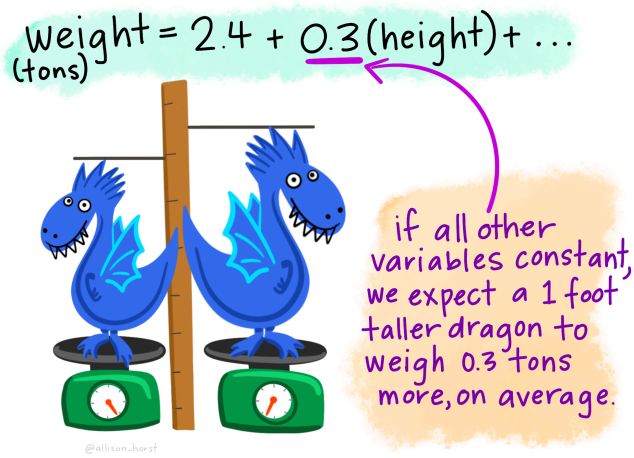
Regression is a very useful tool, which can be extended into different contexts, such as when we have more than one predictor (multiple regression). Believe it or not ANOVA is actually the same thing as regression! You'll see that when we compute regression, we can also compute *F*, just as we do in ANOVA! (But in regression we compute some other useful values about the relationship between our variables.) The actual connection is real – ANOVA and regression are based on the same fundamental statistics, but the details of that story are for another class.

The Straight Line

A linear regression model is basically a linear line, which many of us learned as y = mx + b, where y is our predicted outcome score, x is the IV, b is the intercept (the score in y when x = 0), and m is the slope (when you increase x-value by 1 unit, the y-value goes up by m units).

Let's imagine we have a dataset of dragons with a single continuous predictor (height) and a continuous dependent variable (weight). (Note, we can also have categorial predictors in regression, but we shall touch on those later.) We want to use this dataset to be able to predict the weight of future dragons. First, let's learn how to interpret the coefficients for our predictor variable.

"Artwork by @allison_horst" (CC BY 4.0)



This figure shows us that the weight of a dragon can be predicted by summing 2.4 plus 0.3 multiplied by the height of the dragon. This is an example of a regression equation, which describes the straight line that best fits the data when we plot X against Y.

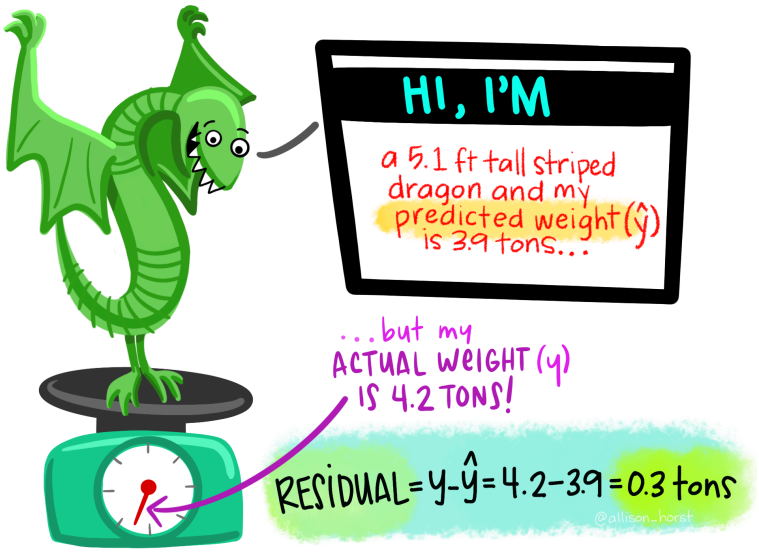
In general terms, the regression line can be described as follows:

$$Y_i = (b_0 + b_1X_i) + error_i$$

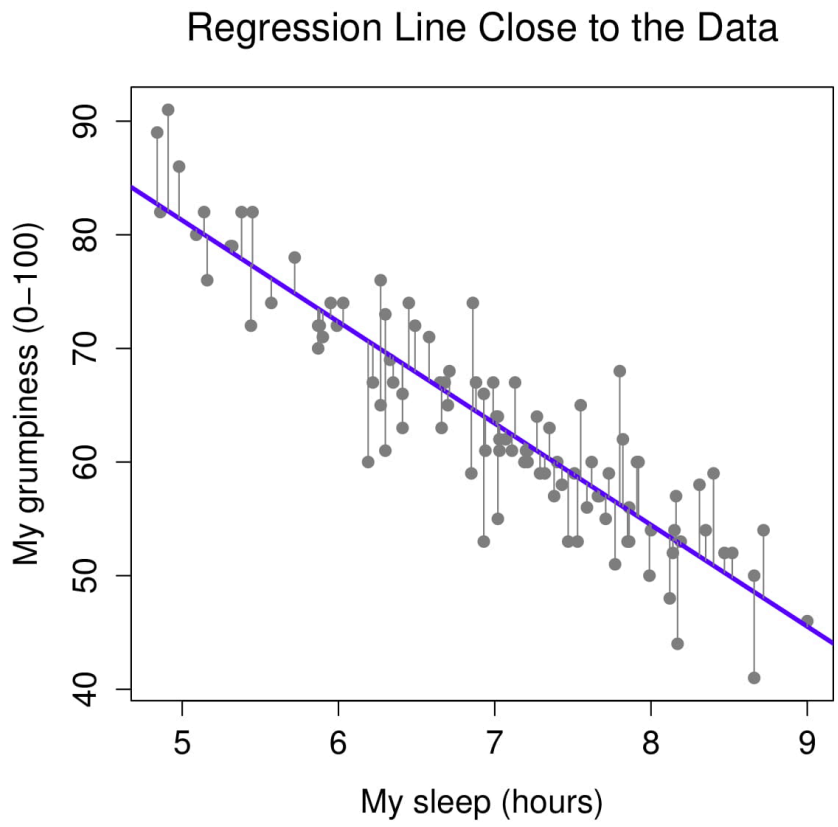
Y is the outcome variable and X is our predictor. There are two **regression coefficients**: b_0 is the intercept (the value of Y when X = 0 – the point at which the regression line crosses the y-axis) and b_1 is the regression coefficient for the predictor. b_1

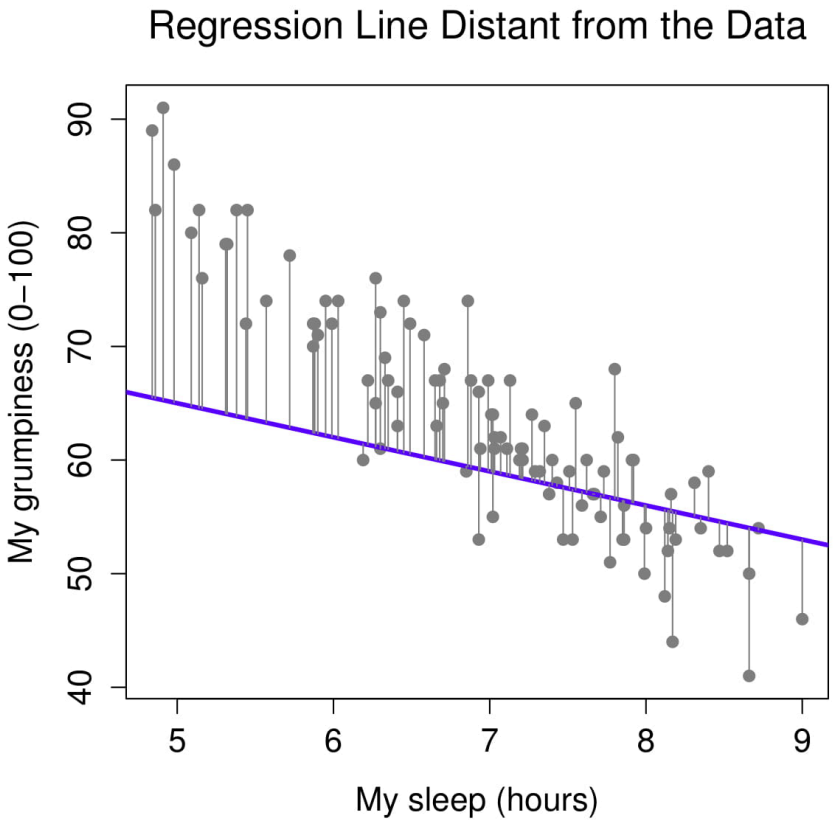
indicates the slope of the regression line (for each unit increase in X, how much does Y increase by?) and the direction and strength of the relationship.

For any given datapoint, the $error_i$ term is called a **residual**. A residual is just the distance between the observed value for that entity and the predicted value, based on the regression model. See the dragon example below!



How are b_0 and b_1 determined? We determine our equation for the line from the scatterplot of scores by figuring out the line that fits closest to all data points. The regression line is the line that minimizes residuals – i.e., results in the smallest distance between the line and data points. Let's visualize the regression line for how Dan's sleepiness affect Dan's grumpiness. Below on the left, we see the regression line (in purple) is very close to the data points and the residuals (the grey lines between the purple line and the data points) are smaller. On the right, we see the regression line is far from a lot of the data points and the residuals are larger.


Regression Line Close to the Data
My grumpiness (0–100)
My sleep (hours)
40
50
60
70
80
90
5
6
7
8
9


Regression Line Distant from the Data
My grumpiness (0–100)
My sleep (hours)
40
50
60
70
80
90
5
6
7
8
9

The specific method to find this straight line is called the **method of least squares**. This is because the line obtained minimizes the sum of the **squared residuals** or the squared deviations from the line (fun fact: the method uses the squared residuals and not the raw residuals because the sums of the raw residuals will just be 0!). Fortunately, we do not have to sit there with a ruler and calculator and do this by hand – statistical software, including jamovi, does it for us!

