Statistics with jamovi

Dana Wanzer

Last Update: 2021-06-26

Contents

Welcome

This is the website for PSYC 290 and PSYC 790 at the University of Wisconsin-Stout, taught by Dana Wanzer. These resources are aimed at teaching you how to use jamovi and null hypothesis significance testing (NHST) to answer research questions.

Portions of this book may have been adapted from "Learning statistics with jamovi: A tutorial for psychology students and other beginners" by Danielle J. Navarro and David R. Foxcroft, version 0.70. Furthermore, the template and style of this book is from PsyTeachR.

Chapter 1

Introduction

This chapter will introduce you to the course (PSYC 290 or PSYC 790), the instructor (Dr. Dana Wanzer), and the textbook.

1.1 PSYC 290

Welcome to PSYC 290 - Interpreting Psychological Research! I'm excited to teach this course, because as a follow-up to PSYC 190 you'll be learning about how to analyze and interpret statistical results.

1.1.1 Learning objectives

By the end of the course, you can expect to

1. Understand the appropriate statistical procedure to apply in basic psychological research.
2. Understand the logic behind basic inferential statistics.
3. Interpret pertinent statistical information in psychology-related journal articles and other publications.
4. Conduct appropriate statistical tests using statistical software.
5. Interpret and evaluate the results of statistical tests.
6. Report the results of statistical analyses using APA style.

In particular, you're going to learn how to analyze data in a statistical program called jamovi, although occasionally we will practice interpreting output from SPSS and R so you can get some experience interpreting statistical output from other statistical software.

1.1.2 Weekly Schedule

This class is purely online and asynchronous in Spring 2020, meaning we will not meet as a class at any point during the semester. Each week will look pretty similar:

1. You will have assigned readings that may include book chapters, chapters from this online textbook, journal articles, or videos.
2. After completing the readings, you will complete a reading reflection to demonstrate you completed the readings, understand the material and its application, help develop the glossary for this textbook, and ask questions.
3. You will have a practice activity each week to help you understand and apply the material. You can complete these repeatedly until you receive full credit.
4. Optionally, you can attend student hours to work on homework, ask questions, and get extra practice examples to extend your understanding. While they are all optional, they are highly recommended and you will be required to attend at least two student hours throughout the semester.
5. Lastly, you will complete either a homework assignment or exam to test your understanding of the material. Homework assignments can be redone with an additional reflection.

1.1.3 Late assignments and re-doing homework

All assignments can be turned in late for absolutely no penalty; however, there are specific dates that are hard deadlines to help you stay on track in this course. Life happens, and sometimes you won't be able to complete an assignment for a week or two and that is completely understandable.

That being said, you should make every effort to stay on top of the coursework in this class. Dedicate hours to work on course activities so you do not fall behind! Please reach out to me if you start feeling overwhelmed or need help getting back on track. I am here to support you!

Furthermore, if you do not get the grade on the homework assignment that you want you can always re-do the homework assignment for up to full credit. There will be additional work you will need to complete to be able to re-do an assignment; more information on this will be available on Canvas.

1.1.4 Getting help in this class

Come to student hours regularly! The GA and myself are *always* available to help you. We will be scheduling regularly recurring student hours each week so you can come ask questions, get help on your homework, or just have a space to come together to work on your assignments in a dedicated online space.

We will also have an online Microsoft Teams team channel so that you can ask questions there. This way everyone can benefit from the answers to questions

students have!

If you have more personal questions, you can message me on Teams or email me at wanzerd@uwstout.edu.

1.2 Dana, your instructor

My name is Dana Wanzer (pronounced DAY-nuh JUAN-zur) and I started teaching at UW-Stout in Fall 2019. I teach statistics (BS and MS programs) and evaluation (MS program) in the psychology department. I *love* statistics! It is one way we can answer our research questions and test our hypotheses.

However, I know not everyone likes statistics. Some of you may not care much about them, and some of you may be scared about taking this course (especially in a pandemic and in an online asynchronous environment). Please know that **I am here for you and I want to make this class an enjoyable learning experience**. If there is anything I can do to help make this class more enjoyable and to help you learn, please reach out to me.

1.3 Navigating this website/book

This book was developed in R/Rstudio using bookdown and is hosted on a platform called GitHub. You can see the code for this book here.

There are some icons at the top of this book that you may find useful:

1. The first button of the toolbar toggles the visibility of the sidebar, which contains the table of contents. You can also hit the `S` key on your keyboard to toggle the sidebar.

2. The second button of the toolbar is the search button, which you can use to search the entire book. You can also hit the `F` (Find) key on your keyboard.

3. The third button is for font/theme settings, which you can use to change font size (smaller or bigger), font family (serif or sans serif), and theme (white, sepia, or night).
4. The fourth button provides information on the keyboard shortcuts.
5. On the right of the toolbar are icons to share on various social media platforms.

At some point when the textbook is more finalized, it will be turned into a PDF that can be downloaded and saved to your computer for use in the future. I will let you know when that happens!

1.3.1 Quiz questions

Throughout this textbook, there will be questions to help you test your knowledge. When you type in or select the correct answer, the dashed box will change color and become solid.

For example:

- What is 2+2?
- We attend the University of Wisconsin- Stout Madison Green Bay
- True or false: Dana thinks statistics is awesome. TRUE FALSE

1.3.2 Errors, mistakes, and suggestions

I am human, therefore I err. If you find an error in the textbook or something you think might be a mistake, please let me know ASAP so I can update this for everyone else. Let me know which section you find the error or mistake in and what the error or mistake is. For example, if there was an error here you could say, "There was an error in 1.2 that the first sentence should really be 'To err is human (Alexander Pope, 1711).'"

In addition, if you have ideas to help make this textbook even better, please let me know. I would love to make this a useful resource to you both during our course and in your future research. Help me in making that a reality!

Chapter 2

Statistics foundations

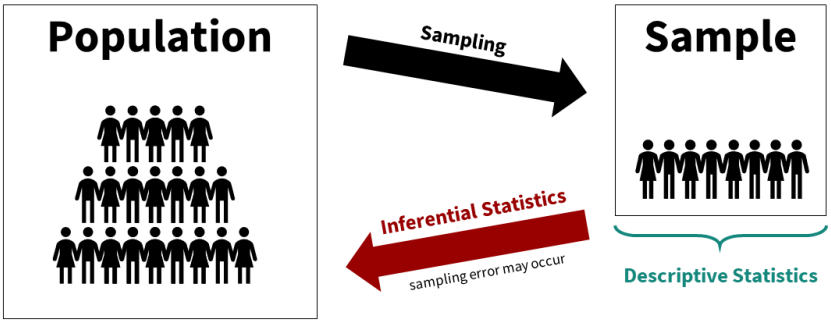
You have learned about both quantitative and qualitative methods. We will be focusing primarily on quantitative methods in this class and in this textbook. By quantitative methods, I mean methods that predominantly collect quantitative data that deals with numbers. We can then analyze that data using statistical procedures, which we will shorthand to "statistics." Understanding what we mean by statistics is the purpose of this chapter.

2.1 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) and measures of dispersion (e.g., range, standard deviation, variance). 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.



2.1.1 An example

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 Schitt's Creek–a very good show–versus watching video lessons on studying techniques–useful, but boring–improved test performance in UW-Stout students.

Our population of interest is therefore all UW-Stout students, roughly 9,500 students total. 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 the 9,500 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 my two 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 Schitt's Creek 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. 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 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 t-test, but we'll learn about that later) or–even better–we perform the statistic in our statistical software (jamovi). It spits out our statistical value and our p-value and we can then infer what the results mean for our population and answers our research question.[1]

2.2 Measures of central tendency and dispersion

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!):

- **Mean**: the sum of all points divided by the total number of points; susceptible to outliers
- **Median**: the middlemost value; less susceptible to outliers and best used when the data is skewed
- **Mode**: most frequent score
  - **Multimodal** or **bimodal**: when two or more values are the most frequent score

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
- **Kurtosis**: the weight of the tails relative to a normal distribution. There are some fancy terms related to kurtosis that you may hear about, but honestly I don't hear them used very frequently by researchers.
  - **Leptokurtic**: light tails; values are more concentrated around the mean

[1] You might be wondering: well, what were the results? Which group performed better? As much as I love Schitt's Creek, most students don't know how to study well, and so the students who watched the video lessons on studying techniques far outperformed the students who watched Schitt's Creek.

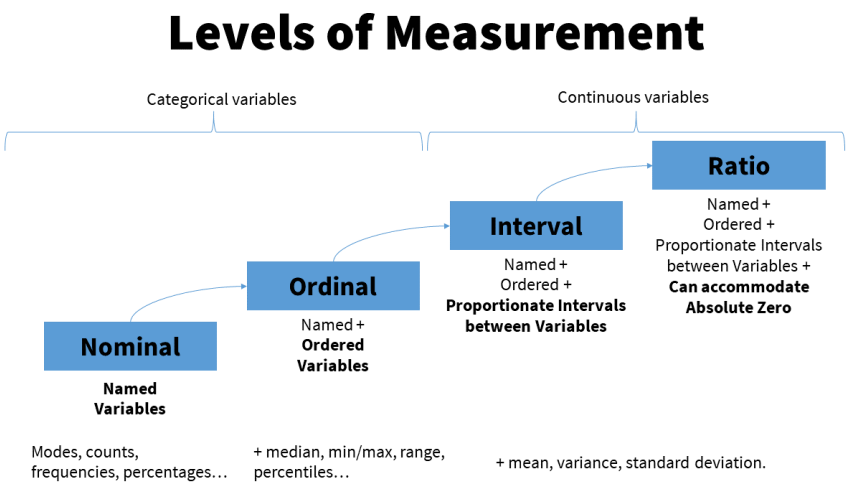
Interested in better techniques for studying? Check out The Learning Scientists. This articledoes a good job of summarizing the research on effective study practices.

  - **Platykurtic**: heavy tails; values are less concentrated around the mean

- **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
- **Quartile**: when a dataset is divided into four equal parts, each part is a quartile (Q1, Q2, Q3, and Q4)
  - **Interquartile range**: the middle 50% (Q1 to Q3)
- **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)
- **Skew**: in a non-normal distribution, it is when one tail of the distribution is longer than another. Present in asymmetric distributions
  - **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

2.3 Levels of measurement

This should be refresher material for you, but it is extremely important you are familiar with the four levels of measurement.



Categorical: variables that have *categories* to the levels, but cannot be analyzed with a mean because the levels are not proportionate. There are two types of categorical variables:

- **Nominal**: a categorical variable in which each level of the variable is named but there is no order to them (e.g., breeds of dogs)
  - **Binary**, **dummy-coded**, or **dichotomous**: a nominal variable with only two levels (general 0 or 1). This is a special type of nominal variable.
- **Ordinal**: a categorical variable in which each level of the variable is named and there is an order to them (e.g., ranks)

Continuous: variables with proportionate intervals between the levels meaning they can be analyzed with a mean, SD, variance. There are two types of continuous variables (although for the purpose of this course we will simply call them continuous variables):

- **Interval**: a continuous variable that has intervals that are directly proportionate (e.g., the distance between 2-3 is the same as the distance between 5-6)
- **Ratio**: a continuous variable like an internal variable but can accommodate an absolute zero, meaning a zero is actually possible (e.g., weight, temperature in Kelvin, reaction time)

2.3.1 Examples of levels of measurement

Confused still on the levels of measurement? Maybe this will help! Notice that studying can be measured at different levels. Depending on the nature of the question and response options, it might be nominal, ordinal, or continuous! Here's an example of data at the continuous, ordinal, and nominal level.

Name	Study_Continuous	Study_Ordinal	Study_Nominal
Name (Character)	*Hours studied per day*	*Likert scale of amount of studying*	*Whether or not they study every day*
Jesus	5.0	A great deal	Yes
Nicky	4.5	A great deal	Yes
Bradford	3.2	A moderate amount	Yes
Sylvia	1.7	A small amount	Yes
Martha	0.2	Rarely	Yes
Lillian	0.0	Never	No
Trayvon	0.0	Never	No

We can make any continuous variable into an ordinal and nominal variable and any ordinal variable into a nominal variable. But if we have a nominal variable

we cannot make it ordinal, nor can we make an ordinal variable continuous. In other words, continuous variables *contain more information*. Often, we want to avoid losing information and *always* keep the variable at the highest level of measurement. Continuous has more information than ordinal has more information than nominal.

Another way to put it: never do a median split and avoid "collapsing" categories when you can. You're losing information from your data by doing so.

2.4 Normal distribution

A very important distribution of data is known as the **normal distribution**. You may have also heard it called a bell-shaped curve. It has really important statistical properties which is why most of the inferential statistics we'll be learning in this class are *parametric statistics* that assume our data has a *normal distribution*.

Some of the important statistical properties of the normal distribution:

- Data are equally distributed on both sides of the mean.
- Skew and kurtosis are equal to 0, which is to say there is no skew or bad kurtosis.
- The mean is equal to the median, and both are the exact center of the distribution of data. In other words, if your mean and median are *not* the same, you know you have skewed data! In fact, if your median $<$ mean then you have positive skew and if your median $>$ mean then you have negative skew.
- We know the percentage of cases within 1, 2, 3, etc. standard deviations from the mean.

2.5 Key Terms

This chapter will cover some basic key terms you should recall from PSYC 190. These terms will come up repeatedly throughout the semester.

2.5.1 Study design terms

Some terms you should be familiar with:

- **Between-group/subject design**: different people are in each condition; participants are only exposed to a single condition
- **Correlational research**: a study in which causality cannot be claimed; correlation does not infer causation! It is, however, one of three necessary conditions to infer causality. It is a *necessary* but *insufficient alone* condition.

- **Cross-sectional research**: also called non-experimental research; the IV is not manipulated and there is no random assignment. Furthermore, data is only collected at one time point (as opposed to longitudinal research)
- **Experimental research**: the IV is manipulated and there is random assignment
- **Falsification**: A key way we separate science from pseudo-science is that we attempt to *falsify* our hypotheses as opposed to try *verify* our hypotheses. Null hypothesis significance testing (NHST) is about falsifying the null hypothesis; we can never truly verify our alternative hypothesis.
- **Hypothesis**: What we think the answer to our research question is (often our alternative hypothesis). The alternative and null hypotheses must be mutually exclusive (a result can't satisfy both) and exhaustive (all possible results are specified)
  - **Alternative hypothesis**: Often that the IV had **an** effect on the DV; can be specified as a two-tailed (an effect) or one-tailed (greater/less than) hypothesis
  - **Null hypothesis**: Often the *nill* hypothesis in that the IV had **no** effect on the DV
- **Qualitative methods**: Broadly, methods that focus on words and meaning (e.g., interviews)
- **Quantitative methods**: Broadly, methods that focus on numbers and statistics (e.g., Likert scales)
- **Quasi-experimental research**: the IV is manipulated but there is no random assignment
- **Randomization**: participants are randomly assigned to conditions
- **Repeated-measures design**: participants are repeatedly measured on the dependent variable, either across conditions or across time
- **Theory**: A description of a behavior that makes predictions about future behaviors
- **Variation**:
  - **Systematic**: researcher something systematically error into the study, especially into one condition over another. For example, by randomly assigning participants into one of two conditions, we are introducing systematic variability between participants. However, it could be unintentional systematic variation; for example, perhaps we have two researchers collecting data and one is mean and the other is nice, and so participants respond differently depending on which researcher collects data from them.
  - **Unsystematic**: random variation

- **Within-group/subject design**: the same person is in all conditions

2.5.2 Variables

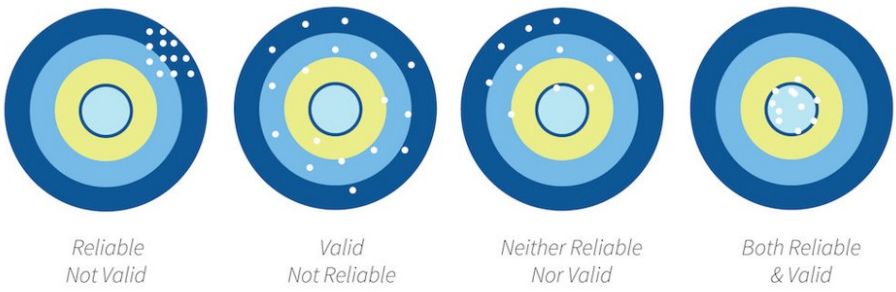
We tend to talk about two different types of variables in our studies:

1. **Independent variable** (IV; also known as the predictor variable): this is the variable that is thought to be the cause of some effect. In experimental research, it is the variable that is manipulated.
2. **Dependent variable** (DV; also known as the outcome variable): this is the variable that is thought to be affected by changes in the IV.

There are other types of variables we may be interested in:

- **Confounding variable**: a variable that affects or is related to both the independent and dependent variable
- **Covariate**: a variable that only affects or is only related to the dependent variable

2.5.3 Reliability and validity



- **Reliability**: the consistency of a measure by time (test-retest reliability), across items (internal consistency) or across different researchers (inter-rater reliability)
- **Validity**: the extent to which a test measures what it claims to measure
  - **Construct validity**: validity of inferences about the higher order constructs that represent sampling particulars. There are multiple types of construct validity; here are a few:
    - **Content validity**: experts using their judgment that something measures what it is supposed to measure
    - **Convergent validity**: correlations among two theoretically related constructs (or measurements) are strong and positive
    - **Divergent validity**: correlations among two theoretically not-related constructs (or measurements) are zero/null

    - **Criterion validity**: content on one test (predictor) correlates with performance on relevant criterion measures (outcome)
- **Statistical validity**: validity of inferences about the correlation between treatment and outcome
- **Internal validity**: validity about whether the observed relationship between A and B reflects a causal relationship between A and B
- **External validity**: validity of inferences about whether the cause-effect relationship holds over variation in persons, places, treatment variables, and measurement variables

2.5.4 Other terms

If other terms come up in the course of the semester that you believe should belong in this key term website, include it in your weekly reflection so I can update this page!

Chapter 3

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 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."
3. It's built on top of the R statistical language, meaning you can begin learning how to code (if you want). I do all of my statistical analyses using R in a different program called RStudio (actually this book was developed in RStudio and hosted on GitHub!). It's 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 (and I!) 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.

3.1 Getting started with jamovi

Throughout this course, you will be watching videos from the Introduction to jamovi LinkedIn Learning course by Barton Poulson, founder of datalab.cc. By the end of this course, you will receive a certificate indicating you watched all the videos that you can put on your LinkedIn profile.

For this chapter, you should first watch the first set of videos (Introduction

and chapters 1-3) on the LinkedIn Learning[1] course before proceeding with the reading.

3.2 Descriptive statistics

As a reminder, **descriptive statistics** are used to summarize, organize, and overall *describe* our sample data.

3.2.1 Data variables

First, 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!

[1] I have linked to the LinkedIn Course, but the links are specific to users in the University of Wisconsin-Stout. You can also find the videos to watch here: https://datalab.cc/jamovi

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. Sweet!

3.2.2 Exploring your data

In the third chapter of the LinkedIn Learning videos, you learned about data exploration, which are descriptive statistics. Exploring our data is 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 tab called `Descriptives`.

You first learned about various descriptive statistics. 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.

In small examples, we might write-up our descriptive statistics into a paragraph[2]:

[2]This comes from Wanzer (2017) Developmentally appropriate evaluations: How evaluation practices differ across age of participants

Youth TIG evaluators and evaluators with higher levels of child experience were less likely to involve tutors and tutees and more likely to observe tutees and tutors. Opposite from the expected findings, Youth TIG evaluators were less likely to involve tutees ($M = 2.57$, $SD = 1.12$) in the evaluation than Other TIG evaluators ($M = 2.96$, $SD = 1.04$), $t\ (334) = 3.19$, $p = .002$. Youth TIG evaluators were also less likely to involve tutors ($M = 2.91$, $SD = 1.09$) in the evaluation than Other TIG evaluators ($M = 3.27$, $SD = .96$), $t\ (334) = 3.13$, $p = .002$. Youth TIG evaluators were also less likely than Other TIG evaluators to involve tutees when designing protocols ($\chi^2\ ([1] = 4.20, p = .041$) and to involve tutors when interpreting results ($\chi^2\ [1] = 6.94, p = .008$).

In examples with many variables, we might write-up our descriptive statistics into a table[3]:

[3]This comes from Wanzer et al. (2020) Experiencing flow while viewing art: Development of the aesthetic experience questionnaire

Table 1
Demographic Information (N = 341)

Variable	*M* (*SD*) or *n* (%)
Age	35.08 (10.55)
Gender	
Female	184 (54.3%)
Male	153 (45.1%)
Other	2 (.6%)
Income	
Less than $25,000	73 (21.5%)
$25,000–$49,999	106 (31.3%)
$50,000–$74,999	70 (20.6%)
$75,000–$99,999	40 (11.8%)
$100,000–$124,999	27 (8.0%)
$125,000–$149,999	15 (4.4%)
$150,000 or more	8 (2.4%)
Education	
Some high school	0 (0%)
High school	31 (9.1%)
Some college	71 (20.9%)
Associate's degree	45 (13.3%)
Bachelor's degree	130 (38.3%)
Some graduate school	16 (4.7%)
Master's degree	33 (9.7%)
Doctoral degree	13 (3.8%)
Frequency of viewing art	
Multiple times per day	55 (17.4%)
Once per day	18 (5.7%)
Multiple times per week	65 (20.6%)
Once per week	40 (12.7%)
Once per month	42 (13.3%)
A few times per year	87 (27.5%)
Once per year	9 (2.8%)
Have you ever worked in a visual-art-related job?	
Yes	66 (19.5%)
No	273 (80.5%)
How much training in visual art have you received? (1 = *no training*, 7 = *a great deal of training*)	2.61 (1.67)
Did you receive at least one of your degrees or a minor in a visual-art-related field?	
Yes	24 (7.8%)
No	284 (92.2%)

3.2.3 Visualizing your 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.

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. For the latter, I will provide some guidance on best practices for visualizing inferential results in the relevant subsequent chapters. For now, let's go over the various `Plots` in the `Descriptives` analysis.

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.

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

Remember: it is incredibly important to always visualize your data! You never know what descriptive statistics may be hiding.

Here's a video walking through why this is so important:

Lastly, there is an option under **Bar Plots** for (you guessed it) a `Bar plot`. It will add error bars to your bar plot, but now that you know what can be hiding under descriptive statistics (bar plots only show the mean and error) you will hopefully avoid using these in the future when you can show the actual data itself.