In our graphs above, *Total error* = $(observed_i - model_i)^2$. In other words, the total error in a model is just the sum of the (observed values minus the model values) squared. So the observed values are the datapoints, and the model values are given by the regression line. The difference between them is called a residual. This total error is the sum of the squared residuals, SS_R, and so the regression line minimizes SS_R.

Fit of the Regression Model

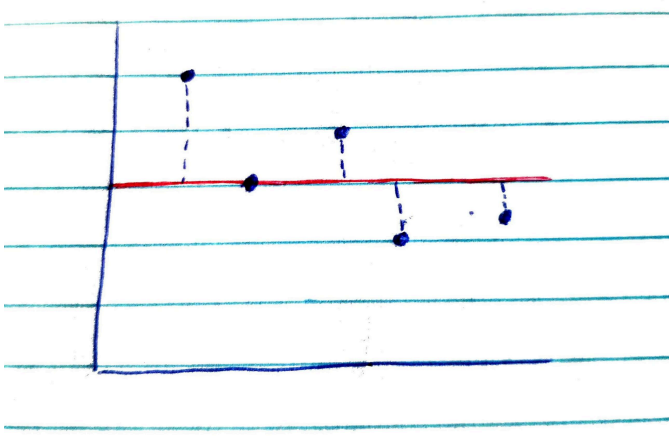
Model Fit

When we have created our regression line, we want to know how well this model fits the data. In some cases, when the datapoints are all very close to the regression line, the model fits the data very well. In other cases, when the datapoints are scattered more widely around the regression line, the model does not fit the data as well. We can quantify this fit in two ways: R^2 and *F*. Before computing these, we need to look at sums of squares in regression:

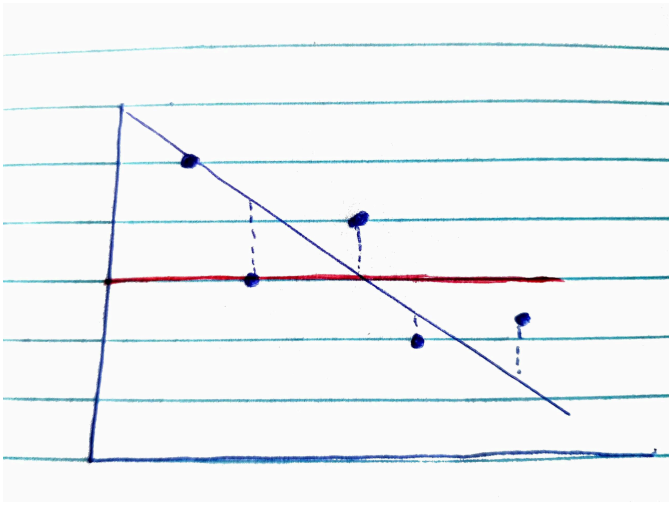
You may recall that in the chapter on one-way ANOVA, we said that:

> The total variance, sum of squares total, SS_t, is an indication of how much all the scores in the experiment vary around the grand mean (i.e., the mean of all the scores). The model sum of squares, SS_m, sometimes reported as SS_b (between-groups sum of squares, for the one-way ANOVA) reflects how much the group means vary around the grand mean. And the residual sum of squares, SS_r, reflects how much participant scores vary around their own group means. So, SS_r is the amount of variability that is left over when we use the model (i.e., the group means) to predict scores, compared to when we just use the grand mean to predict scores.

You might be wondering how this works with regression in the case where we have a continuous predictor. It can help to visualize it. Let's take our mother grumpiness data again. To simplify things for the example, we'll just look at a subset of the data (5 data points), but in reality, all this would actually apply to the whole dataset. In each image below, the blue dots represent each participant's datapoint (X = sleep, Y = grumpiness). The horizontal red line represents the mean grumpiness.



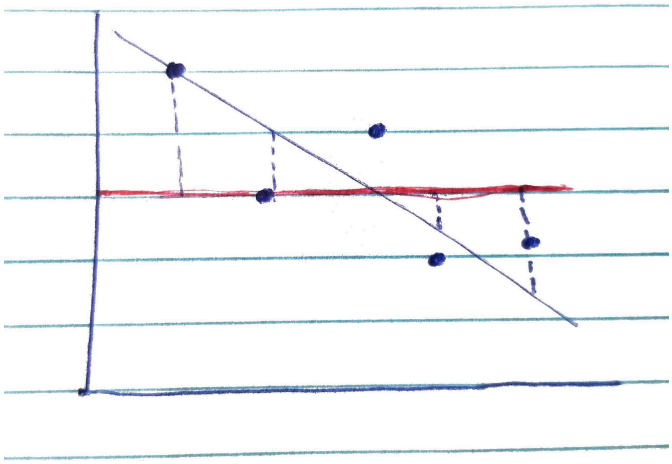
SS_t First, let's imagine a situation where we just use the mean grumpiness scores as our predictor of Y. In other words, if we wanted to predict a grumpiness score in the dataset, we imagine that we do not know anything about the predictor (sleep). If we do not know anything about the predictor, the best information we have to go on if we want to guess what the grumpiness score might be on any given day is the mean of the grumpiness scores. The dashed lines between the datapoints and the red line for the mean grumpiness scores indicate how much error there would be if we used the mean of Y to predict Y scores. If we square and sum those squared errors, we compute SS_t, or the total sum of squares. SS_t tells us how much error there is if we just use the mean of Y to predict Y-scores. SS_t indicates the amount of variability of the Y scores around the mean of Y.



SS_m Finally, we can measure the distance between the regression line (the predicted Y-values) and the mean of Y, for each datapoint. SS_m is the sum of those squared values – the model sum of squares. SS_m reflects the amount by which error is reduced by using the model (the regression line) to predict Y-scores instead of using the mean of Y to predict Y-scores. Again, remember that we are using regression to be able to predict Y, from X. SS_m tells us how much improvement we get in our predictions when we use the predictor to predict Y-scores, compared to when we just use the mean of Y to predict Y-scores.

We can use SS_t, SS_r, and SS_m in different ways to assess the fit of the model, the regression line, to the data.

SS_r Next, let's imagine we use the regression line to predict scores. We (or, in our case, jamovi) finds the best fit line. However, unless there is a perfect relationship between X and Y (a highly unlikely scenario) the datapoints will not lie on the regression line, but above and below it. The vertical distance between each datapoint and the regression line (indicated by the dashed lines in this figure) is the residual. SS_r, the sum of the squared residuals, indicates the amount of error we have when we used the regression line to predict the Y-scores. SS_r indicates the amount of variability of the Y scores around the regression line.



Model Fit: R^2

One measure of model fit is R^2. R^2 represents the proportion of the variance in Y, the outcome, accounted for by the regression model. In the case where there is just one predictor, it is the proportion of the variance in Y explained by the X, the predictor. It is computed as follows:

$$R^2 = \frac{SS_m}{SS_t}$$

As you can see from the equation, R^2 is calculated as the variance explained by the model, divided by the total variance (around the mean of Y) – hence, proportion of the variance!

We can write this value as calculated, or convert it to a percentage by multiplying by 100. For example, it is acceptable to write either R^2 = .36 or R^2 = 36%. This is considered to be a measure of **effect size**.

Model Fit: *F*

Another way to quantify model fit is to compute *F*. As we saw in the earlier chapters on ANOVA, *F* is computed as follows:

$$F = \frac{MS_m}{MS_r}$$

In other words, *F* the improvement due to the model, divided by the amount of error that is remaining (the difference between the model and the observed data). If we have a good model, with all the datapoints close to the dotted line, then MS_m will be large and MS_r will be small, and so *F* will be large. We can test if the *F* is significant (jamovi will do this for us). If it is significant, we shall conclude that the model fits the data well and that the regression line does a good job of describing the relation between X and Y.

Testing Individual Predictors

We can also test whether individual predictors in our model are significant predictors of the outcome. This is particularly useful when we have multiple regression (i.e., when there is more than one predictor variable), because we can test the extent to which an individual predictor predicts the outcome, while controlling for all other predictors in the model. However, we can also do this for simple regression (i.e., with one predictor), but the results in this case will just tell us the same as the *F*.

You will recall that our regression model looks like this:

$$Y_i = (b_0 + b_1X_i) + error_i$$

You may also remember that b_1 is the slope of the regression line (for each unit increase in X, by how many units does Y increase). We can test if the b_1 in our model is significantly different from zero. If our best fit line is just a horizontal line through the dataset (like the red line for the mean, earlier in this section), then, the value of b_1 is going to be zero. The more Y changes with each change in X, the more the value for will b_1 move away from zero (either becoming more positive or more negative), and at some point it will be significantly different from zero. You will see that in jamovi we can run a *t*-test to test if the b_1 slope is significantly different from zero.

In Practice: Regression

Let's run an example with data from lsj-data. Open data from your Data Library in "lsj-data." Select and open `parenthood`. This dataset includes the sleep quality of both Dan and Dan's baby, Dan's grumpiness, and the day of the data collection from 1-100.

We'll be testing how Dan's quality of sleep predicts Dan's grumpiness (later in this chapter we shall also look at how to add a second predictor).

1. Look at the Data

Our data set-up for regression depends on the type of regression and type of data, but in general we'll have one column of our continuous DV and one or more columns of our IV(s). Once we confirm our data are entered and set up correctly in jamovi, we should look at our data using descriptive statistics and graphs, for dan.sleep and dan.grump. The table for the descriptive statistics shows that we have 100 cases and no missing data. The means, medians, standard deviations, and variances are then shown, followed by the minimum and maximum values.

2. Check Assumptions

There are several assumptions for regression and some of them differ a little from some of the assumptions we have encountered up until now. Let's briefly describe each assumption and then we shall look at how to test them in jamovi.

1. The DV is continuous and the IVs are either continuous, categorical, or ordinal.
2. The relation between the variables is linear.
3. Independence of residuals: the residuals are independent – uncorrelated.
4. Normality of residuals: the *residuals* are normally distributed.
5. Homogeneity of the variance (homoscedasticity): at each level of the predictor variable, the variance of the residuals should be constant.
6. No outliers with high leverage: the model is not strongly influenced by a small number of data points (that change the slope of the regression line).

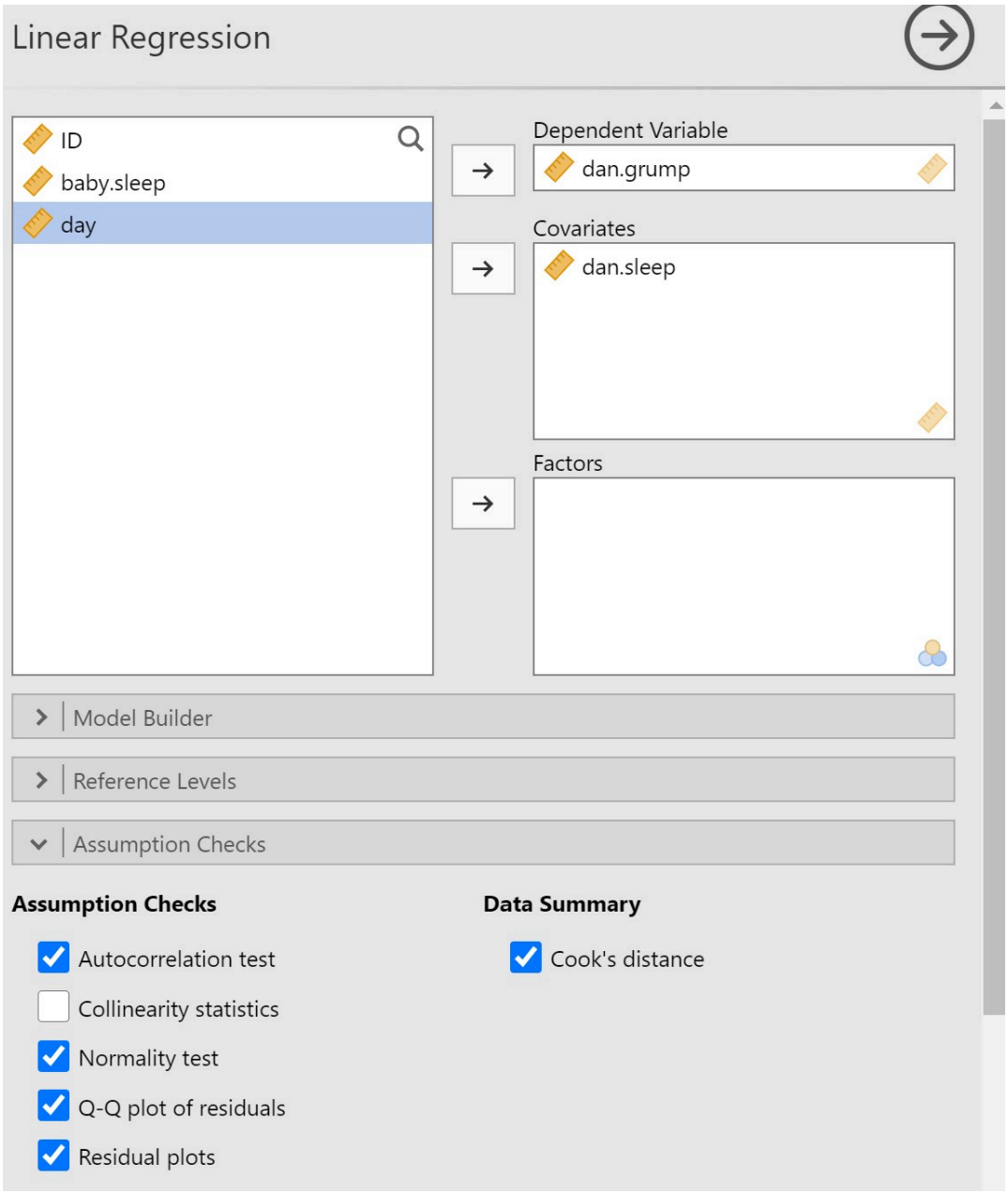
We need to know how the data were collected to check assumption 1, but the rest of the assumptions can be checked in jamovi.

Linear Relationship Between Variables

As we did for correlation, we should check that the relation between X and Y is linear by plotting a scatterplot (either using the scatr add-on module or through the correlation analysis). We saw in the last chapter that the relationship between Dan's sleep and Dan's grumpiness is in fact linear, so let's go on to the next assumption.

To test the next set of assumptions, you'll need to go into the Regression analysis in jamovi. Select dan.grump

as the Dependent Variable (of course it's not really a dependent variable if we did not manipulate something, but that's what jamovi calls the outcome variable). Select dan.sleep as the Covariate (the name jamovi gives to continuous predictor variables). Under Assumption Checks, select Autocorrelation test, Normality test, Q-Q plot of residuals, Residual plots, and Cook's distance. Your set-up in jamovi should look like this:



Independence of Residuals

The Durbin-Watson test for autocorrelation tests for independence of residuals. We want the Durbin-Watson statistic to be as close to 2 as possible. Values less than 1 or greater than 3 are problematic and indicate we are violating this assumption. In our case, the DW test statistic is 2.12 and so very close to 2. Furthermore, jamovi provides a *p*-value and the *p*-value is greater than .05 so the test statistic is not statistically significant, further supporting that we meet the assumption that our residuals are independent.

Durbin–Watson Test for Autocorrelation

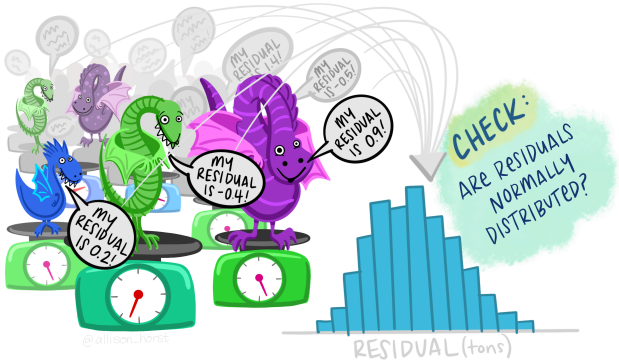
Autocorrelation	DW Statistic	p
−0.0706	2.12	0.528

[3]

If you violate this assumption, it's likely a function of how your data were collected (e.g., you have nested data). We won't be covering what to do in these cases, but if you have nested data you may be interested in multilevel or hierarchical modeling (MLM/HLM).

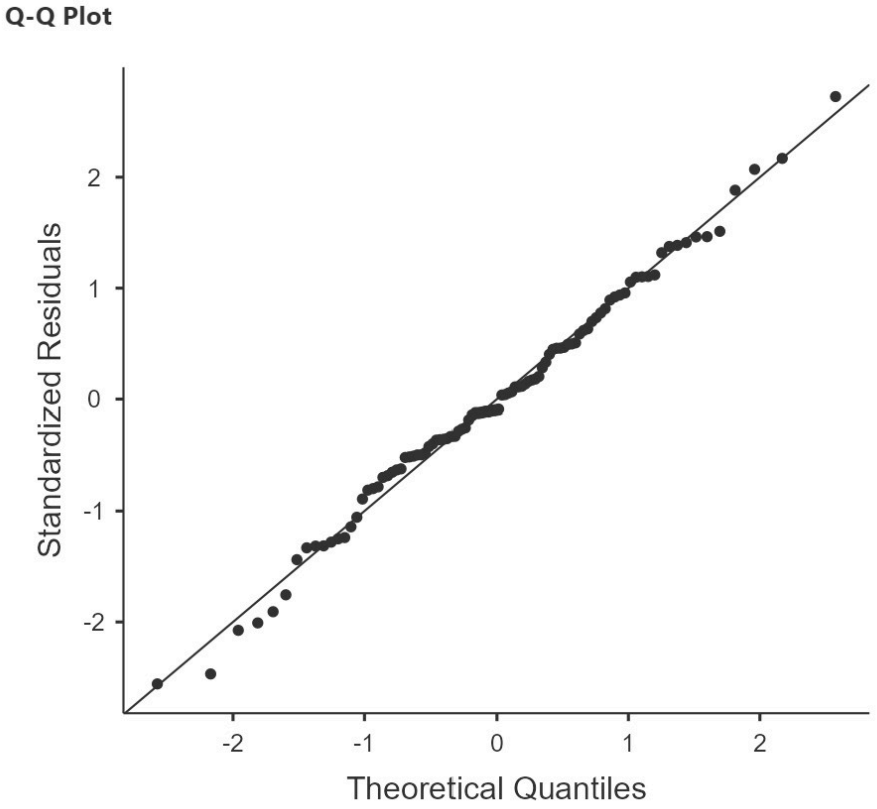
Normality of Residuals

"Artwork by @allison_horst" (CC BY 4.0)



Together, the normality test and Q-Q plot of residuals allow us to assess whether the residuals are normality distributed. The Shapiro-Wilk test simply tests whether the distribution of the residuals is significantly different from a normal distribution. It has the usual caveats as we have discussed in other chapters (i.e., may not be meaningful with small or very large samples), so it is a good idea to use it in tandem with the Q-Q plot. The Q-Q plot for regression shows the standardized residuals (i.e., each residual is converted to a *z*-score) plotted as a function of their theoretical quantiles (what we would expect the residuals to be if they were normally distributed)[1].