3.2.3.1 Expanding your data visualization

Although these can be useful plots, I often do most of my data visualizations in other platforms. For most of my work, I use Excel because I find it pretty easy to make beautiful graphs. Here's an example of a visualization I made in Excel[4]:

[4]This comes from Wanzer et al. (2020) Promoting intentions to persist in computing: An examination of six years of the EarSketch program

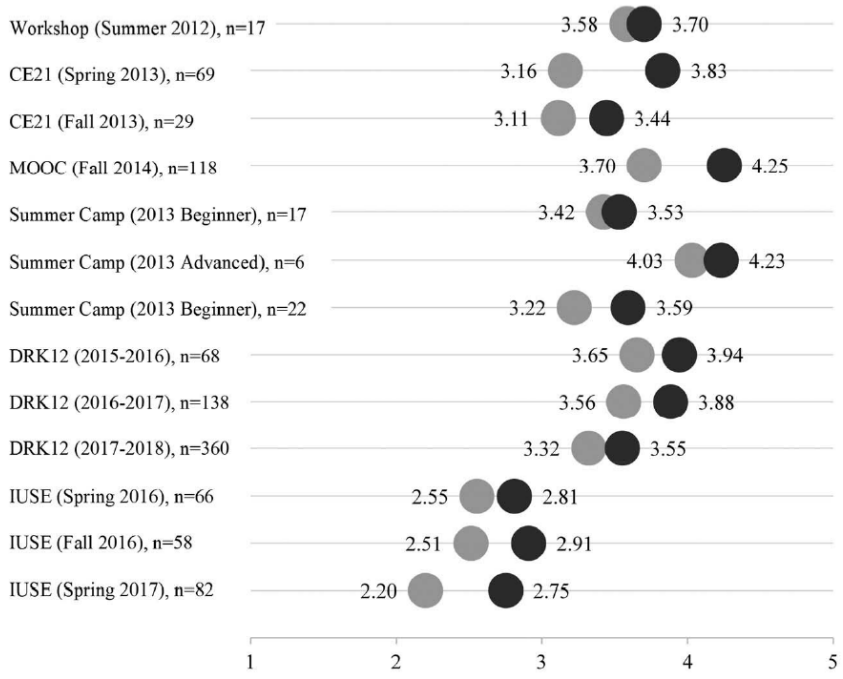


Figure 2. Average intent to persist across all EarSketch studies.

For some more complicated figures, I turn to the `ggplot2` package in R. Here's an example of a visualization I made in R[5]:

[5]This comes from Wanzer (2020) What is evaluation? Perspectives of how evaluation differs (or not) from research

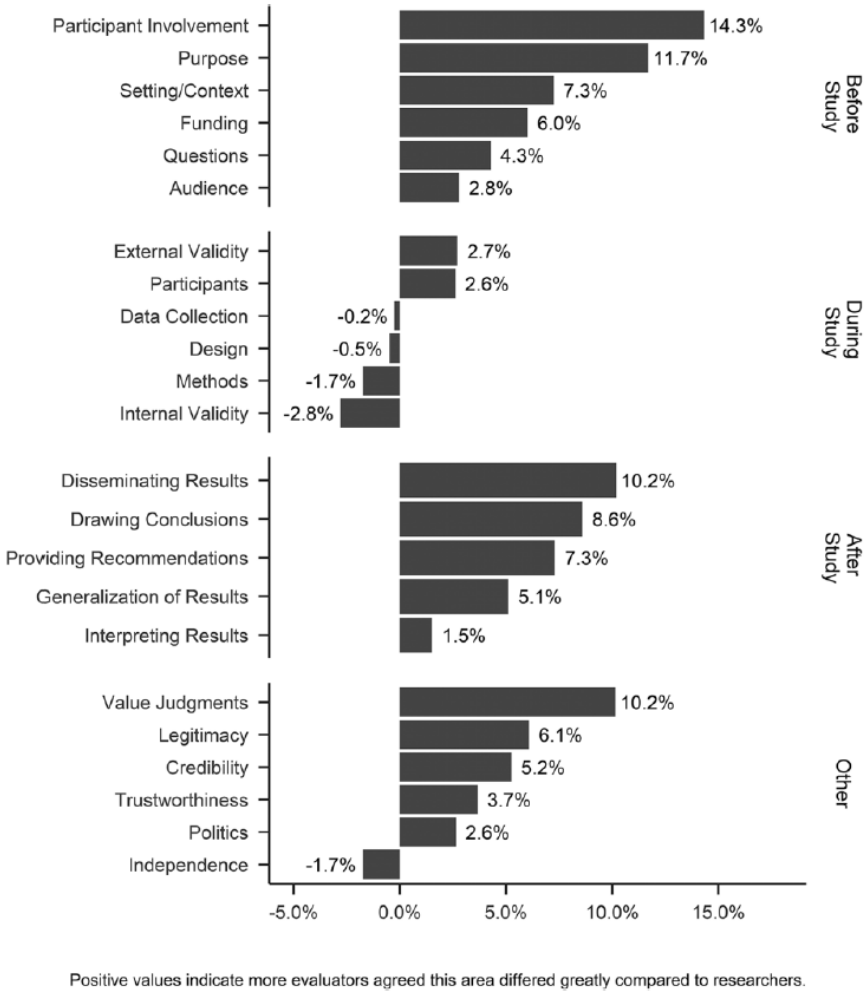


Figure 3. Percentage differences between researchers and evaluators stating the category differed greatly. *Note.* Positive values indicate more evaluators agreed this area differed greatly compared to researchers.

In this class, we'll learn about effective data visualization later in the semester. I will provide some brief tutorials to help you get started, but please note that learning data visualization should be a course unto itself, so we will not be able to cover everything.

3.3 Cleaning data

There are four basic types of cleaning we will be learning about: checking your data is setup correctly, computing new variables, transforming variables, and using filters.

3.3.1 Data setup

As previously mentioned, 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's the things you should be doing for all variables:

1. Make sure the variable name is meaningful to you. You may also want to change it to something that will appear nicely in your data visualizations or tables (e.g., don't write `Q35` but rather `BDI Score`).
2. Add a description to your variable so you have more context. Maybe you write `Average score of all BDI items` for the description of your `BDI Score` 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.

3.3.2 Compute

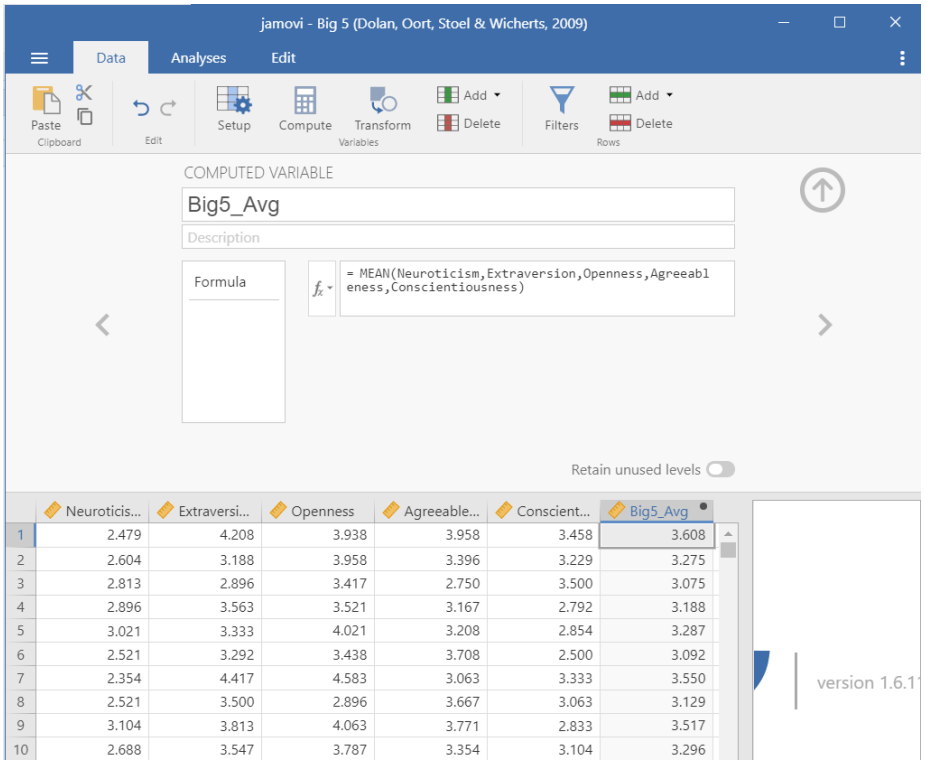
Sometimes you need to create new variables from your raw (meaning uncleaned) data. Perhaps you collected data on a scale that has five items. Normally, we create an average score of all the five items and that new *computed* average score is what we use in our analyses.

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. We would click on the Data tab and choose `Compute`. We would rename the computed variable (e.g., `Big5_Avg`), add in a description, and then create the formula.

In this case, we need to select the function `MEAN`. Below the function, 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.



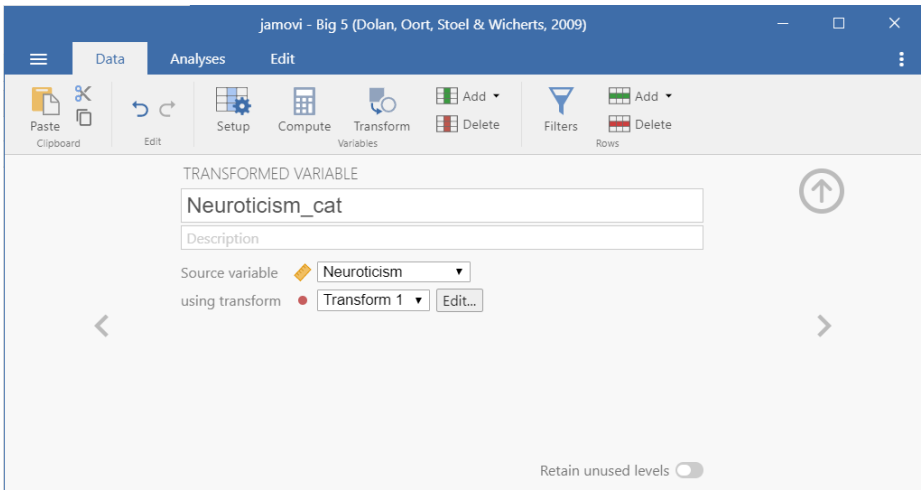
If you'd like to learn more about computed variables in jamovi, check out this jamovi blog post on the topic.

3.3.3 Transform

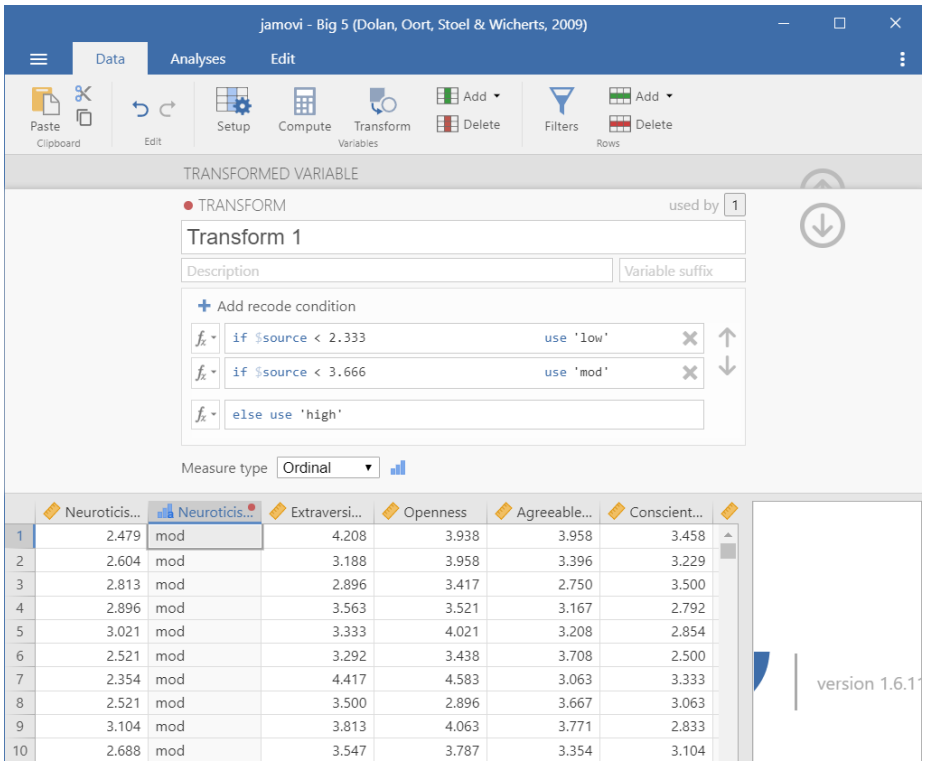
Sometimes we want to take an existing variable and transform it in some way or we want to do a computation across multiple variables (e.g., reverse-score multiple items in a dataset). If you want to learn more about transforming variables, the jamovi blog has a great blog post on the topic.

3.3.3.1 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.



3.3.3.2 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.

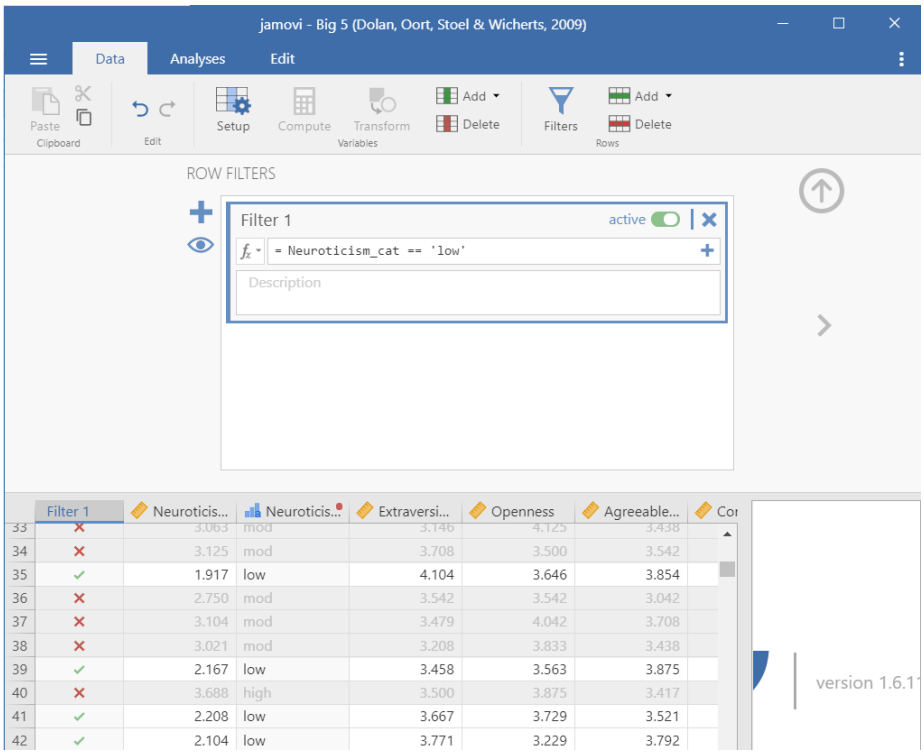
We can click a new variable (e.g., Openness), select Transform, rename the variable, and select the `Low_mod_high` transformation we already used. Voila! The work we did previously can easily be used again in this analysis.

3.3.4 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.

3.3.4.1 Row filters

Maybe we only want to analyze data from people who are low in neuroticism. We would create the following filter:



You'll notice in the dataset it will add 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: `<`, `>`, `<=`, `>=`

3.3.4.2 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).

Chapter 4

Hypothesis testing

Now that we've covered descriptive statistics and are familiar with our statistical software, it's time to turn to inferential statistics. Remember, we conduct inferential statistics because we often cannot collect data from an entire population. Therefore, we collect a sample to draw inferences about the population of interest.

One of the ways we make inferences is using hypothesis testing. Regardless of the inferential statistic we are performing, hypothesis testing goes through the same procedures:

1. Write your hypothesis (aka the alternative hypothesis) and its accompanying null hypothesis.
2. Set the criteria for a decision of whether to support or reject the null hypothesis
3. Perform the test statistic
4. Interpret your results and make a decision

Let's go through each of these in turn, using a hypothetical example.

4.1 An example of hypothesis testing

Imagine a researcher wants to replicate Alburt Bandura's famous Bobo doll experiment. In this study, the researcher randomly assigns 30 6-year-old children to one of two conditions: one group watches a video of an adult showing aggressive behavior toward a Bobo doll and the other group watches a video of an adult passively playing with a Bobo doll. After watching their assigned video, children then went to the same room from the videos with the same Bobo doll. Researchers observed for aggressive behaviors[1].

[1]Note that the study design was actually much more impressive than what I'm describing. They accounted for the children's baseline aggression and the gender of both the child and

4.1.1 1. Write your hypotheses

The first step is to write out our hypotheses. We need to write our **alternative** and **null** hypotheses.

The alternative hypothesis is typically what we expect the results of the study to be. We often expect to see *something*; that there *is* an effect. We usually write this out as H_1.

The null hypothesis is typically what we *don't* expect the results of the study to be. It is often that there was *no* effect of the study. We usually write this out as H_0.

The two hypotheses–our alternative and null hypotheses–must be **mutually exclusive** and **exhaustive**. Mutually exclusive means a potential result of the study cannot support both the alternative and null hypothesis; it must exclusively support only one. Exhaustive means the entire possible universe of results must be captured in our two hypotheses; it must exhaust all possible results.

We might also have **directional** or **non-directional** hypotheses. Directional hypotheses are also called one-tailed hypotheses because only one tail of the distribution would lead us to fail to reject the null hypothesis. Non-direction hypotheses are such that we don't know whether the difference will be greater or less than 0, but we just think there will be a difference; these are also called two-tailed hypotheses because both tails of the distribution would lead us to fail to reject the null hypothesis. This will make a little more sense below and a lot more sense in the next chapter.

What might the hypotheses be for our example study? There should be theory and research to support alternative hypotheses. There is ample research now that viewing aggression leads to aggression through imitation and observed learning. Therefore, the researcher likely has a hypothesis that watching the aggressive adult will lead to greater aggression than watching the passive adult.

Therefore, our hypotheses would be:

H_1: Children watching the video with the adult aggressively playing with the Bobo doll will exhibit *more* aggressive behaviors than children watching the video with the adult playing passively.

H_0: There will be *no difference* in children's aggressive behaviors between the two groups or children watching the video with the adult aggressively playing with the Bobo doll will exhibit *fewer*aggressive behaviors than children watching the video with the adult playing passively.

Note that we now satisfy mutual exclusivity (no possible overlap in the hypotheses) and exhaustiveness (all possible results are captured).

---

the person in the video. If you are interested, you can read more here: https://www.simplypsychology.org/bobo-doll.html

A common error in a directional hypothesis like this is to forget that the null hypothesis is both no difference *and* the opposite. In other words, we have three possible options for our null and alternative hypotheses based on direction (μ is the Greek letter "mu" and we often use it to signify the mean):

	Two-tailed	One-tailed (greater)	One-tailed (less than)
Alternative (H_1)	μ_1 != μ_2	$\mu_1 > \mu_2$	$\mu_1 < \mu_2$
Null (H_0)	μ_1 == μ_2	μ_1 <= μ_2	μ_1 >= μ_2

Since we're talking about mean differences, we could also reformulate the above table slightly differently:

	Two-tailed	One-tailed (greater)	One-tailed (less than)
Alternative (H_1)	μ_{diff} != 0	$\mu_{diff} > 0$	$\mu_{diff} < 0$
Null (H_0)	μ_{diff} == 0	μ_{diff} <= 0	μ_{diff} >= 0

4.1.2 2. Set the criteria for a decision

Our hypotheses simply state "more aggressive" or "no difference/less aggressive." What constitutes *more*? What constitutes *no difference*? We have to specify that.

No difference seems easy. That's a difference of zero, right? Well, not exactly, because it's highly unlikely we would get an *exact* difference of zero. Therefore the question is: which values are close enough to a difference of zero that we'd still say that there is no difference? If our values are within that range, then we would fail to reject the null hypothesis. If our values are *outside* that range, then we would reject the null hypothesis.[2]

Let's try to visualize this. We are saying that the null hypothesis is there is no difference (or less aggression), but at some point no difference turns into *greater* difference. Furthermore, we have a directional hypothesis in that we do not think the difference will be negative, that children watching the adult play aggressively will exhibit fewer aggressive behaviors. Basically, we need to know what the critical value is in the figure below.

```
## Loading required package: viridisLite
```

We figure out that critical value based on what we set as our level of significance, also known as the **alpha level**. Most studies you read use the arbitrary $\alpha = .05$

[2]Note my language carefully here: fail to reject the null hypothesis OR reject the null hypothesis. Note how I am *not* saying support the alternative hypothesis! Through null hypothesis significance testing, we are only ever testing the null hypothesis and therefore can only make conclusions about it. This is why we need replication studies to provide ample support for alternative hypotheses.

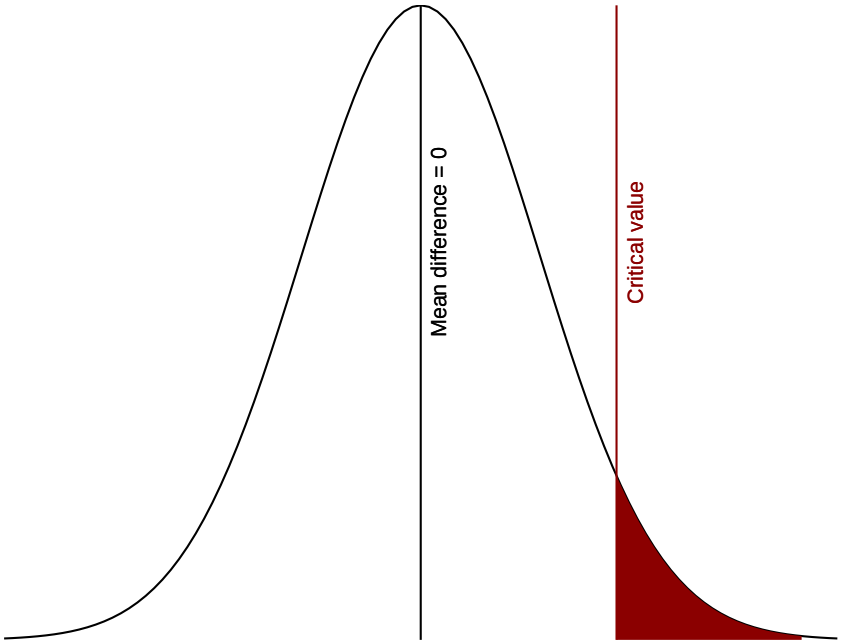


Figure 4.1: Critical area of statistical significance

(5%), although we really should be thinking critically about what alpha level we use (more on that in the next chapter). 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 alpha is the level of which we are saying would be considered "surprising" versus "not surprising." If we got a mean difference that fell in that red area, then we would consider that "surprising" *if we believed the null hypothesis was true.* Basically, if we assume there is a mean difference of 0 (i.e., the null hypothesis), values past 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.*

Now that we understand that a bit better, how do we find out our critical region? We do so based on our understanding of the incredible properties of the normal distribution! Back in the day before computers, some fancy mathematicians and statisticians figured out the exact t-values based on things like the direction of our hypothesis, our alpha, and our degrees of freedom. Let's figure these out for our example:

1. **Direction of our hypothesis**: we have already determined we're using

a one-tailed hypothesis.
2. **Alpha**: let's just go with $\alpha = .05$ for now
3. **Degrees of freedom**: This is calculated by N - 2. We have 30 children total, so 30 - 2 is 28.

We then go to a *t*-value table like this one and find the cell we are looking for to identify our critical t-value (t_{crit}). First, our one-tailed probability is .05 so we're going to look under the sixth column ($t_{.95}$, one-tail = .05, two-tails = .10). Then we need to find the row for our degrees of freedom (df = 28). That leads us to a t_{crit} of **1.701**.