1. If you want to dig a bit deeper into Q-Q plots to understand better how they are created and see some examples of normal and non-normal plots, see https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0.

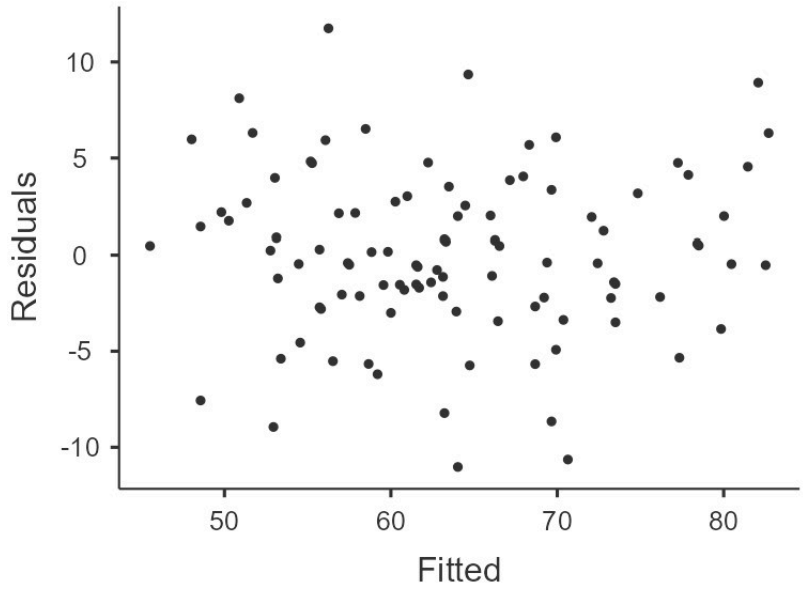


In our sleep-grumpiness example, the Shapiro-Wilk test is not significant and the Q-Q plot also looks like the dots fall pretty close to the straight line, so we can assume normality.

Homoscedasticity

To assess homoscedasticity, we examine the Residual Plots. You will get one plot of the overall model (Fitted) and one for each of your variables (DV and IV(s). We only focus on the Fitted residuals, shown below. In these plots, we want our data to look like a random scattering of dots even dispersed around zero on the y-axis. The plot below, from the sleep-grumpiness dataset, is a good example of homoscedasticity. The datapoints are randomly scattered around with no particular pattern to them.

Residuals Plots

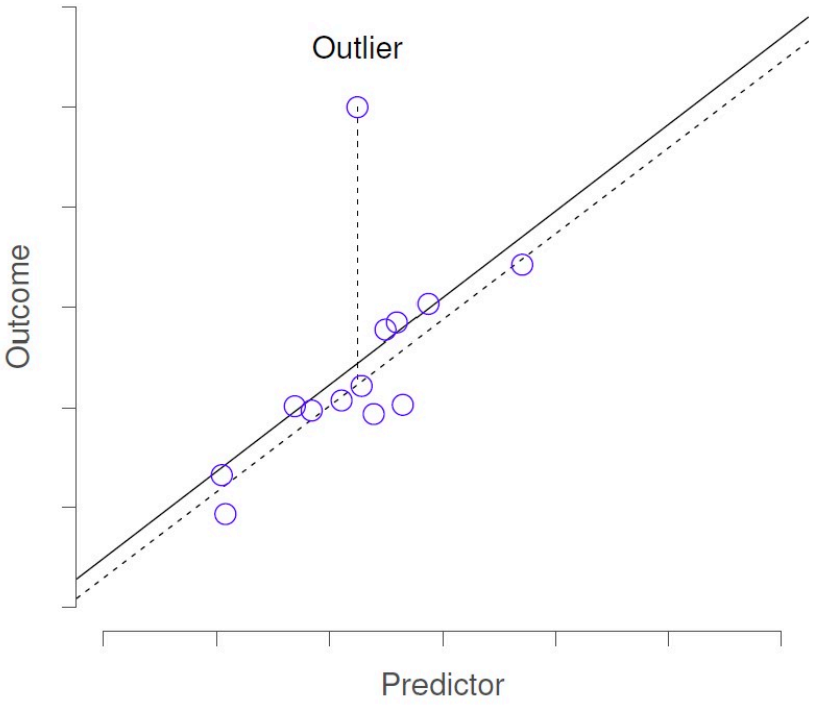


On the other hand, if we have a funnel or fan shape, a bow tie, or some other kind of pattern, then we have heteroscedasticity. (For examples of heteroscedasticity, see this here, approximately 3/4 of the way down the page you will see an example of homoscedasticity – random cloud, and two examples of heteroscedasticity – bow tie shape, fan shape.)

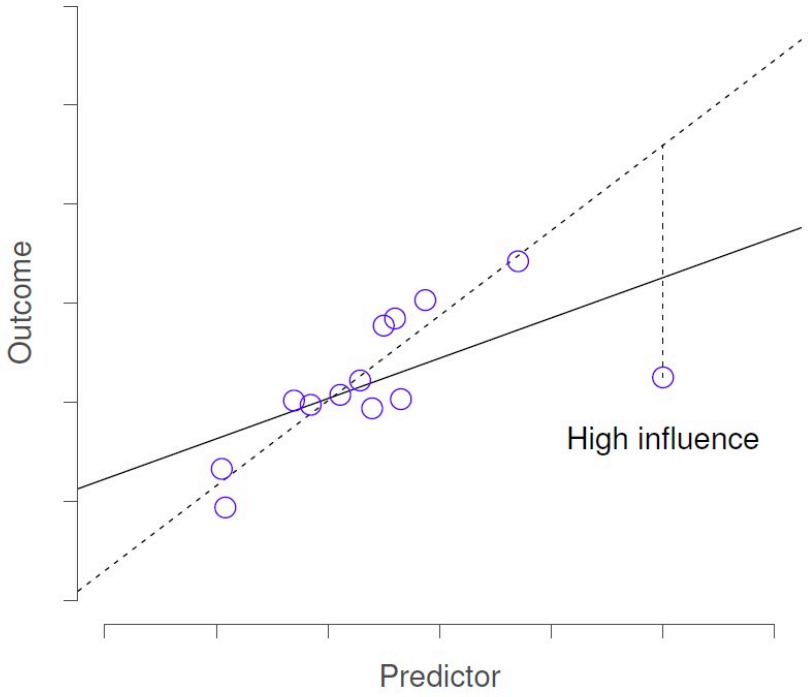
No Outliers with High Leverage

Outliers are just datapoints that are far from the rest of the dataset. Outliers with high leverage are those that additionally "pull" the regression line, and therefore result in a model/regression line, that does not actually accurately represent the majority of the data in the sample.

Below, the first image shows an outlier. If this outlier were removed from the dataset, the regression line (dashed line) would not change much from the original regression line (solid line). The intercept would be slightly lower, but the slope of the line would not change.



In this next figure, the outlier has high leverage and so is called a “high influence” datapoint. Removal of this datapoint would substantially change the slope of the regression line.



We can look at Cook’s distance to check for the presence of these “high influence” (outlier with high leverage) observations. If Cook’s distance is greater than 1, it is often considered large. The table below shows the summary statistics for Cook’s distance for our dataset. We can see the mean, median, standard deviation, and range for all the Cook’s distance values.

Cook's Distance

			Range	
Mean	Median	SD	Min	Max
0.0108	0.00282	0.0188	8.89e−6	0.121

According to this summary, the maximum Cook's distance score is 0.121. Therefore, we do not have any high influence observations.

What if we did? We can go to the Save options in our regression analysis and select Cook's distance. This will add a new column in the dataset indicating the value for Cook's distance for each row in the data. We can then try running the regression again with the observations with Cook's distance value > 1 excluded (using a filter) to examine how that affects the regression coefficients. If it does have a substantial effect, then we need to dig further to understand why that observation is so different from the others (go back to your notes on data collection – was that a participant who was not paying attention or who appeared not to comprehend the instructions, for example?). Note that if we are going to exclude cases, we need to have a good reason for doing so.

3. Perform the Test

1. From the 'Analyses' toolbar select 'Regression' – 'Linear regression.'
2. Move your outcome variable `dan.grump` into the Dependent Variable box and your predictor variable into either Covariates (if it is a continuous variables) or Factors (if it is a categorical variables). In this case, all predictor is continuous so move `dan.sleep` to the Covariates box.
3. If you have categorical predictors with more than two levels, you will use the Reference Levels drop-down menu to specify what you want your reference level to be and whether you want the intercept to be the reference level or the grand man. More information on categorical predictors is beyond the scope of this class.
4. Under Assumption Checks, check all the boxes (except Collinearity statistics – we'll use that one later when we look at multiple regression)!
5. Under Model Fit, select R, R-squared, Asjusted R-squared, and F test.
6. Under Model Coefficients, select Standardized Estimate and the CI for the Standardized Estimate.
7. Optionally, you can ask for plots and tables of the estimated marginal means.

4. Interpret Results

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.903	0.816	0.814	435	1	98	< .001

Model Coefficients - dan.grump

						95% Confidence Interval	
Predictor	Estimate	SE	t	p	Stand. Estimate	Lower	Upper
Intercept	125.96	3.016	41.8	< .001			
dan.sleep	−8.94	0.429	−20.9	< .001	−0.903	−0.989	−0.817

R, R-squared, and adjusted R-squared: We get our R and R-squared values (R-squared literally being R squared). Remember back to correlation: R-squared is the *proportion* of variance in the dependent variable that can be accounted for by the predictor(s). In this case, Dan sleep duration predicts 82% of the variance in Dan's grumpiness.

However, more commonly we report the adjusted R-squared value, which adjusts the R-squared value based on the number of predictors in the model. Adding more predictors to the model will *always* cause R-squared to increase (or at least not decrease) so it's important that we control for that using an adjustment. It's interpreted basically the same, just adjusted for bias. I encourage you to use the adjusted R-squared, *especially* if you have lots of predictors in your model.

Overall Model Test: We also get an *F*-test for the overall model. If you want, you can get the full ANOVA test by selected ANOVA test under Model Coefficients. This is how we know if the overall model is statistically significant. In our case, our *F*-test is statistically significant so we know that the predictor significantly predicts our dependent variable.

Model coefficients: Just like in ANOVA, we first examine if the overall model is significant (overall model test) and then look at individual predictors. Each variable–our intercept and predictor (or predictors when we come to multiple regression, later)–have an Estimated value (the coefficient) and an associated *t*-test. The Estimate value for the intercept just tells us where the regression line crosses the Y-axis. The coefficient for dan.sleep tells us for each unit increase in dan.sleep, whow much does dan.grump change by. We can see that for each hour increase in Dan's sleep, Dan's grumpiness decreases by 8.94. In this case, the *t*-test also shows us that Dan's sleep significantly predicts Dan's grumpiness.

Standardized coefficients: We also asked for standardized estimates (Stand. Estimate), which we get in our model coefficients table. These are *standardized* so that we can compare them to other variables, which may have been measured on a different scale. They give us an idea of the *strength* of the relationship between that predictor and the outcome. Larger values = bigger effects. The standardized estimate is called the standardized regression coefficient or Beta (β), whereas the unstandardized estimate is just called the regression coefficient or B (the letter B, not Beta). We use the standardized estimates to compare the strength of the coefficient to other predictors and we use unstandardized estimates to write our linear equations and predict the value of Y given values of the X.

What about the intercept? You might be wondering what we do with the intercept. Typically, nothing. We only use it to create our equation so that we can predict Dan's grumpiness based on Dan's sleep and the baby's sleep. For example, our equation from our data isL

dan.grump = 125.96 -8.94(dan.sleep)

If Dan had five hours sleep, we would expect Dan's grumpiness to be:

dan.grump = 125.96 -8.94(5) = 81.26

Note: we can only predict values of Y for values of X that are within the range of X values in our dataset. It is very risky to try to predict Y for values of X beyond the X values in our dataset because we do not know if the linear relationship continues for larger or smaller values of X.

Write up the results in APA style

For simple regression (i.e., with a single predictor) we can write the results in text format, as in the example below:

> We explored how hours of sleep for Dan, over 100 days, predicted Dan's daily grumpiness using simple linear regression. Dan's hours of sleep significantly predicted Dan's grumpiness, $F(1, 98) = 435$, $p < .001$, adjusted $R^2 = .81$. For each extra hour of sleep Dan obtained, Dan's grumpiness decreased by 8.94 points ($SE = 0.43$), $\beta = -0.90$, 95% CI [-0.99, -0.82].

Note 1: I did not report the *t*-test for the regression coefficient, B, because in the case of simple regression, this will be the same as the result of the *F*-test. For multiple regression, we should report the *t*-tests for the individual predictors.

Note 2: in many of these write-ups I did not include anything about assumption checking. I normally write up that information as part of my analytic plan in my methods section (e.g., "I checked for multivariate outliers using Cook's distance."). Included in this section, I explain what I will do if I do not meet various assumptions. Then, if I don't meet the assumption in the results section I explain that I did not meet the assumption, explain the results if necessary, explain what I did, and then give the results. In this case, we met all the assumptions (that presumably I described in my methods section) and therefore went straight to the results.

Extending the Regression Model: Multiple Regression

Multiple regression is used when we have multiple predictors (continuous and/or categorical) and a single, continuous outcome variable. For example, in the parenthood datafile that we have been working with this chapter, there is another predictor, baby.sleep, which reflects how much sleep the baby had on each of the 100 days. We can add this predictor to the model. When we have two or more predictors, we can test the extent to which one X variable predicts Y, *while controlling for* the other predictors in the model.

Much of the process for running multiple regression is similar to that for simple regression, but there are some differences, so let's go through it step-by-step.

1. Look at the Data

This is the same as for simple regression. This time, include both dan.sleep *and* baby.sleep as Covariates in jamovi, at the same time.

2. Check Assumptions

We shall test all the same assumptions, and add Collinearity statistics. It is an assumption of multiple regression that the predictor variables are not substantially correlated with each other, i.e., no multicollinearity.

Multicollinearity is a problem for three reasons:

1. **Untrustworthy *B*s**: As multicollinearity increases, so do the standard errors of the *B* coefficient. We want smaller standard errors, so this is problematic.
2. **Limits the size of R**, and therefore the size of R^2, and we want to have the largest R or R^2 possible, given our data.
3. **Importance of predictors**: When two predictors are highly correlated, it is very hard to determine which variable is more important than the other.

Multicollinearity is simply that multiple variables are correlated. We can first just look for general *collinearity*, or the correlations between all our predictors, using the correlation matrix in jamovi. Any correlations greater than .8 or .9 are problematic. You would either need to drop one variable or combine them into a mean composite variable.

However, to test for *multicollinearity*, we examine the VIF and Tolerance values. VIF is actually a transformation of Tolerance (Tolerance = 1/VIF and VIF = 1/Tolerance). In general, we want values of 10 or lower for VIF, which corresponds to Tolerance values greater than .2.

Collinearity Statistics

	VIF	Tolerance
dan.sleep	1.65	0.606
baby.sleep	1.65	0.606

[3]

In our data, our VIF is 1.65 and Tolerance is .61, so we satisfy the assumption of no multicollinearity.

3. Perform the Test

This is the same as before. Remember to ensure that both dan.sleep and baby.sleep are in the Covariates box.

4. Interpret Results

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.903	0.816	0.812	215	2	97	< .001

Model Coefficients - dan.grump

						95% Confidence Interval	
Predictor	Estimate	SE	t	p	Stand. Estimate	Lower	Upper
Intercept	125.9656	3.041	41.4231	< .001			
dan.sleep	−8.9502	0.553	−16.1715	< .001	−0.90475	−1.016	−0.794
baby.sleep	0.0105	0.271	0.0388	0.969	0.00217	−0.109	0.113

For multiple regression, the R-squared tells us the proportion of the variance in Y explained by all the predictors, X_1, X_2, and so on, according to how many predictors we have in the model. Similarly, the *F* statistic give us an overall indication of model fit. The *F* is significant, so that we know that, together, Dan's sleep and the baby's sleep significantly predict Dan's grumpiness. However, these statistics do not tell us how well each individual predictor predicts the outcome. We have to look at the Model Coefficients table for that. In the second table, we see the regression coefficients for each predictor separately, and a *t*-test for each of those coefficients to tell us if they each significantly predict Dan's grumpiness, *while controlling for* the other predictor(s). We can see that dan.sleep is a significant predictor of dan.grump, but baby.sleep is not.

If we wanted to predict dan.grump based on any given dan.sleep and baby.sleep scores, we could use our regression equation:

dan.grump = 125.97 – 8.95(dan.sleep) + .01(baby.sleep)

If Dan's sleep was 5 and baby's sleep was 8, then we'd expect Dan's grumpiness to be:

y = 125.97 – 8.95(5) + .01(8) = 81.3

The rest of the interpretation is the same as for simple regression.

Write Up the Results in APA Style

When we have multiple regression, we have a few more bits of data to write up. With two predictors, we might still choose to report the results in text form, but with three or more predictors, we would probably report our regression coefficients in a table. For the purposes of illustration, I'll show you how to do that. I created the table by copying and pasting the model coefficients table from jamovi into Excel, and then editing it within Excel to fit APA format. This time, we shall report the *F* results *and* the results of the *t*-tests because they give us different information (*F* tells us if the overall model is significant; *t* tells us if the individual predictors are significant, while controlling for other predictors in the model).

> We explored how hours of sleep for Dan and the baby, over 100 days, predicted Dan's daily grumpiness using multiple regression. Together, Dan's hours of sleep and the baby's hours of sleep significantly predicted Dan's grumpiness, $F(2, 97) = 215$, $p < .001$, adjusted $R^2 = .81$. Only Dan's sleep was a significant predictor of Dan's grumpiness, while controlling for baby's hours of sleep (see table 1 below).
>
> **Table 1**
>
> *Dan's and Baby's Sleep as Predictors of Dan's Grumpiness*

Predictor	*B* (*SE*)	*t*	*p*	β (95% CI)
Intercept	125.97 (3.04)	41.42	< .001	
Dan's sleep (hours)	-8.95 (0.55)	-16.17	< .001	-0.90 (-1.02, -0.79)
Baby's sleep (hours)	0.011 (0.27)	0.039	.969	0.002 (-0.11, 0.11)

Categorical Predictors

If you have a single predictor that is categorical, or two or more predictors that are categorical, then regression will simply give you the same results as ANOVA and you should just run an ANOVA. However, if you have one or more continuous predictors, *as well as* one or more categorical predictors (that are between-subjects variables) then regression will allow you to test these predictors simultaneously. In addition, you can build a model that includes interaction terms. You'll notice in the example above that there was no interaction, but we *could* include an interaction between Dan's sleep and baby's sleep, if we wished. Let's say we have a categorical predictor, like whether or not Dan ate breakfast on a given day. We could test whether or not eating breakfast, Dan's sleep, and baby's sleep as predictors of Dan's grumpiness using regression. jamovi (and other statistical software packages) rely on using something called "dummy coding" when including categorical predictors in a regression model. We could also examine some or all of the two- and three-way interactions by building a custom model in the Model Builder of regression jamovi.