Therefore, we can now finalize this step. Our criteria for decision is as such:

- $t_{obt} > t_{crit}$ means we reject the null hypothesis.
- $t_{obt} < t_{crit}$ means we fail to reject the null hypothesis

4.1.3 3. Perform the test statistic

We haven't learned how to calculate the test statistic yet, but no worries we will get there soon. We had 30 participants, 15 in each condition. The researcher performed the experiment and got the following results:

- Children who watched the video of the adult playing aggressively with the Bobo doll displayed an average of 51.1 aggressive behaviors ($SD = 3.5$).
- Children who watched the video of the adult playing passively with the Bobo doll displayed an average of 27.4 aggressive behaviors ($SD = 3.3$).

That means the mean difference is 51.1 - 27.4 = 23.7. We'll learn how to conduct a t-test later, but for now you can just input the numbers into this calculator. It nicely gives you a lot of the values, but the one we are looking for is the test statistic, which is $t_{obt} = 19.08$ (notice we round to two decimals). It also gives us our p-values based on the *null* hypothesis, which in our case is that population 1 < population 2. The p-value is < .00001 but we never go to so many decimals, so we would say p < .001. The probability of getting a *t*-value as large as we did it less than .1% (less than our alpha of .05, so it is statistically significant). Very surprising!

4.1.4 4. Interpret results and draw a conclusion.

Remember that when $t_{obt} > t_{crit}$ we reject the null hypothesis whereas when $t_{obt} > t_{crit}$ we fail to reject the null hypothesis. Since 19.08 > 1.701, we *reject the null hypothesis that there is no difference in conditions or that children in the passive condition displayed more aggression than children in the aggressive condition.*

However, this is when **Type 1** and **Type 2** errors come into play. 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!

I like to use this table 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
$\boldsymbol{p < .05}$ (statistically significant)	Type 1 error	Correct interpretation
$\boldsymbol{p > .05}$ (statistically non-significant)	Correct interpretation	Type 2 error

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 results ($p > .05$) then *either* we made a correct interpretation *or we made a Type 2 error*!

A common mistake is assuming that $p < .05$ means that the alternative hypothesis is true. This is inaccurate because the *p*-value is the probability of our data given the null hypothesis is true. It says nothing about the alternative hypothesis. Similarly, a common mistake is assuming $p > .05$ means the alternative hypothesis is false. This is incorrect for the same exact reason.

Next week we'll learn a lot more about p-values, errors, and more. For now, tuck this piece of information into your brain to remember!

4.1.5 Final note

When you read journal articles, you'll note that they rarely discuss the null or alternative hypothesis. They may explain their research questions or their hypotheses (these hypotheses are their alternative hypotheses), but they rarely discuss the null.

This is not necessarily a *bad* thing. Rather, what may be problematic with it is if researchers apply NHST without critically thinking about what their null hypothesis is or whether they have a one-sided hypothesis, which leads researchers to use defaults when the defaults may not be most appropriate.

Chapter 5

BEAN

What a random chapter title right? Yes, but it's also an important acronym in hypothesis testing. BEAN stands for:

1. **B**eta (for power, which is technically 1 - β)
2. **E**ffect size
3. **A**lpha
4. **N** (sample size)

The following sections will go through each of these in turn before ending with discussion on how all of these things interrelate. Although the BEAN acronym is useful (you'll find out why in a later section), I won't discuss them in that particular order.

5.1 Effect sizes

An **effect size** is a quantitative description of the strength of a phenomenon. There are two basic effect sizes we tend to talk about:

The ***d*** family of effect sizes are standardized mean differences. They start at 0 (no mean difference) and can go up to infinity, with larger values meaning larger standardized mean differences. Some of the effect sizes in this family:

- Cohen's *d* is perhaps the most popular standardized mean difference effect size. Generally, the equation is the mean difference divided by the pooled standard deviation, but in reality the equation differs based on a variety of scenarios and whether you are using a one-sample, independent samples, or paired samples *t*-test.
- Hedge's *g* is a less biased version of Cohen's *d*. Cohen's *d* is particularly problematic for small sample sizes, so Hedge's *g* is generally preferred, but you'll see that not all statistical programs provide this effect size. It's not

that difficult to calculate Hedge's g based on Cohen's d, but just keep this information in mind.

The ***r*** family of effect sizes are measures of strength of association. As you'll read about in the correlation and regression chapters, this family of effect sizes can describe the proportion of variance explained (e.g., r = .8 is 64% variance explained, which is r-squared). Some of the effect sizes in this family:

- r is a correlation. It's a standardized measure of the strength of association where r = -1 or +1 means a perfect relationship and $r = 0$ is no relationship at all.
- 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_p^2 (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.
- ω^2 (omega-squared) is calculated even more differently and is considered the least biased estimate. There is also ω_p^2 and ω_G^2 (generalized omega-squared) but we won't get into that.

If you nerded out over this information, check out this great journal article by Daniel Lakens.

5.1.1 Small, medium, and large effect sizes

What is considered a small, medium, and large effect size? Quite frankly, *it depends.*

You may have seen some heuristics online about what small, medium, and large is for Cohen's d (e.g., .2, .5, and .8) and r (e.g., .1, .3, and .5) but these should not be just used willy nilly without critical thought. In fact, Cohen (who is regularly cited for these heuristics) said that the way we should determine cut-offs is based on looking across studies to find what is considered small, medium, and large *in that particular context.*

Lakens (who also did the great journal article on effect sizes above) has a fantastic new preprint out on Sample Size Justification. In it, he provides an overview of six possible ways to determine which effect sizes are interesting:

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, 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)? 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. This is important for when you are then determining what sample size you need, which will be discussed in a separate section.

5.2 Alpha & p-values

Whereas effect sizes tell us about *practical* significance, they do not tell us about *statistical* significance. That is what p-values are for: they tell us whether our results are *statistically* significant or how surprising they are. The formal definition of a ***p*-value** is that it is the probability of observing data that is as extreme or more extreme than the data you have observed, assuming the null hypothesis is true. There's two things to keep in mind about this definition:

1. The p-value is about the probability of our data. It is not about the probability of our hypothesis.
2. The p-value is based on the assumption that the null hypothesis is true. In null hypothesis significance testing, we are only ever testing against the null. We can never "accept" the alternative hypothesis but rather fail to reject the null. If we fail to reject the null enough times (and rarely reject the null) then it gives weighted evidence towards our alternative hypothesis, but we never prove the alternative hypothesis is true.

Another way we could think of the p-value is: assuming there is no difference (i.e., the null hypothesis is true), how surprising is our data?

You may have heard of p-values before. You may have heard about them in a previous statistics course or you may have heard of them in relation to the "replication crisis." You may have heard about the journal that banned p-values altogether. You may then be wondering why we are learning about p-values if they seem so problematic that they should be banned.

The biggest problem with p-values is that they are misunderstood, even by researchers. They are often misinterpreted. Daniel Lakens has a great blog post on the topic and a great MOOC about improving your statistical inferences. You

did one of the MOOC assignments last week in the "Understanding common misconceptions about *p*-values" assignment. Whereas that assignment focused on understanding misconceptions, I will spend this chapter talking about what *p*-values *are*.

5.2.1 Alpha

The *p*-value is the probability of the "surprisingness" of your data. We've already seen this chart from the last chapter where the area in red is our alpha region. The alpha level is simply the level at which we consider the data *so surprising* that we reject the null hypothesis. This area in red is also considered our critical test region and the region of statistical significance. Statistical significance depends on what we set our alpha at.

```
## Loading required package: viridisLite
```

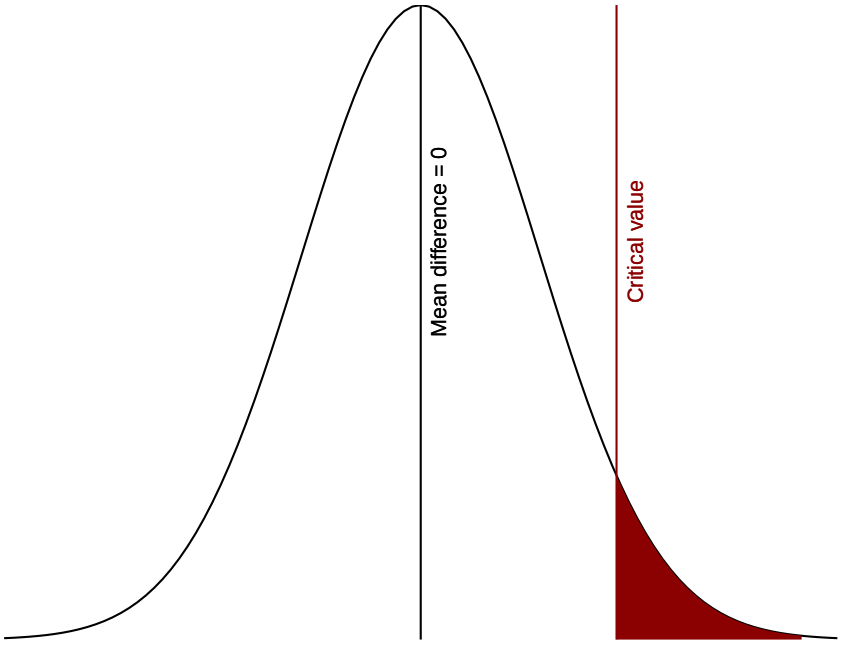


Figure 5.1: Critical area of statistical significance

Why is alpha set at 5% usually? It comes from Fisher (1925), who said something that eventually grew to tradition:

> A deviation exceeding the standard deviation occurs about once in three trials. Twice the standard deviation is exceeded only about once in 22 trials, thrice the standard deviation only once in 370 trials.... The value for which P = .05, or 1 in 20, is 1.96 or nearly 2 ; it is convenient to take this point as a limit in judging whether

> a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant.

Basically, .05 was convenient. It was 1/20. It was around 2 standard deviations from the mean in a normal distribution. For some reason, it caught on (maybe the "formally regarded as significant" was why).

However, a year later, even Fisher acknowledged we shouldn't just arbitrarily use $p = .05$ as our alpha level.[1] Rather, we should consider setting it at higher odds (e.g., $p = .01$). He also argued, "A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance." (Fisher, 1926, p. 504).

In other words, we need to think critically about the alpha level we set *and* we need to test an effect multiple times before we start thinking the alternative hypothesis is true. Let's discuss power before we start putting it all together.

5.2.2 Are p-values bad?

Some have argued that we should abandon the *p*-value; this has led to things like journals completely banning *p*-values altogether. However, I agree with Lakens that "the practical alternative to the *p*-value is the correctly used *p*-value." That's to say: there is nothing *wrong* with the *p*-value inherently, and it can be useful. Rather, what's *wrong* is that many people use them incorrectly.

5.3 Power

Power is the probability that you will observe a significance effect if there is a true effect. In other words, power is the probability of a statistically significant result assuming the alternative hypothesis is true. Power can range from 0-100%, but typically people suggest setting it at 80%. However, in practice, power is often far lower than 80%, something we'll investigate in the final section of this chapter and in the homework.

Let's compare this to our definition of the p-value: the probability of observing data that is as extreme or more extreme than the data you have observed, assuming the null hypothesis is true.

Power is based on the assumption that the alternative hypothesis is true whereas the *p*-value is based on the assumption that the null hypothesis is true. If we want to increase the likelihood of supporting our alternative hypothesis, then we should be doing all we can to increase our power!

Let's start putting all this together.

[1] If you want to read more, this is a short read on the history of the .05 alpha level: https://www2.psych.ubc.ca/~schaller/528Readings/CowlesDavis1982.pdf

5.3.1 Alpha, power, and error rates

Here's all our definitions so far (remember, we can never truly know whether the null or alternative hypothesis is true):

1. Alpha is the value we set at for what constitutes a statistically significant result, assuming the null hypothesis is true.

2. Power is the value we set at for what constitutes a statistically significant result, assuming the alternative hypothesis is true.

3. A type I error is when we get a statistically significant result but in fact the null hypothesis is true.

4. A type II error is when we do not get a statistically significant result but in fact the alternative hypothesis is true.

5. A correct inference is when we either

    1. get a statistically significant result when the alternative hypothesis is true *or*

    2. when we do not get a statistically significant result when the null hypothesis is true.

In the following table, determine where each of the pieces should go. Note that we have six things to populate but only four cells: each cell must contain at least one of the six things. Think critically here before testing your answers!

	H_0 is true	H_1 is true
$p <$ **.05** (statistically significant)	A	B
$p >$ **.05** (statistically non-significant)	C	D

Which cell should each of the following items go?

1. alpha A B C D
2. power A B C D
3. type I error A B C D
4. Type II error A B C D
5. Correct inference #1 A B C D
6. Correct inference #2 A B C D

5.3.2 Playing with alpha & power

Let's play around with some numbers and see what happens!

5.3.2.1 Assuming the null hypothesis is 100% true

Assuming the null hypothesis is 100% true, we could fill in the table with actual numbers. Let's also use the arbitrary values we often set alpha and power at: alpha = 5% and power = 80%. Here's the resulting table:

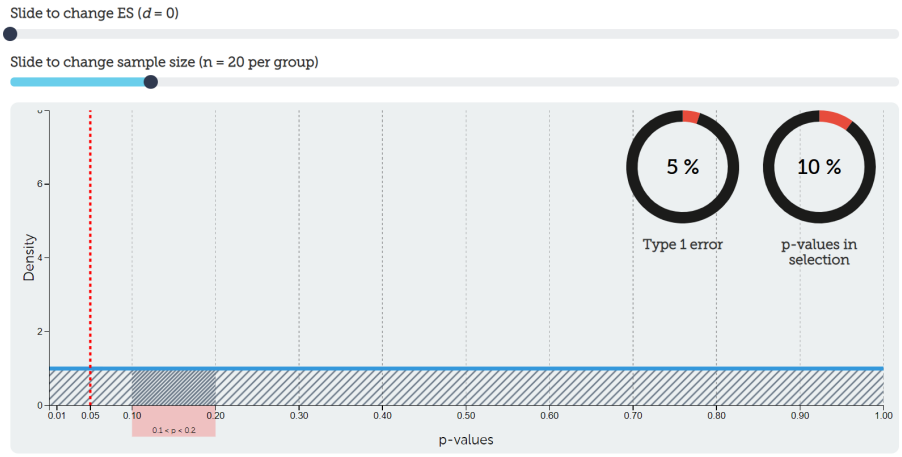
	H_0 is true	H_1 is true
$\boldsymbol{p < .05}$ (statistically significant)	5%	0%
$\boldsymbol{p > .05}$ (statistically non-significant)	95%	0%

How did I get there? First, we're assuming the null hypothesis is 100% true. Therefore, that column must add up to 100%. If the null hypothesis is 100%–and we know our hypotheses must be mutually exclusive–then the alternative hypothesis must be 0% true. Therefore, that column must add up to 0%. The whole table must equal to 100% to exhaust all options.

Therefore, our power doesn't matter at all in this case. If the null is true, then it doesn't matter what power we have to detect the alternative effect *because the alternative effect does not exist*. So we instead use alpha and put it in the upper left cell. Note that our alpha level is the Type I error rate we are setting!

If the whole table must equal to 100%, and the left column must equal to 100% because the null is 100% true, then 100-5 = 95% for the correct inference. In other words, if we tested this effect (that doesn't exist) 100 times, around 95% of the time we would get a non-significant p-value ($p > .05$) and about 5% of the time we would get a significant p-value and be making a Type I error.

We can visualize our *p*-value distribution using this handy interactive calculator. We set our effect size (ES) to be $d = 0$, meaning there is no effect (i.e., the null hypothesis is true). This results in a uniform distribution of p-values. Exactly 5% of p-values would fall between $p = 0$ and $p = .05$ (the shaded region to the right of the red dotted line). That aligns with our Type I error rate as well (5%). Go ahead and play around with the interactive calculator and try moving the slider for sample size! Notice that it does absolutely nothing. We'll understand why when we put everything together.



5.3.2.2 Assuming the alternative hypothesis is 100% true

Let's try out the opposite: assume the alternative hypothesis is 100% true, alpha is 5%, and power is 80%. What would you put in the table?

	H_0 is true	H_1 is true
$p <$ **.05** (statistically significant)	0%	80%
$p >$ **.05** (statistically non-significant)	0%	20%

How did I get those numbers? First, remember that the table must equal to 100% (hypotheses must be exhaustive). Second, remember that the alternative hypothesis is 100% true so that column must equal to 100% (and because hypotheses must be mutually exclusive, the other column must equal to 0%).

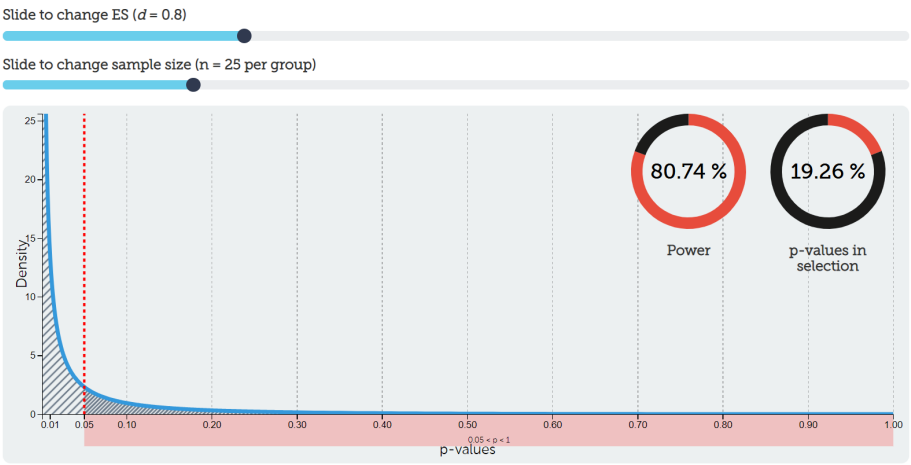
Therefore, it doesn't matter what we set alpha to. We cannot get a Type I error if the alternative hypothesis is true! We can only get a correct inference or make a Type II error.

We set power to 80%, and power is the probability of getting a statistically significant result assuming the alternative hypothesis is true. Therefore it goes in the top right cell. Notice that power is the probability of correctly detecting a statistically significant effect!

With simple arithmetic, 100-80 = 20% is our Type II error. If we were to test for this effect 100 times, about 80 times we would correctly detect the effect and about 20 times we would fail to detect the effect.

Let's visualize this. Go back to our handy interactive calculator and put $d = .8$ as our effect size and $n = 25$ per group to our sample size. Notice now what our distribution of p-values looks like! Rather than a uniform distribution, now we have a steep exponential distribution. I have chosen to highlight all p-values in

the range of $p > .05$, which in that selection is roughly 20% (our Type II error rate from above). The number of p-values $< .05$ is roughly 80%. Play again with the ES slider and sample size slider. Notice now that it makes a difference to our power! You're getting a glimpse into how power depends on our alpha, effect size, and sample size. BEAN!



5.3.2.3 Assuming a 50/50 split on the null and alternative hypotheses

In reality, we never truly know whether the null or alternative hypotheses are true. Maybe we're testing a new effect and we are completely 50/50 of whether the null or alternative hypothesis is true. Let's keep our alpha and power the same (5% and 80%, respectively) and fill out our table now:

	H_0 is true	H_1 is true
$\boldsymbol{p < .05}$ (statistically significant)	2.5%	40%
$\boldsymbol{p > .05}$ (statistically non-significant)	47.5%	10%

How did we get there? Again: the table must equal to 100% and we specified ahead of time that we thought it was about 50% true for each of the hypotheses, so each column must equal to 50%. 50% of 5% is 2.5% and 50% of 80% is 40%. We then fill out the bottom row based on arithmetic.

Imagine this were your study and you got a significant p-value. What could you conclude? Either you reject the null hypothesis or fail to reject the null hypothesis. But which one? In reality, we never know, but there are things we can do to increase the likelihood that our statistically significant result is because the alternative hypothesis is true and not the null hypothesis.

Right now, based on the values we have set (alpha = 5% and power = 80%), it is *16 times more likely* that the alternative hypothesis is true than the null hypothesis is true (40/2.5 = 16). You might be fine with that, but what can we do to increase this likelihood?

5.3.2.4 Increasing power

Let's try it again, but this time let's increase our power to 95% and keep our alpha at 5% (50/50 on the hypotheses). Fill out the table!

	H_0 is true	H_1 is true
$p <$ **.05** (statistically significant)	2.5%	47.5%
$p >$ **.05** (statistically non-significant)	47.5%	2.5%

Now it is *19 times more likely* (47.5/2.5 = 19) that the alternative hypothesis is true than the null hypothesis is true. Awesome! We have now discovered that increasing power increases the likelihood our alternative hypothesis is true.

5.3.2.5 Decreasing alpha

Let's do another example in which we still have a 50/50 on the hypotheses but we reduce our alpha to 1% and keep our power at 80%. Fill out the table!

	H_0 is true	H_1 is true
$p <$ **.05** (statistically significant)	.5%	40%
$p >$ **.05** (statistically non-significant)	49.5%	20%

Now it is *80 times more likely* (40/.5 = 80) that the alternative hypothesis is true than the null hypothesis is true. Awesome! We have now discovered that decreasing alpha increases the likelihood our alternative hypothesis is true.

5.3.2.6 Your turn: Increasing power AND decreasing alpha

Try both of them. Fill out the table on your own. Then fill in the blank:

When we assume the null and alternative hypotheses are 50% likely each, and we set our alpha to 1% and our power to 95%, it is times more likely that the alternative hypothesis is true than the null hypothesis is true.

5.4 Sample size

Sample size is the total number of participants in a study. In a between-subjects study, we often describe how many participants are in each group; although it

is best if there are equal numbers in each group, there are times when that may not be the case.

Often, the biggest question we want to know is: what sample size do I need for my study? Daniel Lakens has a great new preprint out on the topic. We often cannot measure the entire population, but some other ways we can determine the sample size are:

1. Resource constraints: sometimes time and budget limits our sample size
2. Accuracy or an *a priori* power analysis: based on the statistical power we hope to achieve (which is in turn based on the effect size we expect)
3. Heuristics: some prespecified rule or norm that is described in the literature (to be avoided as much as possible)

The first option is more of a research methods discussion and will not be discussed here. The third option is to be avoided as much as possible. Therefore, that leads us to the second option, which is to conduct an *a priori* power analysis.

5.4.1 BEAN: Power analysis

We previously saw how alpha and power relate to one another. In the interactive calculator you also started to discover that effect sizes and sample size also relate to alpha and power. This is the power of the BEAN: if you know three out of the four of BEAN, you can determine the fourth. Power, effect sizes, alpha, and sample sizes all interrelate!

Typically, there are three things we may 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?

There is software out there to help you conduct power analyses. The most popular is G*Power.

For our purposes, we're going to simplify things and use the jpower module in jamovi. This can calculate power for an independent samples t-test, a paired samples t-test, and a one-sample t-test. Our previous example in the last chapter (the Bobo doll experiment) has two groups in a between-subjects design. Next chapter you'll learn how to determine what statistical test you would perform, but for now I will just tell you that we would conduct an independent samples t-test with this experiment's data.