These more advanced techniques are beyond the scope of this book, but it is nice to know they are available because they expand the kinds of research questions you can answer with your data. Ask your statistics instructor if you want to use them! If you are keen to learn more right now, you can also check out this website for a straightforward example of regression with one continuous and one categorical predictor in jamovi.

CHAPTER 12: CHI-SQUARE

Most of this chapter is based on “Statistics with jamovi” by Dana Wanzer, with some minor changes .

Chi-Square

The chi-square (pronounced like kai, not like the tea) is a categorical data analysis which is simply data analysis with categorical data. It's usually used with nominal data, although there are a couple tests we may use with ordinal data. There are two basic types of chi-square tests we'll be covering:

1. χ^2 **goodness-of-fit**: used with one variable to find if the observed frequencies match the expected frequencies
2. χ^2 **test of independence (or association)**: used with two variables to find if the observed frequencies match the expected frequencies. In other words, are the two nominal variables independent or associated with one another?
    1. **Fisher's exact test**: This is an alternative to the χ^2 test of independence that we use when our frequencies are small.
    2. **McNemar's test**: This is an alternative to the χ^2 test of independence when we have a 2×2 repeated-measures design. For example, perhaps we examine pass/fail rates before and after a training.

Because these tests are all with categorical data, there are no assumptions about the distribution of the data. For that reason, these are all *non-parametric statistics*.

Entering Data in jamovi

One thing to note that is unique about the chi-square is that you can either setup your data in the raw format or you can use the frequency tables as your data. If you use the frequency table, then you can move the counts/frequency variable into the Counts box in either the goodness-of-fit or test of independence analyses.

Chi-Square Goodness of Fit

The χ^2 (chi-square) goodness of fit tests whether an observed frequency distribution of a nominal variable matches an expected frequency distribution. Our hypotheses for the test are as follows:

- Null hypothesis: The observed frequencies match the expected frequencies. In other words, the frequencies of the variable are what we would expect.
- Alternate hypothesis: At least one observed frequency doesn't match the expected frequency. In other words, the frequencies of at least one level of the variable are not what we would expect.

Note that these are not how you should describe your hypotheses! You should specify your hypotheses in relation to the nature of your data. For example, if we have a deck of cards and want to see if people don't choose cards randomly, the null hypothesis would be that there is a 25% probability of getting each hearts, clubs, spades, and diamonds.

1. Look at the Data

Let's run an example with data from lsj-data. Open data from your Data Library in "lsj-data." Select and open "randomness." This dataset has participants pull two cards from a deck. For now, we're just going to work with `choice_1`. We're interested in finding out if participants pull cards randomly from the deck.

Data Set-Up

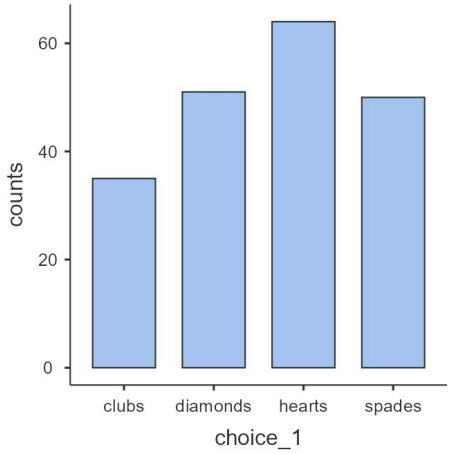
Our data set-up for a chi-square goodness-of-fit test is pretty simple, We just need a single column with the nominal category that each participant is in. In the example, the nominal category we are going to work with is choice_1 (which suit the participant chose).

Describe Your Data

Once we confirm our data is setup correctly in jamovi, we should look at our data using descriptive statistics and graphs. First, our descriptive statistics are shown below. With nominal variables like `choice_1`, we should request Frequency tables, not descriptive statistics like the mean and median. The mean for `choice_1` would be, quite frankly, meaningless. What's the average card type? It can't exist. So we do frequencies instead. Under Plots we can select a Bar plot to visualize the data appropriately.

Frequencies of choice_1

Levels	Counts	% of Total	Cumulative %
clubs	35	17.5 %	17.5 %
diamonds	51	25.5 %	43.0 %
hearts	64	32.0 %	75.0 %
spades	50	25.0 %	100.0 %



Notice how jamovi is pretty smart here and knows not to give us the mean, median, minimum, and maximum. Check the box for Frequency tables to receive those. From our data, we see that most participants pulled a hearts card first (*n* = 64, 32%) followed by diamonds (*n* = 51, 26%), spades (*n* = 50, 25%), and finally clubs (*n* = 35, 18%).

Specify the Hypotheses

We're interested in finding out if participants pull cards randomly from a deck of cards. A typical deck of cards has 52 cards, 13 for each of the four suites (clubs, spades, hearts, diamonds). Because there are 4 suites, then 1/4 is 25% which is our expected frequency of pulling cards randomly from the deck. Under the null hypothesis, we expect that participants pull cards randomly from the deck. In other words, there is a 25% probability of pulling each of hearts, clubs, spades, and diamonds. Under the alternate hypothesis, we expect that participants pull cards not at random from the deck. In other words, participants have a probability other than 25% of pulling at least one of the types of cards.

2. Check Assumptions

The chi-square goodness-of-fit test has just one assumption: **Expected frequencies are sufficiently large**, which is usually greater than 5.

You test for this assumption by checking the "Expected counts" box (see 3. Perform the test, below). You will then see rows of expected counts in your contingency table. Look at the "expected" numbers and check that they are all 5 or greater.

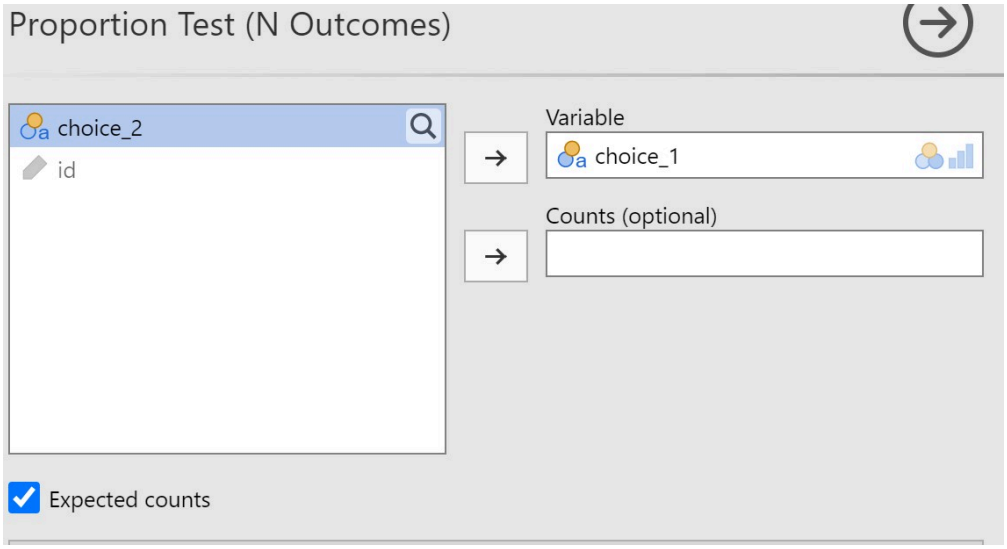
3. Perform the Test

To perform the chi-square goodness of fit test, do the following steps:

1. Go to the Analyses tab, click the Frequencies button, and choose "One sample proportion tests – N outcomes."

2. Move your variable into the Variable box. In this case, move choice_1 into the Variable box.
3. Select Expected counts so you can check for your assumption of expected frequencies.

When you are done, your setup should look like this:



As you will see in the output, jamovi automatically assumed equal proportions of frequencies (in this case 1/4 or 25% chance of pulling each card). However, there might be times when you don't want to make that assumption. Maybe we're testing the whether our sample frequencies match the population frequencies and those are uneven (e.g., whether our 80/20 right/left-handedness split in our sample matches the 90/10 handedness split in the population).

We can use the Expected Proportions in the setup to specify different expected frequencies.

4. Interpret Results

Proportions - choice_1

Level		Count	Proportion
clubs	Observed	35	0.175
	Expected	50.0	0.250
diamonds	Observed	51	0.255
	Expected	50.0	0.250
hearts	Observed	64	0.320
	Expected	50.0	0.250
spades	Observed	50	0.250
	Expected	50.0	0.250

χ^2 Goodness of Fit

χ^2	df	p
8.44	3	0.038

The first table shows us our observed frequencies (our data) and expected frequencies (N/k = 200/4 = 50 which is 25% for each one, like we previously calculated).

The second table gives us our results. Our *p*-value is less than our alpha of .05 so we can reject the null hypothesis that the observed frequencies match our expected frequencies.

Write Up the Results in APA Style

We can write up our results in APA something like this:

> Of the 200 participants in the experiment, 64 selected hearts for their first choice, 51 selected diamonds, 50 selected spades, and 35 selected clubs. A chi-square goodness of fit test was conducted to test whether the choice probabilities were identical for all four suits. The results were statistically significant, $\chi^2(3) = 8.44$; $p = .038$, suggesting that people did not select suits purely at random. Participants chose the hearts (32%) more frequently than expected and the clubs (17%) less frequently than expected.

Note that I described the data in the first sentence, but I could have also described it in more detail in the last sentence as part of my interpretation or I could have even written up the results in a table!

Chi-Square Test of Independence

The X^2 (chi-square) test of independence (or association) tests whether an observed frequency distribution of a nominal variable matches an expected frequency distribution, but unlike the goodness of fit test we are looking at the relationship, independence, or association between two variables. The test of independent tests whether two categorical variables are related or independent.

Our basic hypotheses for the chi-square test of independence is as follows:

- Null hypothesis: the observed frequencies match the expected frequencies. In other words, there are no differences in frequencies of how the levels in one variable relate to the levels in another variable.
- Alternate hypothesis: At least one observed frequency doesn't match the expected frequency. In other words, at least one level has significantly different frequencies in another variable than we would expect.

1. Look at the Data

Let's run an example with data from lsj-data. Open data from your Data Library in "lsj-data." Select and open "chapek9." This dataset indicates the ID number of the participant, the species (robot or human), and their preference of the three things (puppy, flower, or data).

For this example, imagine we are watching a show about the planet *Chapek 9*. On this planet, for someone to gain access to their capital city they must prove they're a robot, not a human. In order to determine whether or not a visitor is human, the planetary beings ask whether the visitor prefers puppies, flowers, or large, properly formatted data files.

Data Set-Up

Our data set-up for a chi-square test of independence is pretty simple, We just need two columns of nominal data, with one row per participant, as we have in the chapek9 example (one column for species and one colour for choice).

Describe the Data

Once we confirm our data is setup correctly in jamovi, we should look at our data using descriptive statistics and graphs. Remember that for nominal variables we should report frequency statistics, not means and medians and such. Bar plots are also a good way of visualizing the data.

Specify the Hypotheses

The question here is whether humans and robots differ in preferring puppies, flowers, or data so we can

determine who is a robot so only robots are let into the city. Therefore, our alternate hypothesis might be something like this: humans and robots have different preferences, or there is an association between species (human or robot) and preference (flowers, data, or puppies).

2. Check Assumptions

1. **Expected frequencies are sufficiently large**, which is usually greater than 5. If we violate this assumption, you can use Fisher's exact test. We test for this assumption by selected "Expected counts" in the Cells tab for the test of independence. You will then see rows of expected counts in your contingency table. Look at the numbers and check that they are all 5 or greater.
2. Data are **independent** of one another, meaning each case contributes to only one cell of the table. If you violate this assumption, you may be able to use the McNemar test. This requires knowing how your data was collected. If it's a within-subjects design with nominal variables, then most likely you want to use McNemar's test. If it's a between-subjects design that should be answered using a chi-square, then you most likely meet this assumption and can perform the chi-square test of independence.

3. Perform the Test

1. Go to the Analyses tab, click the Frequencies button, and choose "Independent Samples – χ^2 test of association."
2. Move your two variables into the rows and columns boxes. In this case, move choice into rows and species into columns. Note that the placement in rows or columns doesn't really matter, but because we typically work with portrait pages I tend to prefer putting in rows whatever variable has more levels. In this case, choice has 3 levels and species only 2 so I like to put choice in rows and species in columns.
3. Under the Statistics tab, select χ^2 under Tests and Phi and Cramer's V under Nominal to get your effect size.
4. Select Expected counts under Cells to test your assumption of expected frequencies. Optionally, you can request the row, column, and total percentages. I often find these easier to report and interpret.
5. Select Bar Plot under plots. You may want to tinker with the settings here: in our case, I recommend using Side by side for Bar Type, and Rows for X-axis (that will put the choice on the x-axis, and I usually like to put the variable with the most levels on the X-axis for ease of interpretation). You can see if you prefer counts or percentages.

Ordinal Variable(s)

If either of your variables are ordinal, instead of selecting Phi and Cramer's V, you should use Gamma or Kendall's tau-b. Kendall's tau-b should only be chosen if you have a square table (e.g., 3×3, 4×4, 5×5, etc.) whereas Gamma can be used with any size table. Kendall's tau-b is also a slightly more conservative estimate compared to Gamma.

4. Interpret Results

Contingency Tables

		species		
choice		robot	human	Total
puppy	Observed	13	15	28
	Expected	13.5	14.5	28.0
	% within row	46.4 %	53.6 %	100.0 %
flower	Observed	30	13	43
	Expected	20.8	22.2	43.0
	% within row	69.8 %	30.2 %	100.0 %
data	Observed	44	65	109
	Expected	52.7	56.3	109.0
	% within row	40.4 %	59.6 %	100.0 %
Total	Observed	87	93	180
	Expected	87.0	93.0	180.0
	% within row	48.3 %	51.7 %	100.0 %

The first table shows us our observed and expected frequencies. We use the expected frequencies to test our assumption that expected frequencies are greater than 5. Our smallest expected frequency is 13.53 so we meet this assumption.

χ^2 Tests

	Value	df	p
χ^2	10.7	2	0.005
N	180		

Nominal

	Value
Phi-coefficient	NaN
Cramer's V	0.244

The second table gives us our results. Our p-value (p = .005) is less than .05 so we can reject the null hypothesis that the observed frequencies match our expected frequencies.

jamovi also gives us our Cramer's V value. Note that it does not provide Phi because we don't have a perfect square table (e.g., 2×2 or 3×3). These are measures of effect size for the chi-square. Cramer's V can be interpreted similar to a correlation (ranges from 0 to 1, with higher scores meaning stronger relationships between the variables).

Write Up the Results in APA Style

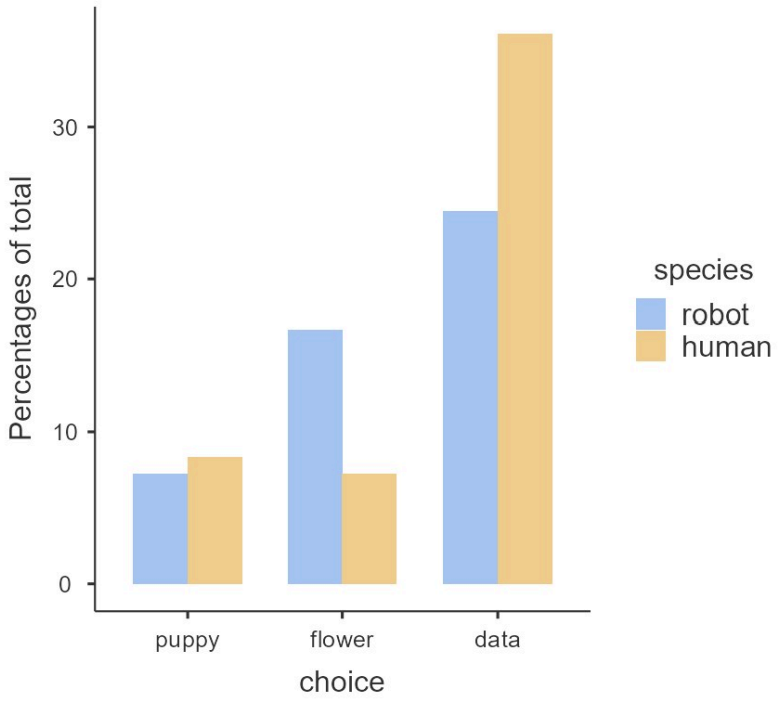
We can write up our results in APA something like this:

> The χ^2 test of independence showed a significant association between species and choice, $\chi^2(2) = 10.72$, p = .005, Cramer's V = .24. Robots, compared to humans, were more likely to say they prefer flowers (robots: 70%; humans: 30%) and humans, compared to robots, were more likely to say they prefer data (humans: 60%; robots: 40%). Robots (46%) and humans (54%) were equally likely to prefer puppies.

Visualize the Results

Here's the plot I was able to produce in jamovi. Note that there are multiple ways to present the plots, so think about which way will best support the results.

Plots



Unfortunately, jamovi does not have any options for changing the colour of the bars or for editing any other details of the plots (e.g., axis labels, legend, etc.). The jamovi plots are fine for this class, but if you are working on a thesis, a poster for presentation at a conference, or a an manuscript for publication, it would be best to use Excel (or some other software for creating your plots).

Fisher's Exact Test

If you violate the assumption that there your expected frequencies are sufficiently large and you have a 2×2 table, you can still perform the test of independence but instead of selecting χ^2 you'll select Fisher's exact test. You'll interpret your results in exactly the same way, but specify that you used Fisher's exact test.

McNemar's Test

McNemar's test is based on the χ^2 (chi-square) test of independence (or association), but is used in a repeated measures or within-subjects design.

We'll go over a brief example so you're familiar with this test, but in practice I don't see this statistic often so we won't go over it any further than this.

1. Look at the Data

For example, suppose we're working with the *Australian Generic Political Party* (AGPP) and your job is to find out how effective AGPP political advertisements are. You gather 100 people and ask them to watch the AGPP ads. You ask participants before and after viewing ads whether they intend to vote for the AGPP.

This data comes from lsj-data. Open data from your Data Library in "lsj-data." Select and open "agpp." This dataset indicates the ID number of the participant and whether they would vote for AGPP before and after viewing the ads.

Data Set-Up

Our data set-up for McNemar's test is pretty simple. We just need two columns of nominal data, with one row per participant and each column being the same variable at two different time points – in our case, we have response_before and response_after as the same variable being measured at two different timepoints.