In the jpower module, choose your statistical test in the drop-down menu; in this case, let's choose independent samples t-test. Next, you specify what you

want to calculate: your N per group (sample size), power, or effect size. It will grey out that box in the three boxes below. Let's discuss them in turn:

1. Minimally-interesting effect size: it shows the lower case Greek letter delta here, but we can essentially think of it as a Cohen's d value. Go back to the effect size section for help in determining your smallest effect size of interest.
2. Minimum desired power: remember from the last section that when we increase power, we increase the likelihood of both obtaining statistically significant results *and* the likelihood that a statistically significant result because the alternative hypothesis is true than that the null hypothesis is true.
3. N for group 1: this is the sample size in one of your two groups.
4. Relative size of group 2 to group 1: if your sample sizes are equal in both groups, leave it at 1. If they aren't, you need to figure out the ratio. For example, if one group is $n = 20$ and the other is $n = 40$ then you would change this box to "2". You can easily calculate this by dividing n_2 by n_1.
5. α (type I error rate): remember from the last section that when we decrease power, we increase the likelihood of obtaining non-significant results when the null hypothesis is true *and* increase the likelihood that a statistically significant result means the alternative hypothesis is true. You shouldn't increase it above .05, but you should consider whether it would be useful to decrease it in your case.
6. Tails: specify whether you have a two-tailed (non-directional) or one-tailed (directional) hypothesis.

There are also options for four types of plots and whether to have explanatory text. For now, keep the explanatory text checked because it will help explain what is going on in the results. The plots are optional and I encourage you to check them out to see if they help you understand what is going on.

5.4.2 Power analysis example #1

Let's return to our example that we used in our interactive calculator before. We are going to calculate Power, set our effect size at $\delta = .8$, N for group 1 at 25, Relative size of group 2 to group 1 at 1 (equal sample sizes), and α (type I error rate) to .05. We'll assume we have a two-tailed hypothesis for now. You should get the following results. This table specifies that we defined the sample size, effect size, and alpha, which results in a power calculation of 79%.

A Priori Power Analysis

	User Defined			
Power	N_1	N_2	Effect Size	α
0.79	25	25	0.80	0.05

The results also provide a useful explanation:

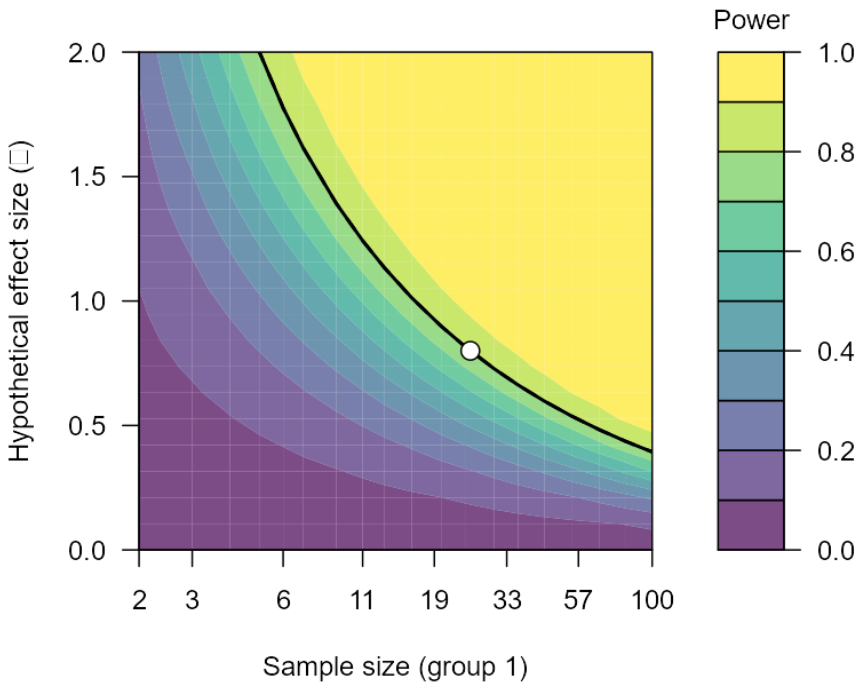
> A design with a sample size of 25 in each group can detect effect sizes of 0.8 with a probability of at least 0.791, assuming a two-sided criterion for detection that allows for a maximum Type I error rate of a=0.05.

This assumes that an effect size of .8 is the smallest effect size of interest. The next table shows us the power to detect various other effect sizes based on our alpha and sample size:

True effect size	Power to detect	Description
$0 < d = 0.566$	50%	Likely miss
$0.566 < d = 0.809$	50% – 80%	Good chance of missing
$0.809 < d = 1.041$	80% – 95%	Probably detect
$d = 1.041$	95%	Almost surely detect

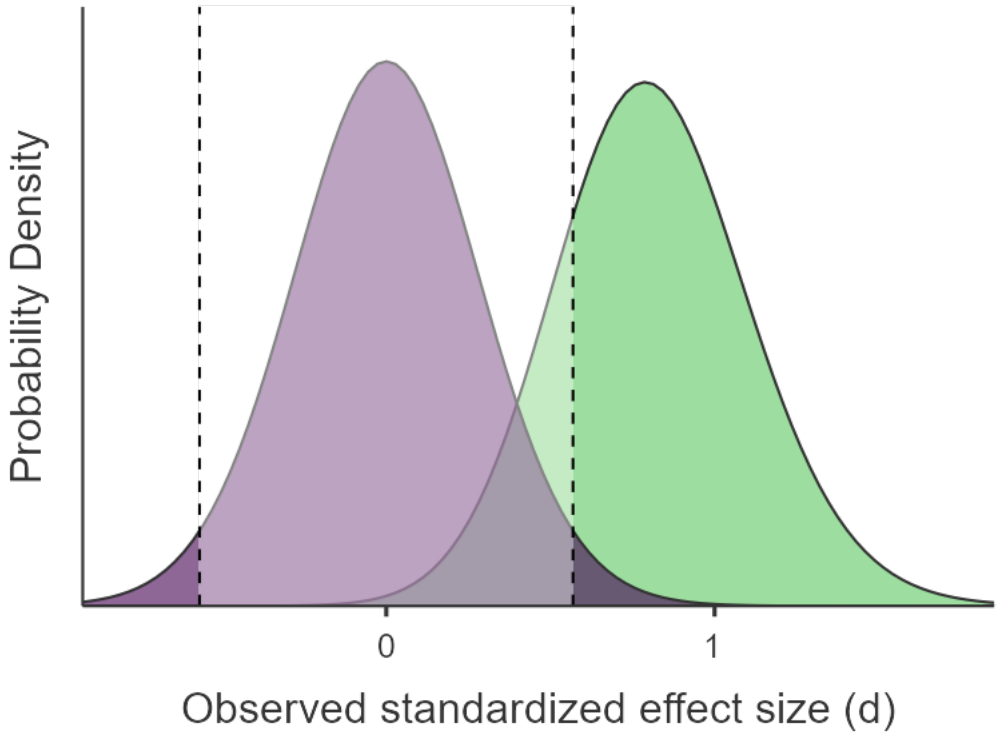
In other words, we are almost sure to detect really large effect sizes, but we'll likely miss really small effect sizes. This gives us a good hint to the relationship among BEAN: holding alpha and sample size constant, as effect sizes go up power goes up.

The Power Contour plot can show a bit more about how power (color), effect size (y-axis) and sample size (x-axis) all relate to one another. Notice how the x-axis is not linear. We are learning some more about the relationship among BEAN: increasing our sample size increases our power, holding alpha and effect size constant.



The next two plots are basically the Power Contour plot, but they shift power to the y-axis and either show effect size or N on the x-axis.

The last plot (Power Demonstration) helps us visualize our Type I and Type II errors and correct inferences nicely. The purple distribution is our null hypothesis distribution (centered at $d = 0$) and the green distribution is our alternative hypothesis distribution (centered at $d = .8$). The vertical dashed lines are the critical values of obtaining $p < .05$ on either side of the null distribution (so 2.5% on either side). The dark green area is therefore our power (80%) and the dark purple areas are our Type I error rate (5%). The light green area to the left of the dashed line is our Type II error rate (20%) and the light purple area is the probability of a correct inference assuming the null is true (95%). Keep in mind that these are the distributions of *both* hypotheses though, and in reality only one is true. We can just never know which is true; we can at best approximate it through repeated testing of effects.



5.4.3 Play with jpower

Play around some more with jpower. Try calculating other things (e.g., sample size or effect size). Play with power and see what increasing it does to your effect size and sample size. Play with effect sizes and see what decreasing them does to your power and sample size. Play with alpha (don't go higher than .05) and see what that does to your power, effect size, and sample size. And lastly, play with your sample size and see what it does to your power and effect size.

Your assignments for this unit will have you conduct power analyses based on various scenarios, so playing around with jpower will help prepare you for them.

Chapter 6

Inferential statistics

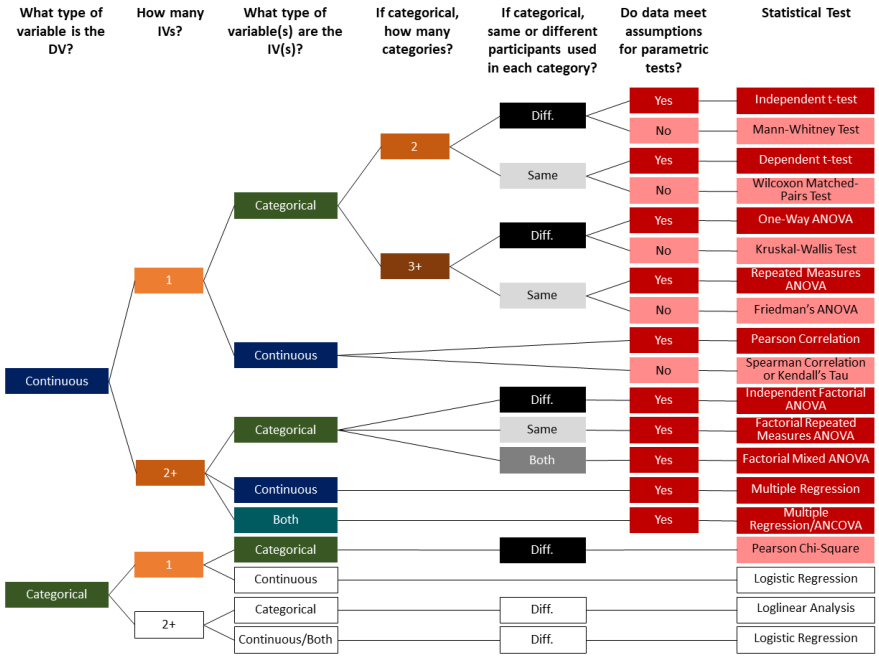
We've learned about hypothesis testing for inferential statistics two chapters ago and learned about some specific components of statistical testing in the last chapter. We have alluded to the fact that there are multiple inferential statistics we can perform, and that is the purpose of this chapter.

Although there are many more types, we are going to cover two basic types of inferential statistics:

1. **Parametric statistics**, which have an assumption of normal distribution
2. **Non-parametric statistics**, which have no assumptions about the distribution of the data

6.1 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. 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, the variable is either categorical (i.e., nominal or ordinal) or it's continuous (i.e., interval or ratio).

Next you specify 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. Backwards mapping involves determining what kind of data is needed to perform a particular statistical test.

Let's do some examples of forwards mapping. You may want to read the example and try your hand at it first and then check your answers!

6.1.1 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, 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 English, Math, or Science tests. The researchers has all participants complete all three tests.

1. What is the DV? What is the level of measurement? This one is tricky in how it's worded. There is one DV and it's simply test score. This is a continuous variable.
2. How many IVs are there? We only have one IV, and it is type of test.
3. What is the level of measurement of the IV? Type of test is a categorical variable.
4. How many categories to the IV? Type of test has three categories: English, Math, and Science.
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

6.1.2 Backwards map: Determine the data you need

Let me start off by saying we don't normally do this. We perform the test based on the data we have. But in our learning, we also want to ensure we learn all the tests. Imagine I gave you a dataset and wanted you to perform two different tests that I told you about.

Here are the variables in the dataset:

1. Mile time (continuous variable ranging from 5-30 minutes)
2. BMI (categorical variable of underweight, normal, or overweight)
3. Happiness at the start of the semester (continuous variable ranging from 0-10)
4. Happiness at the end of the semester (continuous variable ranging from 0-10)

If I told you I wanted you to perform a dependent t-test, what data would you use?

1. Assuming we meet the assumptions for a parametric test, we need to find a situation in which we have 1 continuous variable and 1 categorical variable with 2 levels in which participants are the *same* within each category (i.e., within-subjects variable).
2. We only have three continuous variables: mile time and our two happiness variables.
3. If we rethink happiness, we can realize that it's really a within-subjects variable. We are measuring happiness (our continuous DV) across two time points (start and end of the semester).
4. Therefore, we could perform a dependent t-test with our two happiness data points and see whether happiness differs across time in the semester.

If I told you I wanted you to perform a one-way independent ANOVA, what data would you use?

1. Assuming we meet the assumptions for a parametric test, we need to find a continuous DV and a single categorical IV with 3 or more levels in which participants are *different* across categories (i.e., between-subjects design).
2. We only have three continuous variables: mile time and our two happiness variables.
3. We only have one categorical variable (BMI), and it has 3 levels: underweight, normal, overweight.
4. Now this is where we need to think critically. What would be the most interesting test here? How BMI affects happiness or how BMI affects mile time? Weight and performance on running a mile seem to make most sense here. Therefore, we could look at how BMI affects mile time. Though keep in mind we are not randomizing here and so this is *not* an experimental design!

6.2 Checking assumptions

There are four basic assumptions of most parametric tests:

1. Normal distribution
2. Interval or ratio (i.e., continuous) dependent variable
3. Homogeneity of variances
4. Independent scores

Let's discuss these in turn and how to test for them.

6.2.1 Normal distribution

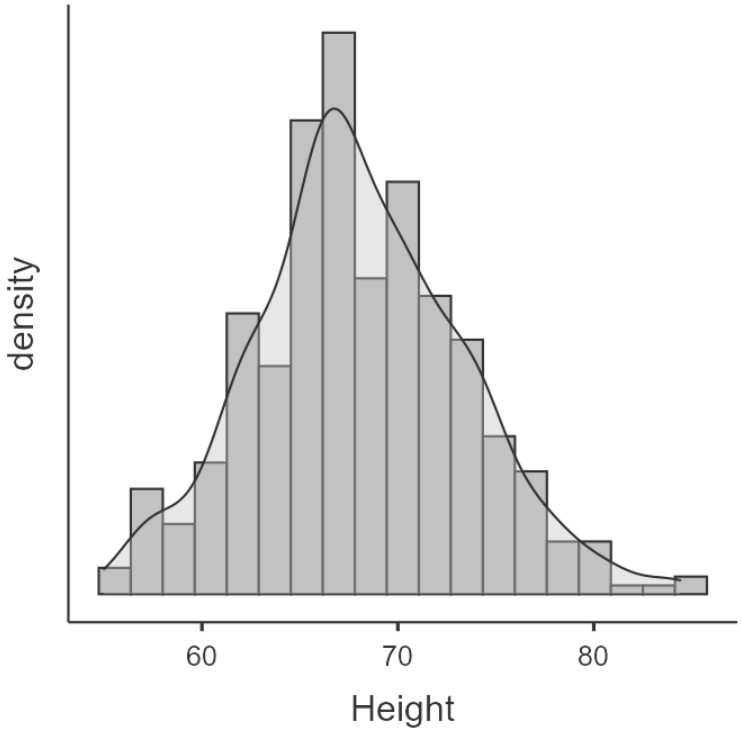
For all our statistics, our dependent variable needs to be normally distributed.[1] We have already covered what the normal distribution is multiple times, so let's move on to how to test for normality. There are four ways to test for normality **and we should test for normality using as many tests as we possibly can!**

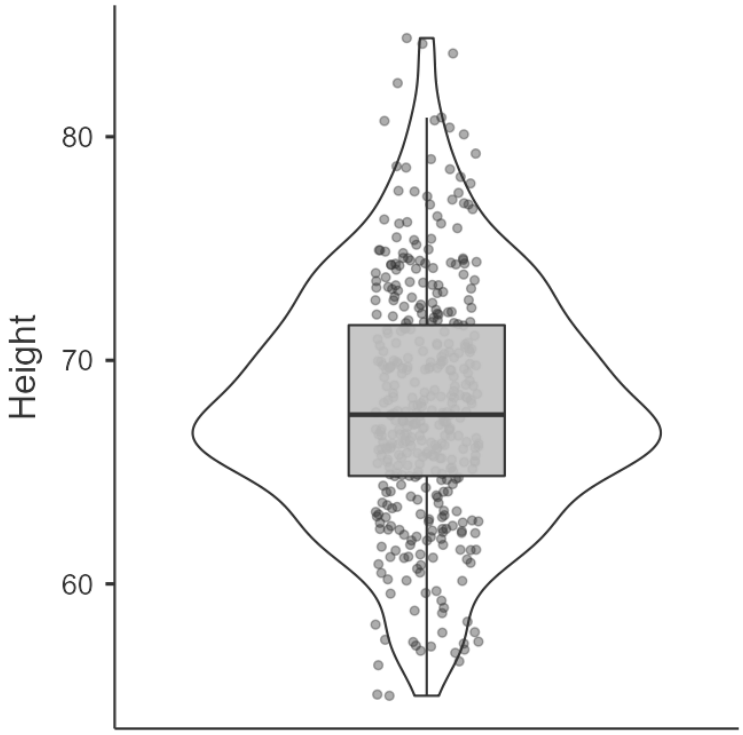
1. Visualize the distribution
2. Test the skew and kurtosis
3. Conduct a Shapiro-Wilk test
4. Visualize the Q-Q plot

6.2.1.1 VIsualize the distribution

In jamovi, we can go to the Explorations option and choose Descriptives. Under Plots, we can choose a histogram and/or density plot (figure on the left) or boxplot and/or violin plot and/or data points (figure on the right). We can just look at this data and visually inspect with our eyes whether the data is normally distributed. Height looks pretty fairly normally distributed in this case.

[1] Technically, it's that the *residuals* need to be normally distributed, but in the case of t-tests and ANOVAs the results are the same if we test for normality of residuals or the dependent variable.


density
60
70
80
Height



6.2.1.2 Test the skew and kurtosis

In jamovi, we can go to the Explorations option and choose Descriptives. Under statistics, choose skew and kurtosis. You'll have to do a bit more work to actually figure out whether the skew and kurtosis is problematic though.

For height, here is our skew and kurtosis:

Descriptives	**Height**
Skewness	.230
Std. error skewness	.121
Kurtosis	.113
Std. error kurtosis	.241

We need to calculate z-scores for skew and kurtosis. We do that by dividing the value by its standard error:

- Skew: .230 / .121 = 1.90
- Kurtosis: .113 / .241 = .47

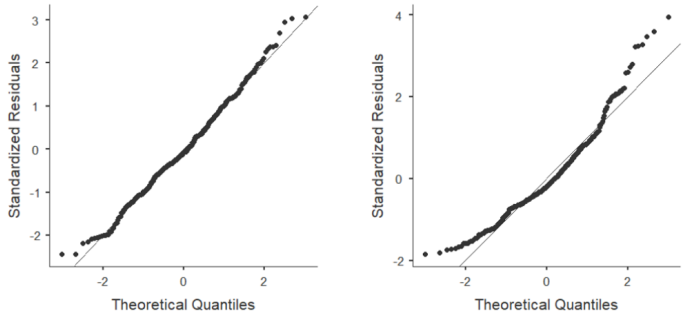
How do we know if it's problematic? **If the *z*-score for skew or kurtosis are less than |1.96| then it is *not* statistically significant and *is* normally distributed.** However, if the $z > |1.96|$ then it *is* statistically significant and *is not* normally distributed.

6.2.1.3 Shapiro-Wilk test

In jamovi, we can go to the Explorations option and choose Descriptives. Under statistics, choose Shapiro-Wilk. It will provide you the Shapiro-Wilk W test statistic and its respective p-value. In our case, Shapiro-Wilk's for height is 68.03, $p = .070$. **If the Shapiro-Wilk's test is *not* statistically significant then it *is* normally distributed.** However, if the Shapiro-Wilk's test *is* statistically significant then it *is not* normally distributed.

6.2.1.4 Q-Q plot

Last, we can visualize the Q-Q plot. In jamovi, we can go to the Explorations option and choose Descriptives. Under plots, choose Q-Q plot. We don't need to go into details of what is being visualized, but what we are looking for is that the data points fall along the diagonal line. On the figure on the left, we can see that the data is pretty well falling on the diagonal line (with small deviations at the tails) so we can say it looks normally distributed. However, on the figure on the right, the data points deviate from the diagonal line pretty significantly and so we can say it does not look normally distributed.



Remember we should look at all pieces of evidence to determine whether we meet the assumption of normal distribution. Typically, all four will support each other, but there are times when some evidence contradicts other evidence. You'll have to use your best judgment there, and often the visual inspection is the one I prioritize (e.g., if it doesn't look normally distributed but then the other tests suggest it is, I would probably be cautious and just say we don't meet the assumption).

6.2.2 Interval/ratio data

If we are performing a test that has a continuous DV, then the variable must be measured at the interval or ratio level. It is important that the data has proportional intervals between levels of the variable, and ordinal variables often do not meet this assumption.

It is very important to avoid treating ordinal variables as continuous variables. We cannot calculate a mean or difference between ordinal values, but we *can* for continuous variables. What is often done–and is often inappropriate to do–is treat Likert-scale items as a continuous DV. What we *can* do is take a sum or average of multiple Likert-scale items and treat that sum or average as a continuous DV.

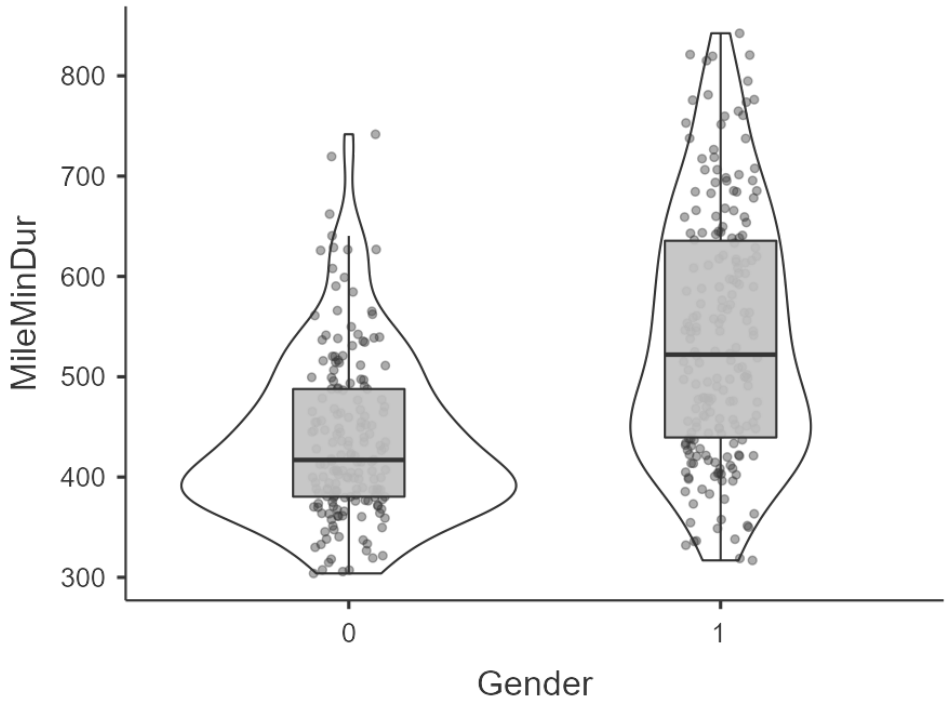
There is no "test" we can perform here. Rather, you will need to improve in recognizing whether data is interval/ratio (continuous) or ordinal (categorical).

6.2.3 Homogeneity of variance

Our third assumption is that the variance in the DV needs to be the same at each level of the IV. If we fail to meet the assumption, we say we have heterogeneity. We can test this assumption in two ways.

6.2.3.1 Visualize the distribution of data across groups

First, we can look at the data points across groups. This can be done by choosing a plot in the Descriptives analysis and adding your IV to the "Split By" box. For example, here's an example of data that violates the assumption of homogeneity of variance (gender by mile):



Similarly, the variance for Gender == 0 (male) is 6796.20 whereas the variance for Gender == 1 (female) is 15401.55. Clearly, there is much greater variability for females than males for time it takes to run the mile.

6.2.3.2 Levene's test

When we perform inferential statistics that have the assumption of homogeneity of variance, in jamovi there will be a check box to check the assumption. It will perform Levene's test. Here's the result of Levene's test for the independent t-test examining the effect of gender on mile duration:

Levene's	F	df1	df2	p
MileMinDur	41.33	1	381	<.001

Like the other tests above, **a non-significant Levene's test means we meet the assumption of homogeneity of variance.** However, if Levene's test is statistically significant, then we fail to meet the assumption of homogeneity of variance and have heterogeneity of variance. In this case, our test is statistically significant so, in combination with our plot above, we say we violated this assumption.

6.2.4 Independent scores

In between-subjects designs (e.g., the independent t-test or one-way ANOVA), data from different participants should be independent meaning that the response of one participant does not influence the response of another participant. We violate this assumption in the case of nested data (e.g., when our sample consists of students in three different classrooms, it is likely that students within classrooms are more similar than we would expect otherwise).

In within-subjects designs (e.g., the dependent t-test or repeated measures ANOVA), we automatically violate the assumption because *of course* the scores of one participant in one condition will relate to their scores on another condition. However, their scores should still not influence any other participant's response.

This is another assumption, like interval/ratio data, that we do not ever *test* but is a function of knowing our data.

6.3 Violated assumptions

What do you do if you have violated assumptions? Let's first talk about the assumptions we don't test (interval/ratio data and independent scores) before we turn to the other two assumptions (normality and homogeneity of variance).

6.3.1 Interval/ratio data

If you are trying to perform a statistical test with a categorical DV, the answer is simple: perform the test that requires a categorical DV and do not try to treat it as continuous. For example, if you have an ordinal DV and a categorical IV, you can perform a chi-square. If you have an ordinal DV and a continuous IV, you can perform a logistic regression (we won't be covering that in this class). Go back to the section on choosing the correct statistical test and you'll see four options of statistical tests that can be performed with a categorical DV.

6.3.2 Independent data

We won't be covering it in this class, but if you violate this assumption then you need to use a statistical test that accounts for nested data or can correlate the errors among dependent data. For example, multilevel modeling (aka hierarchical modeling) is one approach.

6.3.3 Normality or homogeneity of variance

If you violate either normality or homogeneity of variance, there are a few options you can choose.

6.3.3.1 Remove outliers

First, double check that you do not have any outliers. This is one reason why it's so important to visualize your data! There are also ways to test for outliers, but a visual inspection is often sufficient.

What can you do in case of outliers?

1. Ignore them, but this is not a good solution
2. Delete the outliers, but this is not recommended either because you lose data and we now know how important it is to have a large sample size
3. Transform the variable, especially if there are a lot of outliers
4. Winsorize, trim, or modify your data, especially if there are only a few outliers

We'll talk about transformations next, so let's cover Winsorizing or trimming your data. Winsorizing is used when both tails of the distribution have outliers whereas trimming is used when it's just one or a few outliers on one side of the distribution. In both cases, we replace the extreme valuse with the next-most-extreme values. There's more to Winsorizing than what I've described here, so I encourage you to learn more if you are interested.

We can do this through the Transform feature on jamovi. For example, here's what it looks like to trim the Reading variable to get rid of the few scores on the far left of the histogram. We want to take those values less than 60 and replace their scores with a new score of 60.

6.3.3.1.1 Is it an outlier? How do you know if data is an outlier? To look for outliers in single variables (aka univariate outliers), you can just look at your data. To look for multivariate outliers (outliers across multiple variables), you can look at Mahalanobis distance or Cook's distance, which you would need to use the Rj editor to perform in jamovi and we won't cover in this class.

6.3.3.2 Transforming data

If we violate the assumption of normality or homogeneity of variance (or both!) then we can explore whether transformations can improve the normality of our data. There are a variety of different transformations you can try, and here's a list of a few:

Name	Syntax	Corrects for positive skew?	Corrects for negative skew?	Corrects for unequal variances?
Log	log(X)	Yes	No	Yes
Square Root	sqrt(X)	Yes	No	Yes
Reciprocal	1/X	Yes	No	Yes

Name	Syntax	Corrects for positive skew?	Corrects for negative skew?	Corrects for unequal variances?
Reverse Score	(1+MAX) - X then do one of the above transformations	No	Yes	No

When you perform a transformation, then you need to check whether the transformation actually improved the situation. How do you do that? Check normality and homogeneity of variance again with your newly transformed data! You should check with the 4 methods to test for normality and 2 methods to test for homogeneity of variance.

6.3.3.3 Non-parametric tests

If all else fails–meaning there are no outliers or no transformations fix the violated assumption(s)–then you can perform a non-parametric test. These tests have *no* assumption of normally distributed data or homogeneity of variance! As you saw in the chart about choosing the correct statistical test, many of our parametric tests have non-parametric equivalents.

When we cover each individual statistical test (e.g., independent t-test) we will also cover its non-parametric equivalent (e.g., Mann-Whitney test). So stay tuned and just remember you have this option if you violate assumptions!

Chapter 7

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 t-test** has *independent* groups. The participants or things in group 1 are *not* the same as the participants or things in group 2.
3. The **dependent 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.

7.1 One sample t-test

7.1.1 Overview

The one-sample t-test is used to test the difference between our dependent variable mean and the mean of the population.

There are three different types of alternative hypotheses we could have for the one sample t-test:

1. **Two-tailed**
    - H_1: The sample mean has a different mean than the population mean.
    - H_0: There is no difference in means between the sample and population.
2. **One-tailed**

- H_1: The sample has a greater mean than the population.
- H_0: The mean for the sample is less than or equal to the mean for the population.

3. **One-tailed**

- H_1: The sample has a smaller mean than the population.
- H_0: The mean for the sample is greater than or equal to the mean for the population.

7.1.2 Look at the data

For this chapter, we're going to work with data from lsj-data. Open data from your Data Library in "lsj-data". Select and open "zeppo". This dataset is hypothetical data of 20 psychology students taking Dr. Zeppo's introductory statistics class. Dr. Zeppo wants to know if the psychology students tend to get the same grade as everyone else ($M = 67.5$) or whether they get a higher or lower grade. As psychologists, we're going to assume psychology students get higher grades. Therefore our hypotheses can be written up as such:

- H_1: Psychology students get higher grades than the population of Dr. Zeppo's students.
- H_0: There is no difference in student grades between psychology students and the population of Dr. Zeppo's students.

7.1.2.1 Data set-up

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 is the first ten rows of our data from the zeppo dataset.

7.1.2.2 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. Our overall data consists of 20 cases and the students in our dataset have a mean grade of 72.30 ($SD = 9.52$). The minimum and maximum values look accurate; theoretically, student grades should range from 0-100. Lastly, the distribution of data looks possibly not normally distributed. Although we have a pretty small sample size, we can proceed with our analyses. First, though, we need to check our assumptions.

	ID	x
1	1	50
2	2	60
3	3	60
4	4	64
5	5	66
6	6	66
7	7	67
8	8	69
9	9	70
10	10	74

Figure 7.1: One-sample t-test data in jamovi

Descriptives

	x
N	20
Missing	0
Mean	72.30
Median	75.00
Standard deviation	9.52
Minimum	50
Maximum	89

Figure 7.2: Descriptive statistics

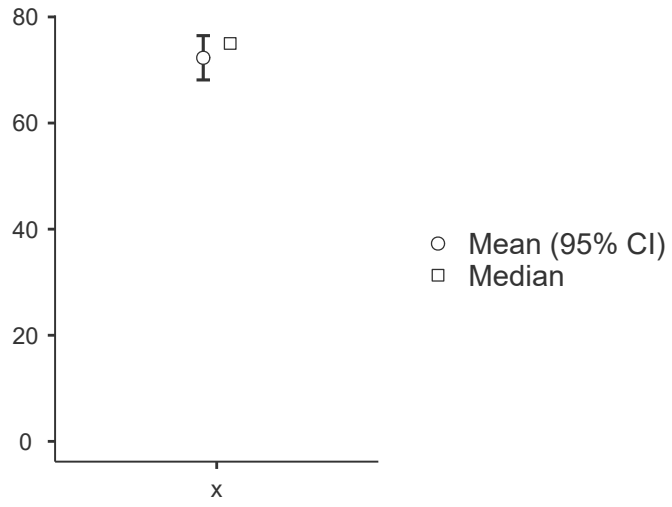


Figure 7.3: Plot of the data

7.1.3 Check assumptions

7.1.3.1 Assumptions

As a parametric test, the independent t-test has the same assumptions as other parametric tests:

1. The dependent variable is **normally distributed**
2. The dependent variable is **interval or ratio** (i.e., continuous)
3. Scores are **independent** between groups

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

However, we can and should test for the first assumption. Fortunately, the independent samples t-test in jamovi has two check boxes under "Assumption Checks" that lets us test for normality.

7.1.3.2 Checking assumptions

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.

7.1.3.2.1 Testing normality jamovi easily allows us to check for normality using the Shapiro-Wilk test and the Q-Q plot. The Shapiro-Wilk test was not statistically significant (W = .96, $p = .586$); therefore, this indicates the data is normally distributed. Furthermore, the lines are fairly close to the diagonal line in the Q-Q plot. We can conclude that we satisfy the assumption of normality.

Remember that we can also test for normality by **looking at our data** (e.g., a histogram or density plot, which you can see above) and by examining **skew and kurtosis**. However, you will need to view them using Exploration –> Descriptives, not in the t-tests menu. Here is our skew and kurtosis:

- **Skew**: $-.53/.51 = -1.04$
- **Kurtosis**: $.07/.99 = .07$

Remember that we divide the value by its standard error to determine the z-score. If the absolute value of it is below 1.96 then we assume it is normally distributed. Both skew and kurtosis meet the assumption of normality. In addition, so did all our other pieces of evidence of normality: Shapiro-Wilk's, visual examination of the distribution, and the Q-Q plot. Therefore we can assume we met the assumption of normality.

7.1.4 Perform the test

Now that we've satisfied the assumptions, we can perform the one sample t-test. Here are the steps for doing so in jamovi:

1. Go to the Analyses tab, click the T-Tests button, and choose "One Sample T-Test".
2. Move your dependent variable `x` to the Dependent Variables box.
3. Under Tests, select `Student's`. We'll learn about Wilcoxon rank when we discuss violated assumptions.
4. Under Hypothesis, input the population mean. In our case, it is `67.5` (this is the mean given to us by Dr. Zeppo). Also, select the hypothesis that matches your hypothesi. In our case, select `> Test value` because we believe psychology students have a higher mean than the test value (population mean).
5. Under Additional Statistics, select `Mean difference`, `Effect size`, `Descriptives`, and (optionally) `Descriptives plots`.
6. Under Assumption Checks, select both options: `Normality test` and `Q-Q plot`.

When you are done, your setup should look like this:

Normality Test (Shapiro-Wilk)

	W	p
x	0.96	0.586

Note. A low p-value suggests a violation of the assumption of normality

Q-Q plots

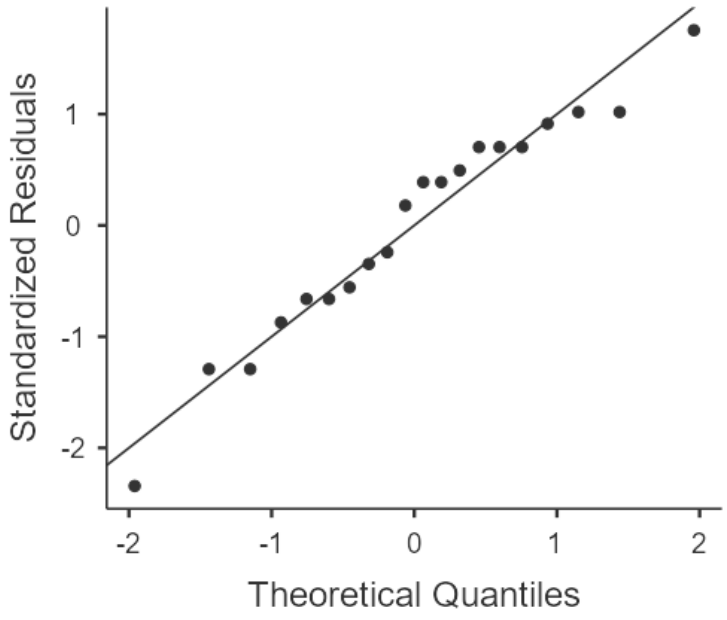


Figure 7.4: Testing normality in jamovi

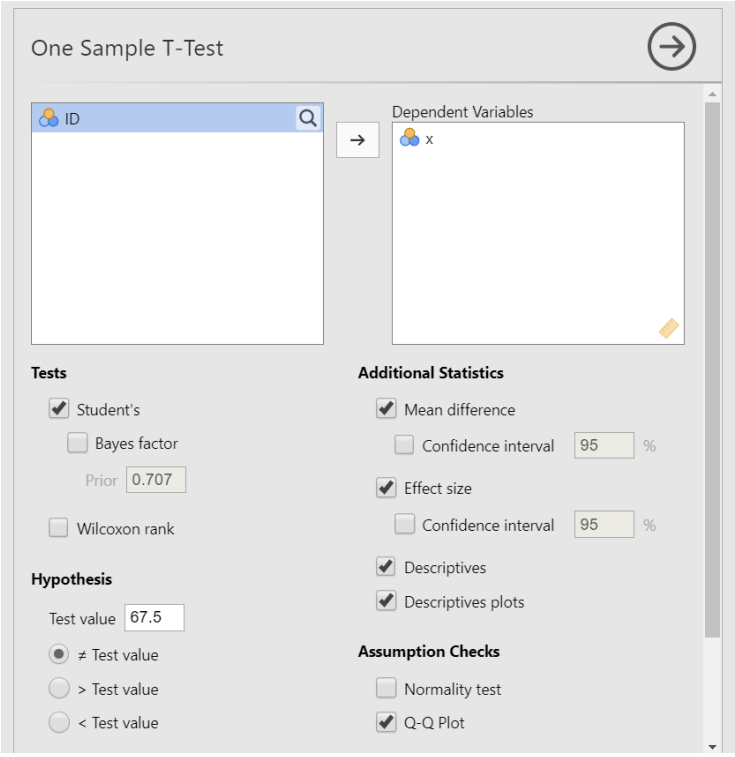


Figure 7.5: One sample t-test setup in jamovi

7.1.5 Interpret results

Once we are satisfied we have satisfied the assumptions for the independent t-test, we can interpret our results.

One Sample T-Test

One Sample T-Test

		Statistic	df	p	Mean difference		Effect Size
x	Student's t	2.25	19.00	0.036	4.80	Cohen's d	0.50

Note. H_a population mean ≠ 67.5

Descriptives

	N	Mean	Median	SD	SE
x	20	72.30	75.00	9.52	2.13

Figure 7.6: One sample t-test results in jamovi

Our p-value is less than .05, so our results are statistically significant. 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. Therefore, we reject our null hypothesis that the population mean is less than or equal to the sample mean of psychology students.

7.1.5.1 Write up the results in APA style

When writing up the results of a statistical test, we should always include the following information:

1. Description of your research question.
2. Description of your data. 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 effect size.
4. Interpretation of the results, including any other information as needed.

We can write up our results in APA something like this:

> Dr. Zeppo's psychology colleague hypothesized that his psychology students have a higher grade than the population of his students. Psychology students ($M = 72.30$, $SD = 9.52$, $n = 20$) had significantly higher grades than the population of Dr. Zeppo's students ($M = 67.50$), $t\ (19) = 2.25$, $p = .046$, $d = .50$.

Let's analyze that against the 4 things we need to report:

> **#1** Dr. Zeppo's psychology colleague hypothesized that his psychology students have a higher grade than the population of his students. **#4** Psychology students **#2** ($M = 72.30$, $SD = 9.52$, $n = 20$) had significantly higher grades than the population of Dr. Zeppo's students ($M = 67.50$), **#3** $t\ (19) = 2.25$, $p = .046$, $d = .50$.

Note that this is not the only way we can write up the results in APA format. The key is that we include all four pieces of information as specified above.

7.1.5.2 Visualize the results

By selecting `Descriptives plots` in the setup, you get the figure below. Personally, I don't think this is a very good plot. It's not very informative. It just provides the mean (circle), 95% confidence interval (blue bars), and the median.

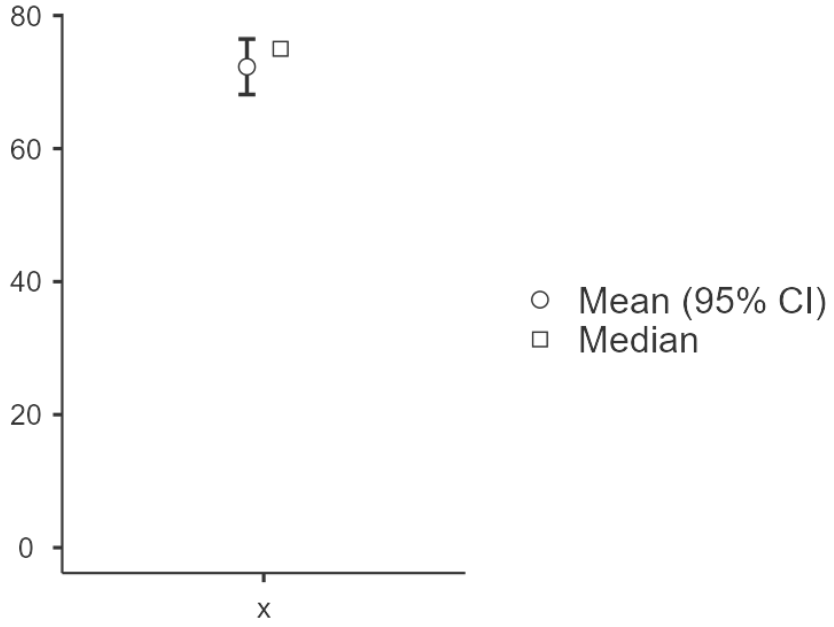


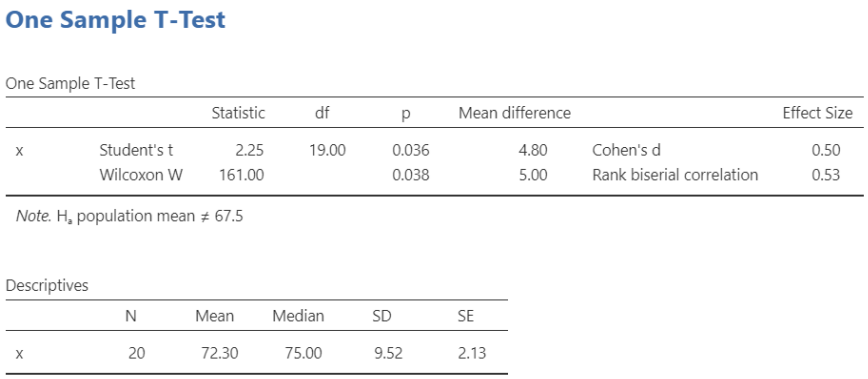
Figure 7.7: One sample t-test descriptives plot

Another default option in jamovi is via the Descriptives analysis. You can ask for the boxplot, violin plot, and data of your dependent variable split by your independent variable. This is a much better option. Not only does it show us our mean (black bars) and interquartile range (via the boxplot), but it also shows our distribution (violin plot) and data points (grey dots). This is much more informative. You can see this in the Look at your data section above.

7.1.6 In case of violated assumptions

If you fail to meet the assumption of normality, and no transformation fixes the data, then you can use the Wilcoxon W test.

The Wilcoxon W is not calculated based on the mean but rather the median. It has no assumptions about the distribution of data. Therefore, it is a non-parametric test. Here is what the output for the student's t-test and Wilcoxon W look like in jamovi:



One Sample T-Test

One Sample T-Test

		Statistic	df	p	Mean difference		Effect Size
x	Student's t	2.25	19.00	0.036	4.80	Cohen's d	0.50
	Wilcoxon W	161.00		0.038	5.00	Rank biserial correlation	0.53

Note. H_a population mean ≠ 67.5

Descriptives

	N	Mean	Median	SD	SE
x	20	72.30	75.00	9.52	2.13

Figure 7.8: All one sample t-test results in jamovi

7.1.6.0.1 Wilcoxon W in jamovi To conduct this in jamovi, under Tests select `Wilcoxon W`. You will interpret the results similarly to the one sample t-test:

> Dr. Zeppo's psychology colleague hypothesized that his psychology students have a higher grade than the population of his students. Using a Wilcoxon W test, psychology students ($Mdn = 75.00$, $SD = 9.52$, $n = 20$) had a higher grade than the population of Dr. Zeppo's students ($M = 67.5$), $W = 161$, $p = .038$, $r_{bn} = .53$.

Note that we no longer report the mean but rather the median. That is because Wilcoxon W is based on the median, not the mean score.

7.1.7 Additional information

7.1.7.1 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 and Bernadette: (Anastasia - Bernadette) = (74.53 - 69.06) = (5.48).

However, if we flipped it and tested the difference between Bernadette and Anastasia, our mean difference would be -5.48 and our t-statistic would be -2.12.

All that is to say, *your positive or negative t-statistic is arbitrary.* So do not fret!

However, it is important the sign of your t-statistic matches what you report. For example, notice the difference:

1. Anastasia's students had **higher** grades than Bernadette's, t (31) = **2.12**, p = .043, d = **.74**.
2. Bernadette's students had **lower** grades than Anastasia's, t (31) = **-2.12**, p = .043, d = **-.74**.

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

7.1.8 Your turn!

Open the `Sample_Dataset_2014.xlsx` file that we will be using for all Your Turn exercises. You can find the dataset here: Sample_Dataset_2014.xlsx Download

Perform one sample t-tests based on the following research questions. Think critically about whether you should be using a one-tailed or two-tailed hypothesis and check your assumptions so you know which test to use!

To get the most out of these exercises, try to first find out the answer on your own and then use the drop-down menus to check your answer.