Specify the Hypotheses

Given you want to see if the AGPP political advertisements are, you want to see if participants are more likely to vote for AGPP after viewing the advertisement. Therefore, the alternative hypothesis is that intentions to vote for the AGPP don't match the expected frequency of intentions to vote.

2. Check Assumptions

You came here because you violated the assumptions of the test of independence's assumption of independence. You should have a within-subjects design to perform this test. We meet this assumption so we can move on!

3. Perform the Test

1. Go to the Analyses tab, click the Frequencies button, and choose "Paired

Samples – McNemar test."
2. Move response_before into rows and response_after into columns. Note that the placement in rows or columns doesn't really matter.
3. Under the Statistics tab, select χ^2 under Tests.
4. Optionally, you can request under to show the row and column percentages.

4. Interpret Results

Contingency Tables

	response_after		
response_before	no	yes	Total
no	65	5	70
yes	25	5	30
Total	90	10	100

McNemar Test

	Value	df	p
χ^2	13.3	1	< .001
N	100		

The first table shows us our observed frequencies.

The second table gives us our results. Our *p*-value is less than .05 so we can reject the null hypothesis that the observed frequencies match our expected frequencies. Unfortunately, looking at our table it also shows that the ads had a negative effect: people were less likely to vote AGPP after seeing the ads.

Write Up the Results in APA Style

We can write up our results in APA something like this:

> McNemar's test indicated that support for AGPP changed from before to after reviewing the AGPP advertisement, $\chi^2(1) = 13.33$, $p < .001$. Most participants continued to not vote for AGPP after the ad (n = 65) and a few continued to vote for AGPP after the ad (n = 5). However, many participants who originally stated they would vote for AGPP changed to no longer voting for AGPP after the ad (n = 25); only five people who originally would not vote for AGPP changed to vote for AGPP after the ad.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*(3), 379–384. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16405133

Field, A. (2013). *Discovering Statistics Using IBM Statistics (4th ed.)*. SAGE Publications.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Navarro, D. J., & Foxcroft, D. R. (2019). Learning Statistics with jamovi: A Tutorial for Psychology Students and Other Beginners. (Version 0.70). DOI: 10.24384/hgc3-7p15

Sauder, D. C., & DeMars, C. E. (2019). An Updated Recommendation for Multiple Comparisons. *Advances in Methods and Practices in Psychological Science, 2*(1), 26–44. https://doi.org/10.1177/2515245918808784

Vermeulen, S. (n.d.). Personal communication.

Wanzer, D. (2022). Statistics with jamovi. https://danawanzer.github.io/stats-with-jamovi/

Glossary

A priori predictions

Predictions made prior to analysing the data, usually based on prior research and theory

Alternate hypothesis

Statement of the expected relationship between variables, usually that there is a relationship or effect

Balancing participant variables

Measuring the participant variable (which is an extraneous variable) and ensuring an even distribution of participants with different values on that variable across conditions, while using random assignment

Between-subjects or independent design

Different entities (in Psychology, usually people) are in two or more experimental conditions/groups

Confound

A variable that varies systematically with the independent variable. It can changes in the dependent variable across conditions, so that one does not know if the different scores on the dependent variable across conditions are due to the manipulation of the independent variable or due to the confound (or both).

Content validity

The extent to which a measure reflects the variable of interest (and not some other construct)

Continuous variables

Allow for fractions (at least in theory)

Correlational research

Variables are measured, not manipulated

Counterbalancing

Change the order of conditions across participants in a balanced way, in a within-subjects design (e.g., half the participants are exposed to condition A followed by condition B; half the participants are exposed to condition B followed by condition A)

Demand characteristics

Features of the experimental situation provide cues to the participant as to how they are expected to behave. The participant may conform to these expectations, leading to loss of internal validity. Thus, demand characteristics may be a confound.

Dependent variable

In an experiment, the variable that is measured and that we expect to change as a result of the different values of the independent variable

Descriptive statistics

Used to summarize, organize, and describe data

Discrete variables

Can only be measured in whole number amounts

Dispersion

Spread of scores in a dataset

Ecological validity

Extent to which research results can be generalized to common real-world behaviours and natural situations (i.e., a special type of external validity)

Experimental research

One or more variables (independent variable) is systematically manipulated to see its (their) effect on another variable (the dependent variable)

External validity or generalizability

Extent to which the results of a study generalize to other populations and settings

Extraneous variables

Any variables that vary at random in the experiment (they could be participant variables or aspects of the experiment itself)

Factorial design

A design in which there are at least two independent variables (e.g., a 3 x 2 factorial design has one independent variable with three levels and another independent variable with two levels)

Familywise or experiment wise error rate

The type I error rate across a family of tests

Hypothesis or prediction

The expectation (written as a statement) of what will happen in the context of a particular study

Independent variable

A variable that is manipulated by the experimenter in an experiment

Inferential statistics

Used on sample data to infer things about the population from which the sample was drawn

Interaction

In a multifactorial ANOVA, the interaction refers to how the effect of one independent variable on the

dependent variable changes according to the level of the other independent variable (it is ***not*** the effect of one independent variable on the other independent variable)

Interval data

Each score indicates an actual amount, and there are equal units separating any two adjacent scores. Zero scores is possible, but does not necessarily indicate a zero amount

Kurtosis

The weight of the tails relative to a normal distribution.
Leptokurtic: light tails; values are more concentrated around the mean
Platykurtic: heavy tails; values are less concentrated around the mean

Limit the population

Do not include in the study participants who have particular scores on an extraneous variable of concern

Main effect

In a multifactorial ANOVA, a main effect is the effect of a single independent variable on the dependent variable

Manipulation check

Checks whether the intended manipulation of the independent variable actually occurred (this is *not* the same as measuring the effects of the independent variable on the dependent variable)

Marginal means

The means of the levels of each variable while collapsing across the levels of the other variable

Matching, matched-groups design

Used in an experiment, to hold a variable(s) constant across groups, where pairs (if there are two groups) of participants scoring the same or similarly on a particular variable are randomly assigned to the different conditions

Mean

Sum of all scores divided by the number of scores

Median

Middle score in a dataset (when scores ordered from lowest to highest)

Mode

Most frequent score

Nominal data

A label is used to describe levels of a variable - numbers do not mean anything in a mathematical sense (they just represent a category)

Null hypothesis

Statement of the expected relationship between variables if there is no actual relationship between the variables; this is usually, but not always a statement that there will be no relationship or effect

Omnibus test

A test that tests for an overall difference between group means, but does not tell us which groups differ significantly from each other

Ordinal data

Data that are rank ordered (e.g., 1st, 2nd, 3rd, etc.)

Outcome or criterion

In non-experimental contexts, the variable that we *think* might be changing as a result of changes in the predictor

p-value

The probability or the likelihood of getting that value for the test statistic or more extreme, in the long run, when the null hypothesis is true

Planned comparisons

Comparisons between pairs of means that are based on a priori predictions (vs. post hoc)

Post hoc tests

Comparisons between pairs of means, conducted when there were no *a priori* predictions

Power

The probability that the statistical test will detect an effect, given that there is an effect in the population

Predictor

In non-experimental contexts, the variable that we *think* might be causing change

Ratio data

Scores measure an actual amount, there is a true zero, and ratio statements can be made

Reliability

Ability of a measure to produce the same results under the same conditions (e.g., specifically, test-retest reliability refers to whether the measurements are stable across repeated testings)

Residuals

Distance between each datapoint's Y-score and the Y-value that would be predicted based on the regression model

Restriction of the range

In a correlational design, when difference between the lowest and highest scores on one of the measured variables is small

Sampling distribution of the mean

The distribution of means that we would get if we randomly sampled an infinite number of times from the population, samples of the same size (the size of the sample in our study), when the null hypothesis is true, and plotted those means on a frequency distribution

Sensitivity

The extent to which a measure allows for precision in measurement (e.g., a rating scale from 1 to 3 is going to be less sensitive, and hence precise, than a rating scale from 1 to 7; in the former, two people might both respond with a low score of 1 but feel quite differently about whatever the item is asking them)

Simple main effects

In a factorial design, the effects of one independent variable on the dependent variable at each level of the other independent variable

Skew

In a non-normal distribution, it is when one tail of the distribution is longer than another

Negative skew: when the tail points to the negative end of the spectrum; in other words, most of the values are on the right side of the distribution

Positive skew: when the tail points to the positive end of the spectrum; in other words, most of the values are on the left side of the distribution

Strong manipulation

Levels of the independent variable are substantially different from one another

Test statistic

systematic variation / unsystematic variation

Theory

a general principle or set of principles that explains known findings about a particular topic

Third variable problem

In correlational research, the difficulty in knowing whether an unmeasured or uncontrolled, third variable, led to an *apparent* association between the two measured variables

Type I error

Concluding that the alternate hypothesis is correct when it is in fact false (false positive, alpha)

Type II error

Concluding that the alternate hypothesis is false when it is in fact correct (false negative, beta)

Within-subjects or repeated measures or dependent design

The same entitites (in Psychology, usually people) take part in all the different conditions

Version History

This page provides a record of changes made to this learning resource, Research Methods and Statistics with jamovi. Each update increases the version number by 0.1. The most recent version is reflected in the exported files for this resource.

If you identify an error in this resource, please report it using the TRU Open Education Resource Error Form.

Version	Date	Change	Details

TRU Open Education Resource Error Form

Report an Error

Name(Required)

First and Last Name

Email(Required)

Material you would like to report an error for(Required)

Where did you find the error?

What did you find and how should we fix it?(Required)

submit