1. **Do the students in our dataset have a higher Writing score than passing (M = 70)?**
   - Should you use a one-tailed or two-tailed hypothesis? one-tailed two-tailed
   - Which statistic should you use based on your assumptions? Student one sample t-test Wilcoxon rank one sample t-test
   - Do the students in our dataset have a higher Writing score than passing (M = 70)? yes no
2. **Do the students in our dataset have the same national average height of college students (M = 68 inches)?**
   - Should you use a one-tailed or two-tailed hypothesis? one-tailed two-tailed
   - Which statistic should you use based on your assumptions? Student one sample t-test Wilcoxon rank one sample t-test

- Do the students in our dataset have the same national average height of college students)? yes no

7.2 Independent t-test

7.2.1 Overview

The independent t-test is used to test the difference in our dependent variable between two different groups of observations. Our grouping variable is our independent variable. In other words, we use the independent t-test when we have a research question with a **continuous dependent variable** and a **categorical independent variable with two categories in which different participants are in each category**.

The independent t-test is also called the independent samples t-test and the Student's t-test.

There are three different types of alternative hypotheses we could have for the independent t-test:

1. **Two-tailed**
    - H_1: Group 1 has a different mean than Group 2.
    - H_0: There is no difference in means between the two groups.
2. **One-tailed**
    - H_1: Group 1 has a greater mean than Group 2.
    - H_0: The mean for Group 1 is less than or equal to the mean for Group 2.
3. **One-tailed**
    - H_1: Group 1 has a smaller mean than Group 2.
    - H_0: The mean for Group 1 is greater than or equal to the mean for Group 2.

7.2.2 Look at the data

For this chapter, we're going to work with data from lsj-data. Open data from your Data Library in "lsj-data". Select and 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?" We don't have a hypothesis that one does better than the other. Therefore our hypotheses can be written up as such:

- H_1: There is a difference in student grades between Anastasia's and Bernadette's classes.

- H_0: There is no difference in student grades between Anastasia's and Bernadette's classes.

7.2.2.1 Data set-up

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 is 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

Figure 7.9: Independent t-test data in jamovi

In the data above, what is your **independent variable**? ID grade tutor

In the data above, what is your **dependent variable**? ID grade tutor

7.2.2.2 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. Our overall data consists of 33 cases and the students in our dataset have a mean grade of 71.55 (SD = 7.80). The minimum and maximum values look accurate; theoretically, student grades should range from 0-100. Lastly, the distribution of data looks nice and normally distributed. Although we have a pretty small sample size, especially within each group, we can proceed with our analyses. First, though, we need to check our assumptions.

Descriptives

	grade
N	33
Mean	71.55
Median	72
Standard deviation	7.80
Minimum	55
Maximum	90

Figure 7.10: Descriptive statistics

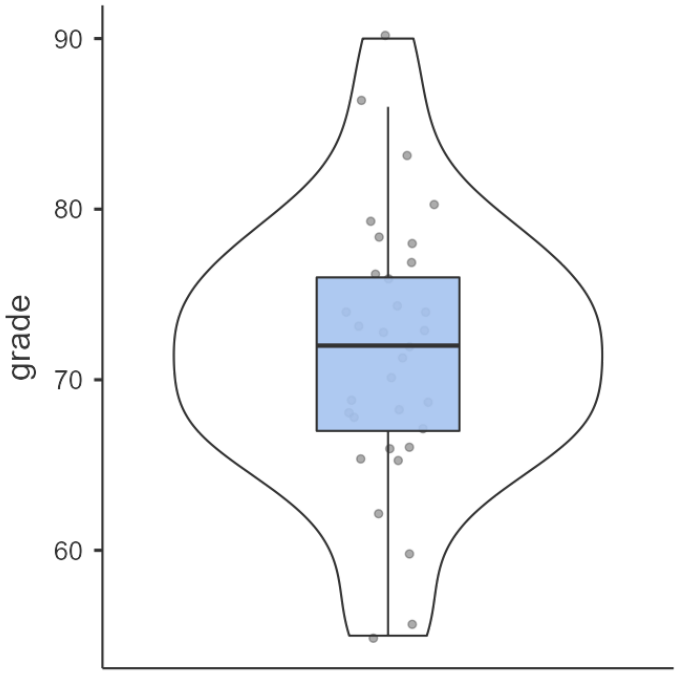


Figure 7.11: Plot of the data

7.2.3 Check assumptions

7.2.3.1 Assumptions

As a parametric test, the independent t-test 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 independent samples t-test in jamovi has two check boxes under "Assumption Checks" that lets us test for both assumptions.

7.2.3.2 Checking assumptions

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.

7.2.3.2.1 Testing normality jamovi easily allows us to check for normality using the Shapiro-Wilk test and the Q-Q plot. The Shapiro-Wilk test was not statistically significant (W = .98, p = .827); therefore, this indicates the data is normally distributed. Furthermore, the lines are fairly close to the diagonal line in the Q-Q plot. We can conclude that we satisfy the assumption of normality.

Remember that we can also test for normality by **looking at our data** (e.g., a histogram or density plot, which you can see above) and by examining **skew and kurtosis**. However, you will need to view them using Exploration –> Descriptives, not in the t-tests menu. Here is our skew and kurtosis:

- **Skew**: .06/.41 = .15
- **Kurtosis**: .33/.80 = .41

Remember that we divide the value by its standard error to determine the z-score. If the absolute value of it is below 1.96 then we assume it is normally distributed. Both skew and kurtosis meet the assumption of normality. In addition, so did all our other pieces of evidence of normality: Shapiro-Wilk's,

Assumptions

Normality Test (Shapiro-Wilk)

	W	p
grade	0.98	0.827

Note. A low p-value suggests a violation of the assumption of normality

Plots

grade

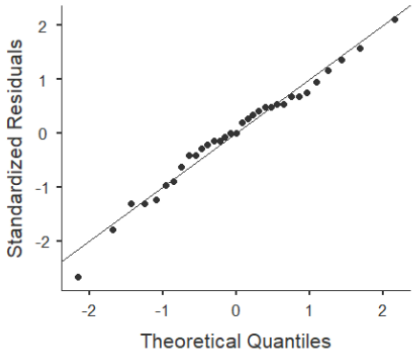


Figure 7.12: Testing normality in jamovi

visual examination of the distribution, and the Q-Q plot. Therefore we can assume we met the assumption of normality.

Descriptives

	grade
Skewness	0.06
Std. error skewness	0.41
Kurtosis	0.33
Std. error kurtosis	0.80

Figure 7.13: Plot of the data

7.2.3.2.2 Testing homogeneity of variance We test for homogeneity of variance using the Levene's test. The Levene's test was not statistically significant (F [1, 31] = 2.49, p = .125); therefore, this indicates our data satisfies the assumption of homogeneity of variance. However, I would add a caveat that we have a small sample of data (n = 15 for Anastasia and n = 18 for Bernadette) and the standard deviations are quite different from one another (SD = 9.00 vs 5.77, respectively). We should have tried to collect more data.

Assumptions

Homogeneity of Variances Test (Levene's)

	F	df	df2	p
grade	2.49	1	31	0.125

Note. A low p-value suggests a violation of the assumption of equal variances

[3]

Figure 7.14: Testing homogeneity of variance in jamovi

7.2.4 Perform the test

Now that we've satisfied the assumptions, we can perform the independent t-test. Here are the steps for doing so in jamovi:

1. Go to the Analyses tab, click the T-Tests button, and choose "Independent Samples T-Test".

2. Move your dependent variable `grade` to the Dependent Variables box and your independent variable `tutor` to the Grouping Variable box.
3. Under Tests, select `Student's`. We'll learn about Welch's and Mann-Whitney U under the violated assumptions section.
4. Under Hypothesis, select the hypothesis that matches your research question. In our case, select `Group 1  Group 2` because we have a two-sided hypothesis.
5. Under Additional Statistics, select `Mean difference`, `Effect size`, `Descriptives`, and (optionally) `Descriptives plots`.
6. Under Assumption Checks, select all three options: `Homogeneity test`, `Normality test`, and `Q-Q plot`.

When you are done, your setup should look like this

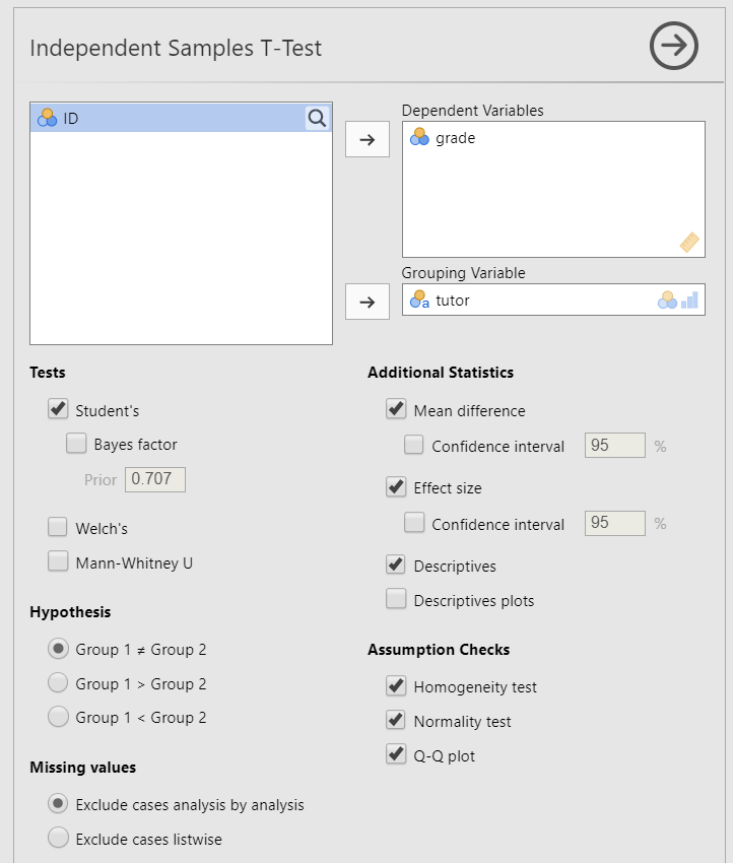


Figure 7.15: Independent t-test setup in jamovi

7.2.5 Interpret results

Once we are satisfied we have satisfied the assumptions for the independent t-test, we can interpret our results.

Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference		Effect Size
grade	Student's t	2.12	31.00	0.043	5.48	2.59	Cohen's d	0.74

Group Descriptives

	Group	N	Mean	Median	SD	SE
grade	Anastasia	15	74.53	76.00	9.00	2.32
	Bernadette	18	69.06	69.00	5.77	1.36

Figure 7.16: Independent t-test results in jamovi

Our p-value is less than .05, so our results are statistically significant. 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. Therefore, we reject our null hypothesis that there is no difference between the two groups.

7.2.5.1 Write up the results in APA style

When writing up the results of a statistical test, we should always include the following information:

1. Description of your research question.
2. Description of your data. 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 effect size.
4. Interpretation of the results, including any other information as needed.

We can write up our results in APA something like this:

> The research question was whether there was a difference in student grades between Anastasia's and Bernadette's classes. Anastasia's students ($M = 74.53$, $SD = 9.00$, $n = 15$) had significantly higher grades than Bernadette's students ($M = 69.06$, $SD = 5.77$, $n = 18$), $t\,(31) = 2.12$, $p = .043$, $d = .74$.

Let's analyze that against the 4 things we need to report:

> **#1:** The research question was whether there was a difference in student grades between Anastasia's and Bernadette's classes. **#4**

> Anastasia's students **#2** ($M = 74.53$, $SD = 9.00$, $n = 15$) had significantly higher grades than Bernadette's students **#2** ($M = 69.06$, $SD = 5.77$, $n = 18$), **#3** $t\ (31) = 2.12$, $p = .043$, $d = .74$.

Sometimes, people like to put the statistics inside a parentheses. In that case, you need to change the parentheses around the degrees of freedom as brackets. Here's another example write-up of the results in APA style:

> **#1** I tested the difference in grades between Anastasia's students **#2** ($M = 74.53$, $SD = 9.00$, $n = 15$) and Bernadette's students ($M = 69.06$, $SD = 5.77$, $n = 18$). **#3** An independent samples t-test showed that the 5.48 mean difference between the tutor's student was statistically significant ($t\ [31] = 2.12$, $p = .043$, $d = .74$). **#4** Therefore, we reject the null hypothesis that there is no difference in grades between the two classes.

Note that these are not the only way we can write up the results in APA format. The key is that we include all four pieces of information as specified above.

7.2.5.2 Visualize the results

By selecting `Descriptives plots` in the setup, you get the figure below. Personally, I don't think this is a very good plot. It's not very informative. It just provides the mean (circle), 95% confidence interval (blue bars), and the median.

Another default option in jamovi is via the Descriptives analysis. You can ask for the boxplot, violin plot, and data of your dependent variable split by your independent variable. This is a much better option. Not only does it show us our mean (black bars) and interquartile range (via the boxplot), but it also shows our distribution (violin plot) and data points (grey dots). This is much more informative.

Oftentimes, people display results in a simple bar chart, often adding error bars (either 95% CI or SE error bars). But this is also not a great chart because it lacks information about the underlying distribution of data. Therefore, for the independent t-test I recommend the visualization shown above.

7.2.6 In case of violated assumptions

If you fail to meet one or both of the assumptions of normality (and no transformations fixed your data) and homogeneity of variances, jamovi has the alternative statistics easily built in. Here's what statistic you should choose based on satisfying assumptions:

Table 7.1: Independent t-test to perform based on assumptions

	Normality: satisfied	**Normality: not satisfied**
Homogeneity of Variance: satisfied	Student's t-test	Mann-Whitney U

	Normality: satisfied	**Normality: not satisfied**
Homogeneity of Variance: not satisfied	Welch's t-test	Mann-Whitney U

The Welch's t-test has three main differences from the independent samples t-test: (a) the standard error (SE) is not a pooled estimate, (b) the degrees of freedom are calculated very different (not N - 2), and (c) it does not have an assumption of homogeneity of variance. Note that Welch's t-test is *not* a non-parametric test because it still has the assumption of a normal distribution.

The Mann-Whitney U is not calculated based on the mean but rather the median and compares ranks of values across the two groups: it has no assumptions about the distribution of data or homogeneity of variances. Therefore, it is a non-parametric test. Here is what the output for all three tests look like:

7.2.6.1 Welch's t-test

To conduct this in jamovi, under Tests select `Welch's`. You will interpret the results similarly to the independent t-test:

> Using a Welch's t-test, there was not a statistically significant difference in grades between Anastasia's students (M = 74.53, SD = 9.00, n = 15) and Bernadette's students (M = 69.06, SD = 5.77, n = 18), t (23.02) = 2.03, p = .054, d = .72.

Why is it no longer statistically significant? Which result should you trust? In reality, the difference in p-values is likely due to chance. However, the independent t-test and Welch's test have different strengths and weaknesses. If the two populations really do have equal variances, then the independent t-test is slightly more powerful (lower Type II error rate) than the Welch's test. However, if they *don't* have the same variances, then the assumptions of the independent t-test are violated and you may not be able to trust the results; you may end up with a higher Type I error rate. So it's a trade-off.

Which should you use? I tend to prefer always using Welch's t-test because if the variances are equal, then there will be practically no difference between the independent and Welch's t-test. But if the variances are not equal, then Welch's t-test will outperform the independent t-test. For that reason, defaulting to the Welch's t-test makes most sense to me.

7.2.6.2 Mann-Whitney U test

If you do not satisfy the assumption of normality (regardless of whether you satisfy the assumption of homogeneity of variance), you should either try to transform your data to be normally distributed or you will need to use a non-parametric test. In this case, if you originally wanted to perform an independent t-test, the non-parametric equivalent test is the Mann-Whitney U test.

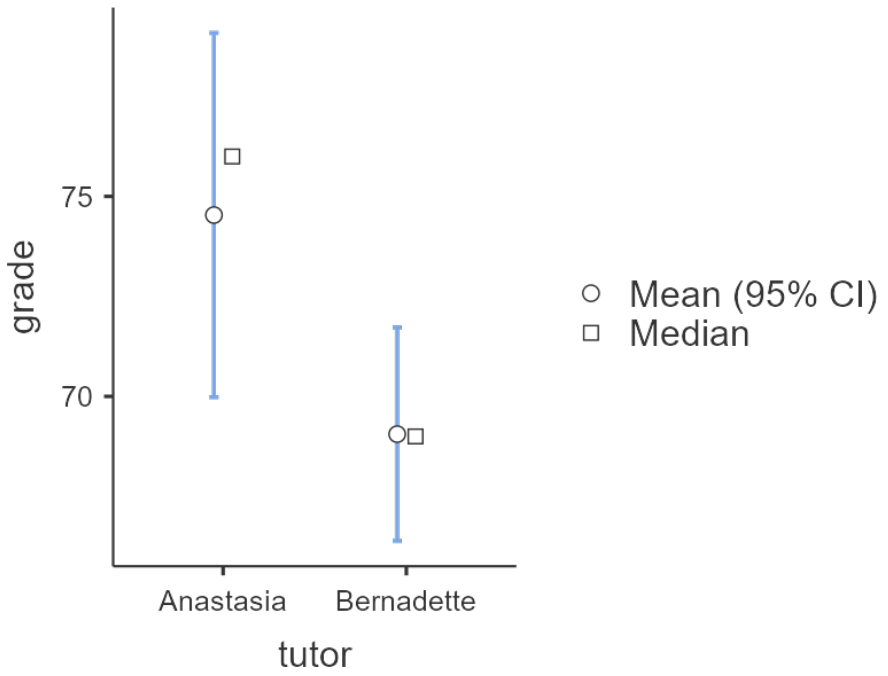


Figure 7.17: Independent t-test descriptives plot

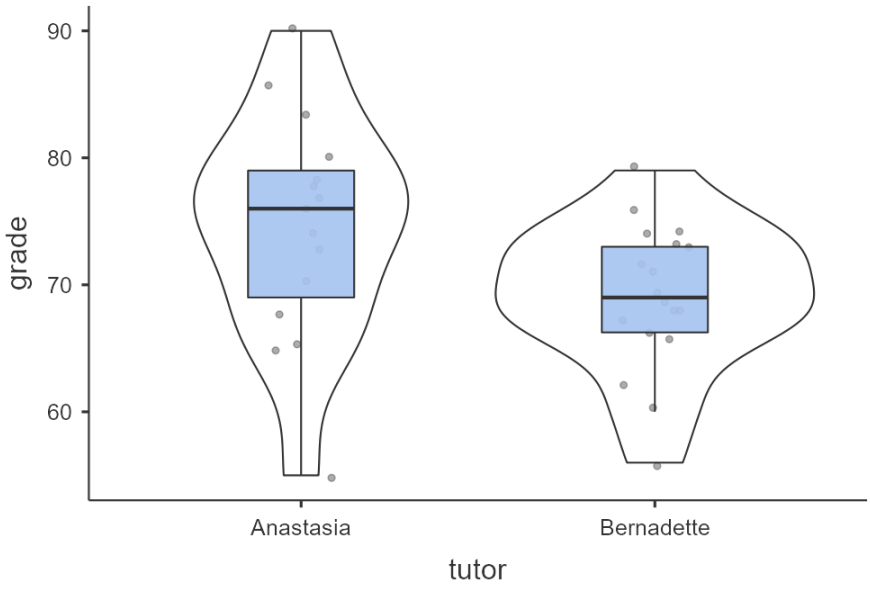


Figure 7.18: Independent t-test descriptives plot

Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference		Effect Size
grade	Student's t	2.12	31.00	0.043	5.48	2.59	Cohen's d	0.74
	Welch's t	2.03	23.02	0.054	5.48	2.69	Cohen's d	0.72
	Mann-Whitney U	79.50		0.046	6.00		Rank biserial correlation	0.41

Group Descriptives

	Group	N	Mean	Median	SD	SE
grade	Anastasia	15	74.53	76.00	9.00	2.32
	Bernadette	18	69.06	69.00	5.77	1.36

Figure 7.19: All independent t-test results in jamovi

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*.

> 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 ($Mdn = 69$, $n = 18$), $U = 79.50$, $p = .054$, $r_{pb} = .41$.

7.2.7 Your turn!

Open the `Sample_Dataset_2014.xlsx` file that we will be using for all Your Turn exercises. You can find the dataset here: Sample_Dataset_2014.xlsx Download

Perform independent t-tests based on the following research questions. Think critically about whether you should be using a one-tailed or two-tailed hypothesis and check your assumptions so you know which test to use!

To get the most out of these exercises, try to first find out the answer on your own and then use the drop-down menus to check your answer.

1. **Does height differ by gender (Gender: male = 0, female = 1)?**
    - Should you use a one-tailed or two-tailed hypothesis? one-tailed two-tailed
    - Which statistic should you use based on your assumptions? independent t-test Welch's t-test Mann Whitney U
    - Does height differ by gender? yes no
2. **Do athletes (Athlete: athletes = 1, non-athlete = 0) have faster sprint times than non-athletes?**
    - Should you use a one-tailed or two-tailed hypothesis? one-tailed two-tailed
    - Which statistic should you use based on your assumptions? independent t-test Welch t-test Mann Whitney U
    - Do athletes have faster sprint times than non-athletes? yes no
3. **Do students who live on campus (LiveOnCampus: on campus = 1, off campus = 0) have higher English scores than students who live off campus?**
    - Should you use a one-tailed or two-tailed hypothesis? one-tailed two-tailed

- Which statistic should you use based on your assumptions? independent t-test Welch t-test Mann Whitney U
- Does students who live on campus have higher English scores? yes no

4. **Does athletic status relate to math scores?**
    - Should you use a one-tailed or two-tailed hypothesis? one-tailed two-tailed
    - Which statistic should you use based on your assumptions? independent t-test Welch's t-test Mann Whitney U
    - Does athletic status relate to math scores? yes no

7.3 Dependent t-test

7.3.1 Overview

The dependent t-test is used to test the difference in our dependent variable between two categories in which participants are the *same* across categories. Our category variable is our independent variable. In other words, we use the dependent t-test when we have a research question with a **continuous dependent variable** and a **categorical independent variable with two categories in which the same participants are in each category**.

The dependent t-test is also called a dependent samples t-test or paired samples t-test.

There are three different types of alternative hypotheses we could have for the dependent t-test:

1. **Two-tailed**
    - H_1: There is a difference in means between the two time points or conditions.
    - H_0: There is no difference in means between the two time points or conditions.
2. **One-tailed**
    - H_1: The mean at time 1 or condition 1 is greater than the mean at time 2 or condition 2.
    - H_0: The mean at time 1 or condition 1 is less than or equal to the mean at time 2 or condition 2.
3. **One-tailed**
    - H_1: The mean at time 1 or condition 1 is smaller than the mean at time 2 or condition 2.

- H_0: The mean at time 1 or condition 1 is greater than or equal to the mean at time 2 or condition 2.

7.3.2 Look at the data

For this chapter, we're going to work with data from lsj-data. Open data from your Data Library in "lsj-data". Select and open "Chico". This dataset 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? First, let's determine our hypotheses:

What is Dr. Chico's alternative hypothesis? test 1 scores are different from test 2 scores test 1 scores are greater than test 2 scores test 1 scores are less than test 2 scores

7.3.2.1 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.

	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
7	student7	59.8	60.8
8	student8	50.8	51.6
9	student9	62.5	64.3
10	student10	61.9	63.2

Figure 7.20: Dependent t-test data in jamovi

In the data above, what is your **independent variable**? id grade test score grade_test1 grade_test2

In the data above, what is your **dependent variable**? id grade test score grade_test1 grade_test2

7.3.2.2 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. 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.

Descriptives

	grade_test1	grade_test2
N	20	20
Missing	0	0
Mean	56.98	58.38
Median	57.70	59.70
Standard deviation	6.62	6.41
Minimum	42.90	44.60
Maximum	71.70	72.30

Figure 7.21: Descriptive statistics

7.3.3 Check assumptions

7.3.3.1 Assumptions

As a parametric test, the dependent t-test has the same assumptions as other parametric tests minus the homogeneity of variance assumption 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*

Descriptives

	grade_test1	grade_test2
N	20	20
Missing	0	0
Mean	56.98	58.38
Median	57.70	59.70
Standard deviation	6.62	6.41
Minimum	42.90	44.60
Maximum	71.70	72.30

Figure 7.22: Descriptive statistics

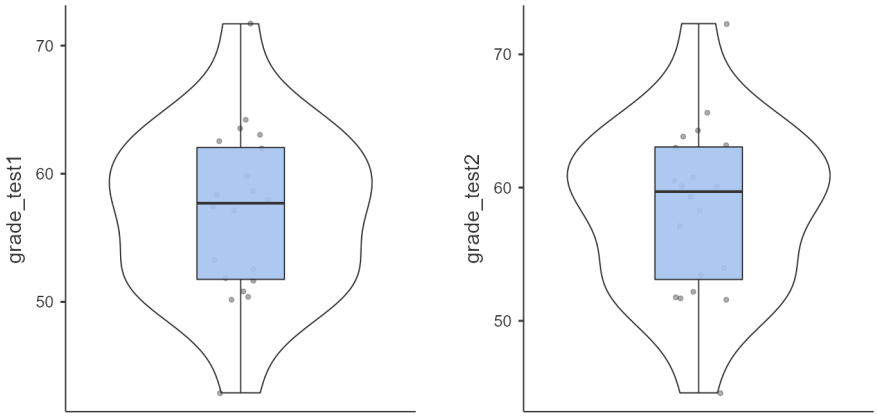


Figure 7.23: Plots of the data

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

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.

7.3.3.2 Checking assumptions

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.

7.3.3.2.1 Testing normality Notice how our dependent variable is really the difference in scores, and therefore that is what we are testing for normality. We test for normality using the Shapiro-Wilk test and the Q-Q plot. The Shapiro-Wilk test was not statistically significant ($W = .97$, $p = .678$); therefore, this indicates the data is normally distributed. 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). We can conclude that we satisfy the assumption of normality.

7.3.4 Perform the test

Now that we've satisfied the assumptions, we can perform the dependent t-test. Here are the steps for doing so in jamovi:

1. Go to the Analyses tab, click the T-Tests button, and choose "Paired Samples T-Test".
2. Move both measurements of your dependent variable (`grade_test1` and `grade_test2`) to the Paired Variables box.
3. Under Tests, select `Student's`. We'll learn more about the Wilcoxon rank option, which is an option if you fail to meet the assumption of normality.
4. Under Hypothesis, choose the correct hypothesis: Measure 1 is not equal to Measure 2 Measure 1 > Measure 2 Measure 1 < Measure 2
5. Under Additional Statistics, select `Mean difference`, `Effect size`, `Descriptives`, and (optionally) `Descriptives plots`.
6. Under Assumption Checks, select both options: `Normality test` and `Q-Q plot`. You'll check these assumptions first, which is discussed above.

When you are done, your setup should look like this

Normality Test (Shapiro-Wilk)

			W	p
grade_test1	-	grade_test2	0.97	0.678

Note. A low p-value suggests a violation of the assumption of normality

Plots

grade_test1 - grade_test2

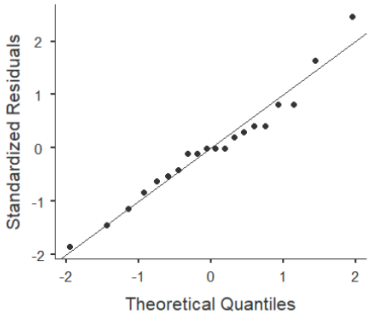


Figure 7.24: Testing normality in jamovi

7.3.5 Interpreting results

Once we are satisfied we have satisfied the assumptions for the dependent t-test, we can interpret our results.

Our p-value is less than .05, so our results are statistically significant. Therefore, we reject the null hypothesis that there is no difference between the two groups.

7.3.5.1 Write up the results in APA style

When writing up the results of a statistical test, we should always include the following information:

1. Description of your research question.
2. Description of your data. 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 effect size.
4. Interpretation of the results, including any other information as needed.

We can write up our results in APA something like this:

> Dr. Chico tested whether students performed better on the second test compared to the first test. The 20 students in performed better

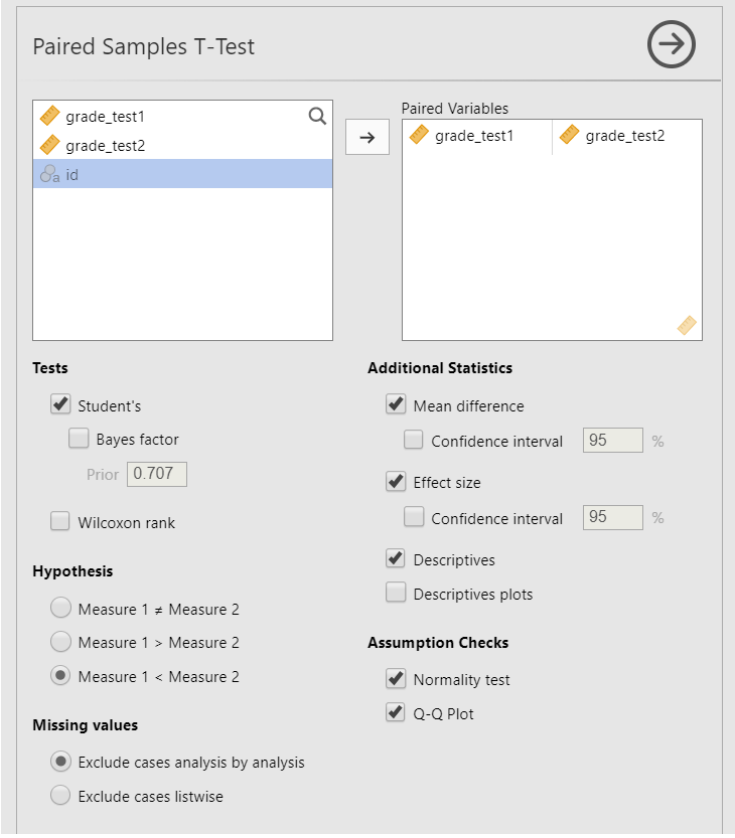


Figure 7.25: Dependent t-test setup in jamovi

Paired Samples T-Test

Paired Samples T-Test

			statistic	df	p	Mean difference	SE difference		Effect Size
grade_test1	grade_test2	Student's t	−6.48	19.00	< .001	−1.40	0.22	Cohen's d	−1.45

Note. H_a Measure 1 < Measure 2

Descriptives

	N	Mean	Median	SD	SE
grade_test1	20	56.98	57.70	6.62	1.48
grade_test2	20	58.38	59.70	6.41	1.43

Figure 7.26: Dependent t-test results in jamovi

> on the second test ($M = 58.38$, $SD = 6.41$) than they did on the first test ($M = 56.98$, $SD = 6.62$), $t(19) = 6.48$, $p < .001$, $d = 1.45$.

Remember in the previous chapter that our t-test can be negative but we can always flip the interpretation.

7.3.5.2 Visualize the results

By selecting `Descriptives plots` in the setup, you get the figure below. Personally, I don't think this is a very good plot. It's not very informative. It just provides the mean (circle), 95% confidence interval (blue bars), and the median.

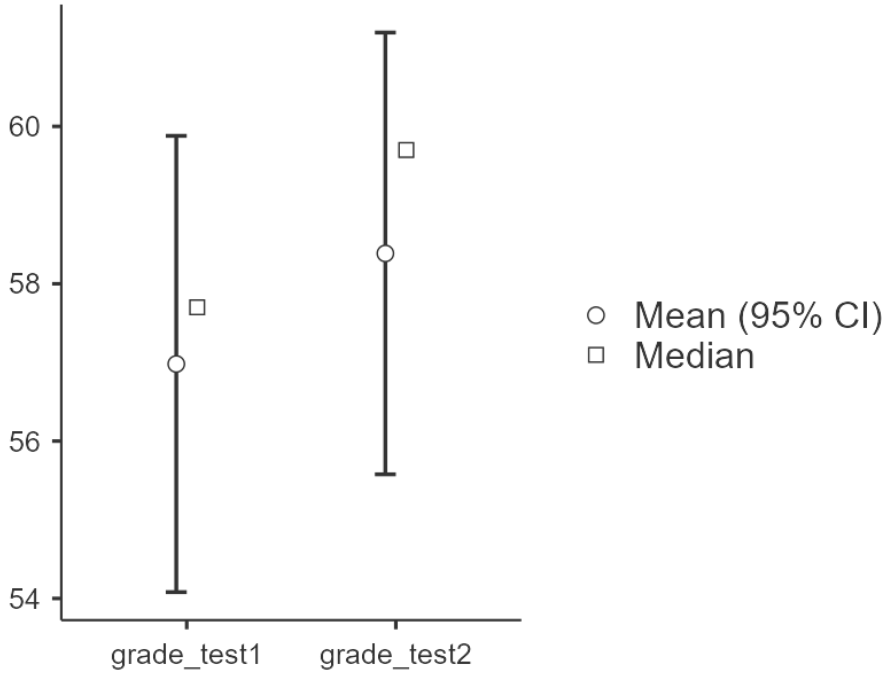


Figure 7.27: Dependent t-test descriptives plot

Another option is to use the Descriptives plots available in jamovi, which we see above in the Look at the data section above. I wish there were a way to combine them into one graph, but unfortunately there isn't within jamovi. Instead, you'll have to go into the Rj editor and use R code to reshape the data from wide format to long format and then call the descriptives syntax to produce the plot. You *could* copy-paste the data into a new dataset, but I always try to avoid manual work when doing analyses because something *always* goes wrong.

Here's what this looks like in jamovi:

To make our lives a bit easier, here is the code in the Rj editor that you can

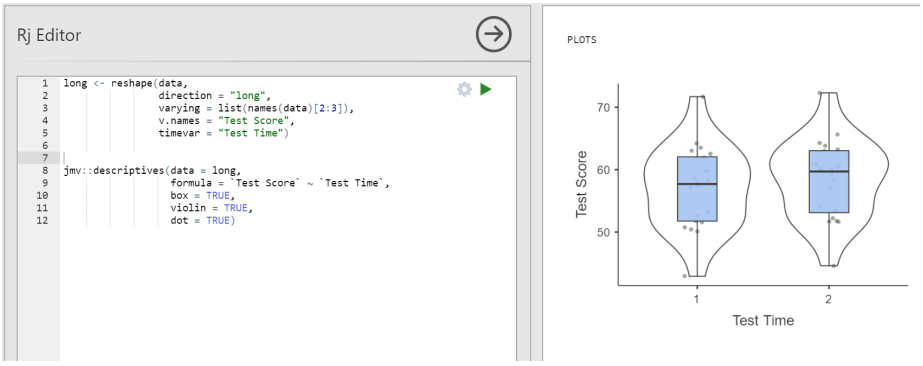


Figure 7.28: Revised plot using the Rj editor

copy-paste (that kind of copy-pasting I allow!) into your own jamovi. You'll need to edit things like the list of names in line 3, the name of your DV (line 4), the name of your IV (line 5), and re-specify those names in line 9.

```
long <- reshape(data,
                direction = "long",
                varying = list(names(data)[2:3]),
                v.names = "Test Score",
                timevar = "Test Time")


jmv::descriptives(data = long,
                  formula = `Test Score` ~ `Test Time`,
                  box = TRUE,
                  violin = TRUE,
                  dot = TRUE)
```

7.3.6 In case of violated assumptions

If you violated the assumption of normality and no transformation fixed your data, then you can perform the non-parametric version of the dependent t-test called the Wilcoxon Rank test. As a reminder, non-parametric tests do not make assumptions about the distribution of data because it deals with the *median* not the *mean*.

Here is the output for both the dependent t-test and the Wilcoxon rank test:

7.3.6.1 Wilcoxon rank

To conduct this in jamovi, under Tests select `Wilcoxon rank`. You will interpret the results similarly to the dependent t-test:

> Using Wilcoxon rank test, students' test scores were significantly

Paired Samples T-Test

Paired Samples T-Test

			Statistic	df	p		Effect Size
grade_test1	grade_test2	Student's t	−6.48	19.00	< .001	Cohen's d	−1.45
		Wilcoxon W	2.00 [a]		< .001	Rank biserial correlation	−0.98

Note. H_a Measure 1 < Measure 2

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

Descriptives

	N	Mean	Median	SD	SE
grade_test1	20	56.98	57.70	6.62	1.48
grade_test2	20	58.38	59.70	6.41	1.43

Figure 7.29: All dependent t-test results in jamovi

higher at the second test ($Mdn = 59.70$) than at the first test ($Mdn = 57.70$), W = 2.00, $p < .001$.

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

7.3.7 Your turn!

Open the `Sample_Dataset_2014.xlsx` file that we will be using for all Your Turn exercises. You can find the dataset here: Sample_Dataset_2014.xlsx Download

Perform dependent t-tests based on the following research questions. Think critically about whether you should be using a one-tailed or two-tailed hypothesis and check your assumptions so you know which test to use!

To get the most out of these exercises, try to first find out the answer on your own and then use the drop-down menus to check your answer.

Note: Technically, none of our data is suitable for a dependent t-test in this dataset. We will pretend that the four test score variables (`English`, `Reading`, `Math`, and `Writing`) are really four measurements of the same underlying test. In reality, we would analyze this data using correlation.

1. **Do students perform better on the English test than they do the Writing test?**
    - Should you use a one-tailed or two-tailed hypothesis? one-tailed two-tailed

- Which statistic should you use based on your assumptions? dependent t-test Wilcoxon rank
- Do students perform better on the English test than they do the Writing test? yes no

2. **Does students' English scores relate to their Reading scores?**
   - Should you use a one-tailed or two-tailed hypothesis? one-tailed two-tailed
   - Which statistic should you use based on your assumptions? dependent t-test Wilcoxon rank
   - Does students' English scores relate to their Reading scores? yes no

Chapter 8

Chi-Square

The chi-quare is a categorical data analysis which is simply data analysis with categorical data. It's usually 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^2test of independence that we use when our frequencies are small.
    2. **McNemar's test**: This is an alternative to the χ^2test of independence that we have a 2x2 repeated-measures design. For example, perhaps we examine pass/fail rates before and after a training.

Notice how we don't have an assumption about a normal distribution. For that reason, these are all *non-parametric statistics*.

8.1 Chi-Square Goodness-of-Fit

8.1.1 Overview

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 chi-square goodness-of-fit test is as follows:

- H_0: The observed frequencies match the expected frequencies.

- H_1: At least one observed frequency doesn't match the expected frequency.

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.

8.1.2 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.

- H_0: Participants pull cards randomly from the deck. In other words, there is a 25% probability of pulling each hearts, clubs, spades, and diamonds.
- H_1: Participants pull cards not at random from the deck. In other words, participants have a different probability than 25% of pulling at least one of the types of cards.

8.1.2.1 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.

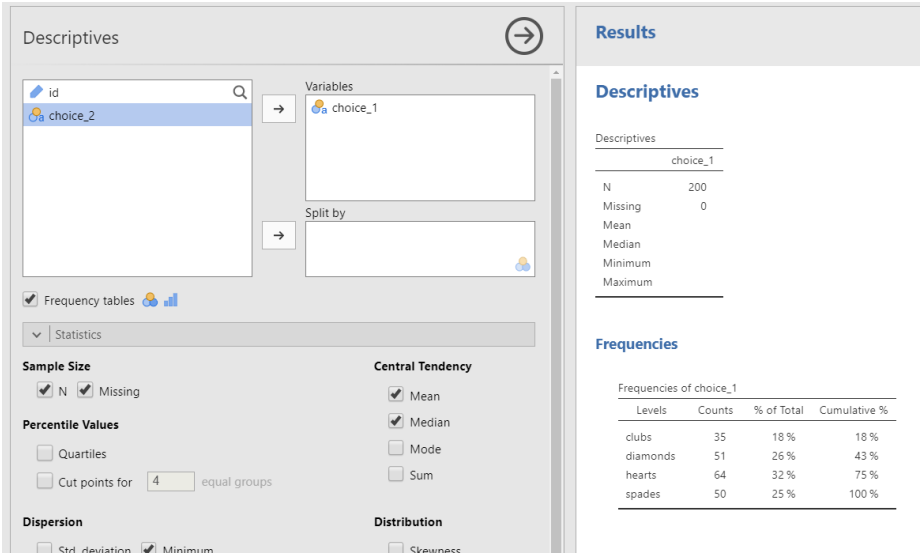
	id	choice_1
1	subj1	spades
2	subj2	diamonds
3	subj3	hearts
4	subj4	spades
5	subj5	hearts
6	subj6	clubs
7	subj7	hearts
8	subj8	diamonds
9	subj9	spades
10	subj10	diamonds

8.1.2.2 Describe your data

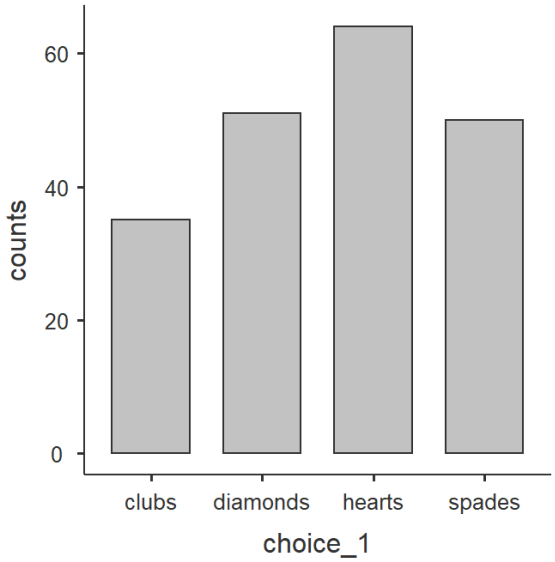
Once we confirm our data is setup correctly in jamovi, we should look at our data using descriptive statistics and graphs. First, our descriptive statistics

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.

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%).



A bar plot can visualize these descriptive statistics nicely.



8.1.3 Check assumptions

The chi-square goodness-of-fit test has the following assumptions:

1. **Expected frequencies are sufficiently large**, which is usually greater than 5. If you violate this assumption, you can use Fisher's exact test.

8.1.4 Perform the test

We'll be examining

1. From the 'Analyses' toolbar select 'Frequences' - 'One sample proportion tests - N outcomes'.
2. Move `choice_1` into the Variable box.
3. Select `Expected counts`.

When you are done, your setup should look like this

8.1.5 Interpret results

The first table shows us our observed frequencies (our data) and expected frequencies (N/k = 200/4 = 50). 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.

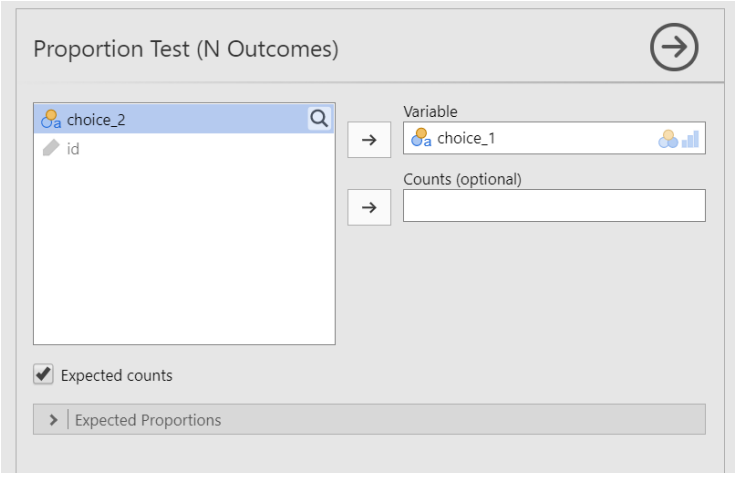


Figure 8.1: Chi-square goodness-of-fit setup in jamovi

8.1.5.1 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, 51selected 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 (χ^2 (3) = 8.44; p = .038), suggesting that people did not select suits purely at random.

8.1.5.2 Visualize the results

The bar chart from above does a decent job of visualizing the results. Although it would be difficult to do in jamovi, we could do a revised chart in Excel pretty easily. Instead of counts, perhaps we care more about percentages, and adding a line for the expected frequency (25%). Here's an example, using instructions from this tutorial:

Proportion Test (N Outcomes)

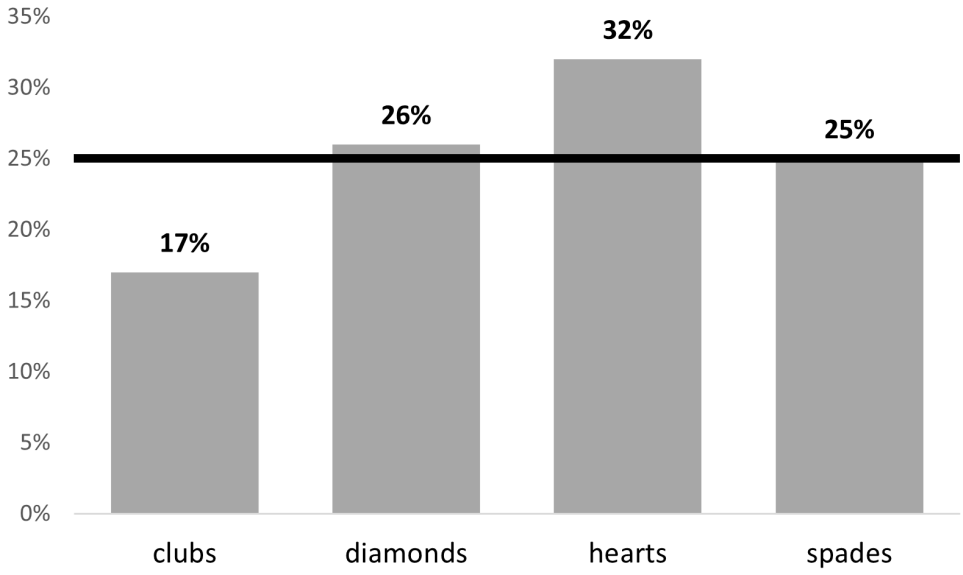
Proportions - choice_1

Level		Count	Proportion
clubs	Observed	35	0.17
	Expected	50.00	0.25
diamonds	Observed	51	0.26
	Expected	50.00	0.25
hearts	Observed	64	0.32
	Expected	50.00	0.25
spades	Observed	50	0.25
	Expected	50.00	0.25

χ^2 Goodness of Fit

χ^2	df	p
8.44	3	0.038

Figure 8.2: Chi-square goodness-of-fit results in jamovi



8.1.6 Additional information

8.1.6.1 Different Expected Frequencies

As you can tell, jamovi automatically assumed equal proportions of frequencies. However, perhaps we think our deck is loaded or we have the actual population frequencies and want to see if our distribution matches the population distribution. We can use the `Expected Proportions` in the setup to specify different expected frequencies.

For example, maybe we think our deck is a little stacked in favor of red cards–or we think our participants are more likely to choose red cards than black cards. We can specify our expected proportions and then interpret the results. In this case, participants do not seem more likely to choose red cards based on the expected frequencies we provided.

8.1.7 Your turn!

Open the `Sample_Dataset_2014.xlsx` file that we will be using for all Your Turn exercises. You can find the dataset here: Sample_Dataset_2014.xlsx Download

To get the most out of these exercises, try to first find out the answer on your own and then use the drop-down menus to check your answer.

1. **Are there equal numbers of athletes and non-athletes?** (`Athlete` variable)

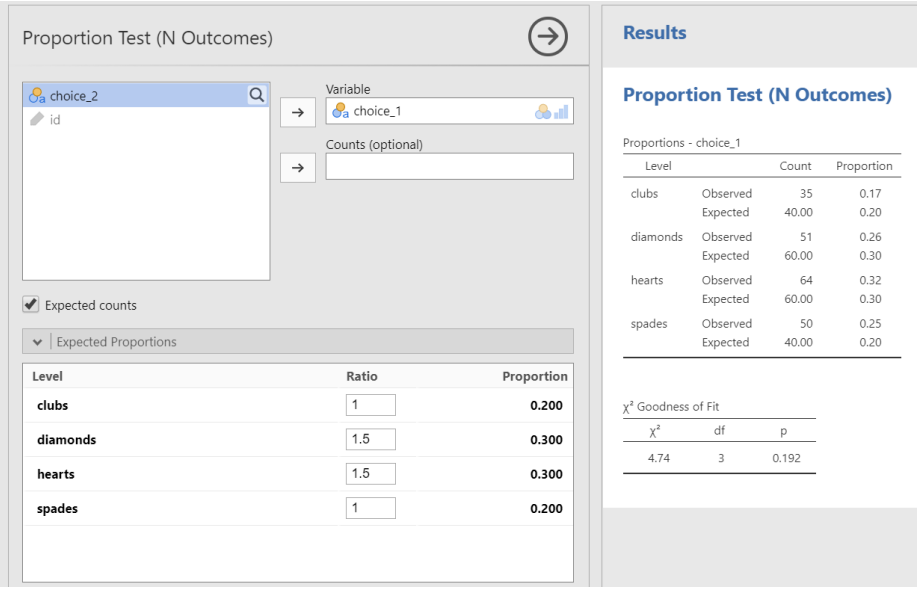


Figure 8.3: Chi-square goodness-of-fit - Different expected proportions

- Do you meet the assumptions? yes no, expected frequencies are too small no, data are not independent
- Are the observed frequencies similar to the expected frequencies? yes no
- What is your chi-square value, rounded to two decimal places:

2. **I happen to know the school this data comes from has 40% athletes and 60% non-athletes. Does our data match the school population?**
   - Change your Expected Proportions ratio to .6 for non-athletes and .4 for athletes.
   - Are the observed frequencies similar to the expected frequencies? yes no
   - What is your chi-square value, rounded to two decimal places:
3. **Are there equal numbers of freshmen, sophomores, juniors, and seniors?** (`Rank` variable)
   - Do you meet the assumptions? yes no, expected frequencies are too small no, data are not independent
   - Are the observed frequencies similar to the expected frequencies? yes no

- What is your chi-square value, rounded to two decimal places:

8.2 Chi-Square Test of Independence

8.2.1 Overview

The χ^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. Our hypotheses for the chi-square goodness-of-fit test is as follows:

- H_0: The observed frequencies match the expected frequencies.
- H_1: At least one observed frequency doesn't match the expected frequency.

8.2.2 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 natives ask whether the visitor prefers puppies, flowers, or large, properly formatted data files.

- H_0: Humans and robots have similar preferences.
- H_1: Humans and robots have different preferences.

8.2.2.1 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. Here's our data for our example we'll be working with, which you can find in the lsj-data called `chapek9`:

ID	species	choice
1	robot	flower
2	human	data
3	human	data
4	human	data
5	robot	data
6	human	flower
7	human	data

ID	species	choice
8	robot	data
9	human	puppy
10	robot	flower

8.2.2.2 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. Remember that for nominal variables we should report frequency statistics, not means and medians and such. Bar plots continue to be a good way of visualizing the data.

Frequencies

Frequencies of species

Levels	Counts	% of Total	Cumulative %
robot	87	48 %	48 %
human	93	52 %	100 %

Frequencies of choice

Levels	Counts	% of Total	Cumulative %
puppy	28	16 %	16 %
flower	43	24 %	39 %
data	109	61 %	100 %

8.2.2.3 Check assumptions

The chi-square test of independence has the following assumptions:

1. **Expected frequencies are sufficiently large**, which is usually greater than 5. If you violate this assumption, you can use Fisher's exact test.

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.

8.2.3 Perform the test

1. From the 'Analyses' toolbar select 'Frequencies' - 'Independent Samples - χ^2 test of association'.
2. 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.
4. Optionally, you can request under the Cells tab to show the expected counts and the row, column, and total percentages.

When you are done, your setup should look like this

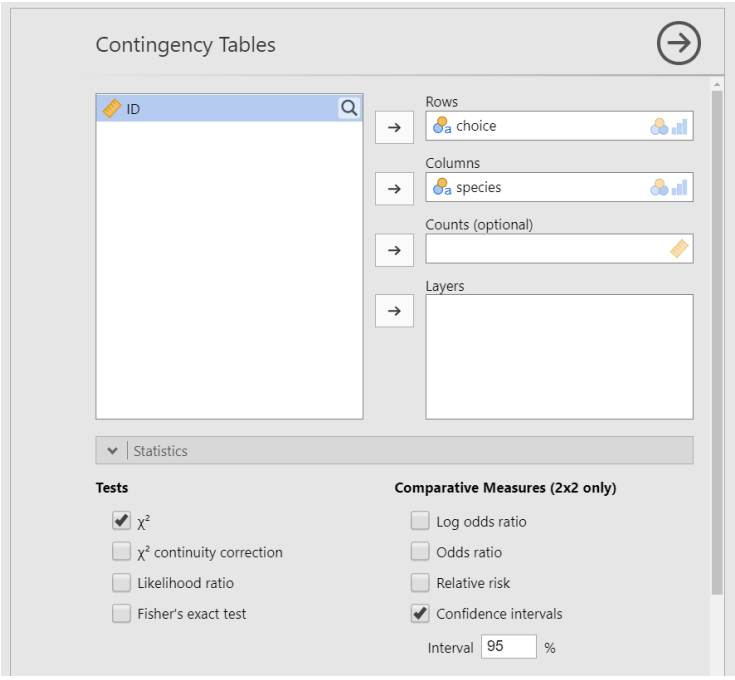


Figure 8.4: Chi-square test of independence setup in jamovi

8.2.4 Interpreting results

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. 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., 2x2 or 3x3).

8.2.4.1 Write up the results in APA style

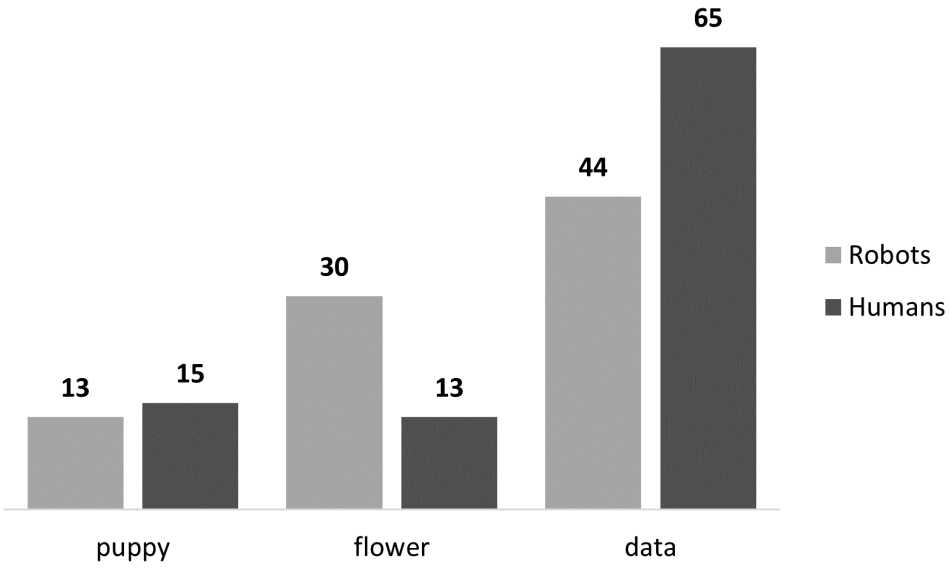
We can write up our results in APA something like this:

> Pearson's χ^2 test of independence showed a significant association between species and choice, χ^2 (2) = 10.72, p = .005, Cramer's V = .24. Robots appeared to be more likely to say they prefer flowers and humans appeared to be more likely to say they prefer data. Robots and humans appeared to be equally likely to prefer puppies.

I would either write-up the observed frequencies above or, ideally, I would share the contingency table with my observed frequencies.

8.2.4.2 Visualize the results

This one is also a visualization that would likely do better in Excel than in jamovi. There are two that I think work well for this dataset and our research questions. The first is a clustered column chart:



The second is a stacked bar chart with connected lines:

Contingency Tables

Contingency Tables

	species		
choice	robot	human	Total
puppy	13	15	28
flower	30	13	43
data	44	65	109
Total	87	93	180

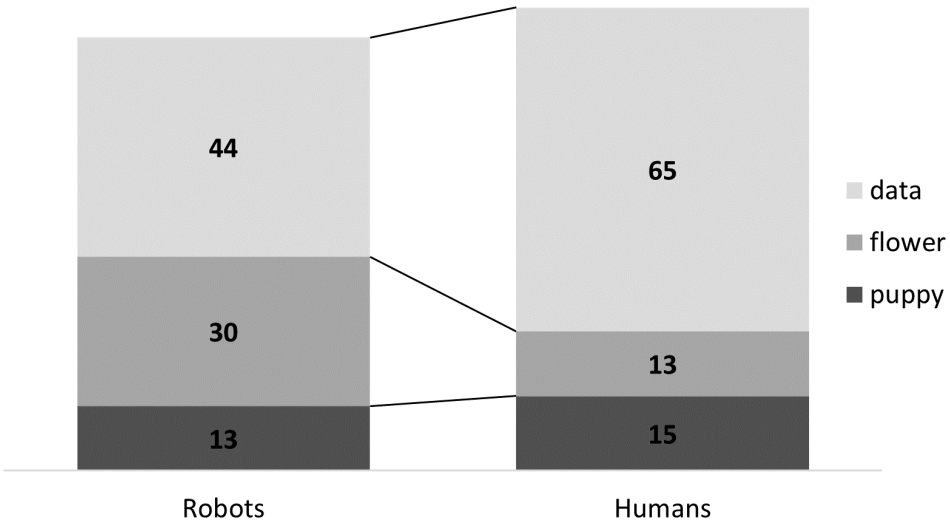
χ^2 Tests

	Value	df	p
χ^2	10.72	2	0.005
N	180		

Nominal

	Value
Phi-coefficient	NaN
Cramer's V	0.24

Figure 8.5: Chi-square test of independence results in jamovi



8.2.5 In case of violated assumptions

8.2.5.1 Fisher's exact test

If you violate the assumption that there your expected frequencies are sufficiently large and you have a 2x2 table, you can still perform the χ^2 test of independence but instead of selecting χ^2you'll select `Fisher's exact test`. You'll interpret your results exactly the same but specify you used the Fisher's exact test.

8.2.6 Additional information

8.2.6.1 Ordinal variable(s)

If either of your variables are ordinal, instead of selecting `Phi and Cramer's V` you should select `Gamma` or `Kendall's tau-b`. Which do you choose? `Kendall's tau-b` should only be chosen if you have a square table (e.g., 3x3, 4x4, 5x5) whereas `Gamma` can be done with any size table. `Kendall's tau-b` will be a slightly more conservative estimate compared to `Gamma`.

8.2.7 Your turn!

Open the `Sample_Dataset_2014.xlsx` file that we will be using for all Your Turn exercises. You can find the dataset here: Sample_Dataset_2014.xlsx Download

To get the most out of these exercises, try to first find out the answer on your own and then use the drop-down menus to check your answer.

1. **Is Athlete related to Gender?**
    - Do you meet the assumptions? yes no, expected frequencies are too small no, data are not independent
    - Which test should you perform? Chi-square Fisher's exact test
    - Are the observed frequencies similar to the expected frequencies? yes no
    - What is your chi-square value, rounded to two decimal places:
2. **Is Gender related to Rank?**
    - Do you meet the assumptions? yes no, expected frequencies are too small no, data are not independent
    - Which test should you perform? Chi-square Fisher's exact test
    - Are the observed frequencies similar to the expected frequencies? yes no
    - What is your chi-square value, rounded to two decimal places:

8.3 McNemar's Test

8.3.1 Overview

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. Our hypotheses for the McNemar test is as follows:

- H_0: The observed frequencies match the expected frequencies.
- H_1: At least one observed frequency doesn't match the expected frequency.

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.

8.3.2 Look at the data

8.3.2.1 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. Here's our data for our example we'll be working with, which you can find in the lsj-data called `agpp`:

ID	response_before	response_after
subj.1	no	yes
subj.2	yes	no
subj.3	yes	no
subj.4	yes	no
subj.5	no	no
subj.6	no	no
subj.7	no	no
subj.8	no	yes
subj.9	no	no
subj.10	no	no

8.3.3 Perform the test

Let's run an example with data 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.

1. From the 'Analyses' toolbar select 'Frequencies' - '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.

When you are done, your setup should look like this

8.3.4 Interpret results

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.

8.3.4.1 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, χ^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

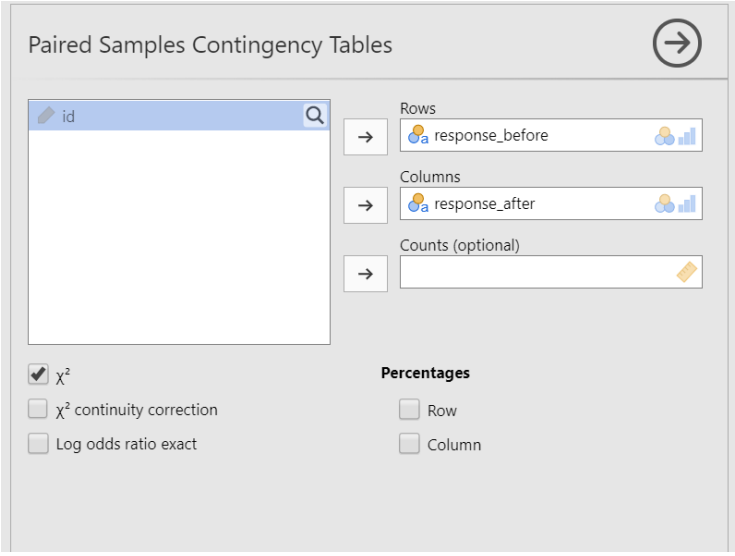


Figure 8.6: McNemar's test setup in jamovi

= 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.

Paired Samples Contingency Tables

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	13.33	1	< .001
N	100		

Figure 8.7: McNemar's test results in jamovi

Chapter 9

ANOVA

ANOVA stands for **AN**alysis **O**f **VA**riance. ANOVAs analyze the variation between and within groups. There are multiple types of ANOVAs based on the nature of the independent variable(s):

Nature of IV(s)	**Type of ANOVA**
1 between-subjects IV	One-way ANOVA
1 within-subjects IV	Repeated-measures ANOVA
2+ between-subjects IVs	Independent factorial ANOVA
2+ within-subjects IVs	Repeated measures factorial ANOVA
2+ IVs mixed between/within	Mixed factorial ANOVA
1+ IV with a continuous covariate	ANCOVA

Furthermore, there are ANOVAs for when there are multiple dependent variables (called the MANOVA or **M**ultiple **ANOVA**, as well as the MANCOVA or **M**ultiple **ANCOVA**) but we will not discuss them in this class.

9.1 One-way ANOVA

9.1.1 Overview

The one-way analysis of variance (ANOVA) is used to test the difference in our dependent variable between three or more different groups of observations. Our grouping variable is our independent variable. In other words, we use the one-way ANOVA when we have a research question with a **continuous dependent variable** and a **categorical independent variable with three or more categories in which different participants are in each category**.

The one-way ANOVA is also known as an independent factor ANOVA.

One thing to keep in mind is the one-way ANOVA is an omnibus statistic that tests against the null hypothesis that the three or more means are the same. It does not tell us where the mean differences are (e.g., that $1 > 2$); for that, we need planned contrasts or post-hoc procedures, which you'll learn about later. Therefore, the null and alternative hypotheses for the one-way ANOVA are as follows:

- H_0: There is **no difference** in means between the groups. In other words, the means for the three or more groups are the **same**.
- H_1: There is **a difference** in means between the groups. In other words, the means for the three or more groups are **different**.

9.1.1.1 Why not multiple t-tests?

In the example above, we have three groups: fall, spring, and summer. We could 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 to reduce our Type I error rate. If I had performed three separate t-tests, set my alpha (Type I error rate) at 5%, and knew for a fact the effects do not actually exist, then each test has a probability of a Type I error rate of 5%. Because we are running three tests, our alpha rate 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 would be statistically significant by chance alone and would be a false positive. 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 I'll get into later.

9.1.2 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.

9.1.2.1 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).

	ID	drug	therapy	mood.gain
1	1	placebo	no.therapy	0.5
2	2	placebo	no.therapy	0.3
3	3	placebo	no.therapy	0.1
4	4	anxifree	no.therapy	0.6
5	5	anxifree	no.therapy	0.4
6	6	anxifree	no.therapy	0.2
7	7	joyzepam	no.therapy	1.4
8	8	joyzepam	no.therapy	1.7
9	9	joyzepam	no.therapy	1.3
10	10	placebo	CBT	0.6
11	11	placebo	CBT	0.9
12	12	placebo	CBT	0.3
13	13	anxifree	CBT	1.1
14	14	anxifree	CBT	0.8
15	15	anxifree	CBT	1.2
16	16	joyzepam	CBT	1.8
17	17	joyzepam	CBT	1.3
18	18	joyzepam	CBT	1.4

9.1.2.2 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

Descriptives

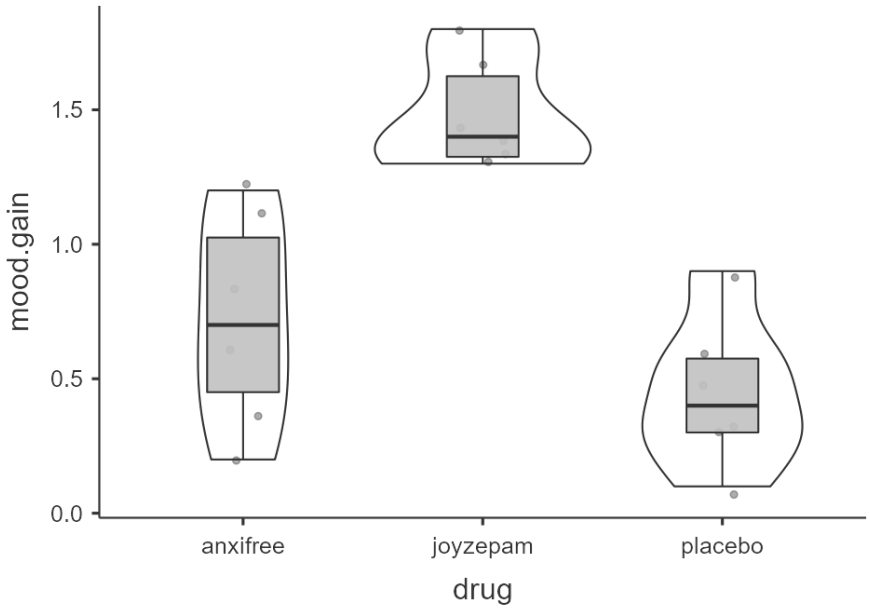
	drug	mood.gain
N	18	18
Missing	0	0
Mean		0.88
Median		0.85
Standard deviation		0.53
Minimum		0.10
Maximum		1.80

Frequencies

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!



9.1.3 Check assumptions

9.1.3.1 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.

9.1.3.2 ANOVA is robust to violations

Although we should meet the assumptions as much as possible, 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 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 small (such as 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!

9.1.3.3 Checking assumptions

We test for normality using the Shapiro-Wilk test and the Q-Q plot. The Shapiro-Wilk test was not statistically significant (W = .96, $p = .605$); therefore, this indicates the data is normally distributed. Furthermore, the lines are fairly close to the diagonal line in the Q-Q plot. We can conclude that we satisfy the assumption of normality.

We test for homogeneity of variance using the Levene's test. The Levene's test was not statistically significant (F [2, 15] = 1.45, $p = .266$); therefore, this indicates our data satisfies the assumption of homogeneity of variance. However, I would add a caveat that we have a small sample of data ($N = 18$); we should have tried to collect more data.

Assumption Checks

Normality Test (Shapiro-Wilk)

	W	p
mood.gain	0.96	0.605

Note. A low p-value suggests a violation of the assumption of normality

Homogeneity of Variances Test (Levene's)

	F	df1	df2	p
mood.gain	1.45	2	15	0.266

[3]

Figure 9.1: Testing assumptions in jamovi

9.1.4 Perform the test

Note: Do not use the one-way ANOVA analysis in jamovi! The options there are too limited for our use. Instead, be sure you use ANOVA!

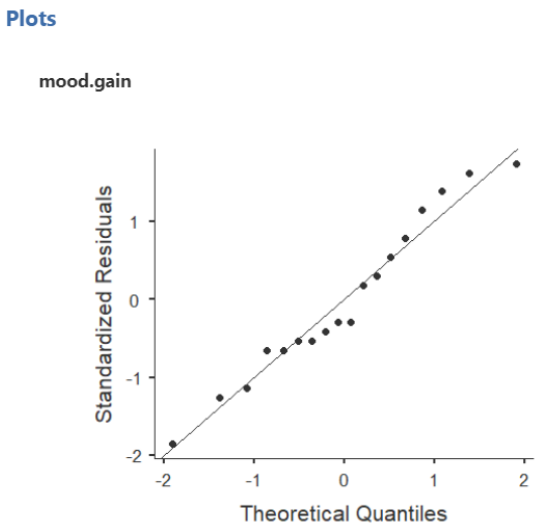


Figure 9.2: Testing assumptions in jamovi

1. To perform a one-way ANOVA in jamovi, go to the Analyses tab, click the **ANOVA** button, and choose "ANOVA". You might be asking why we aren't choosing "One-Way ANOVA" and that's because the options there are too limited.

2. Move your dependent variable `mood.gain` to the Dependent Variable box and your independent variable `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. Ignore the Contrasts and Post Hoc Tests drop-down menus for now. See below for more information on them.

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`.

When you are done, your setup should look like this:

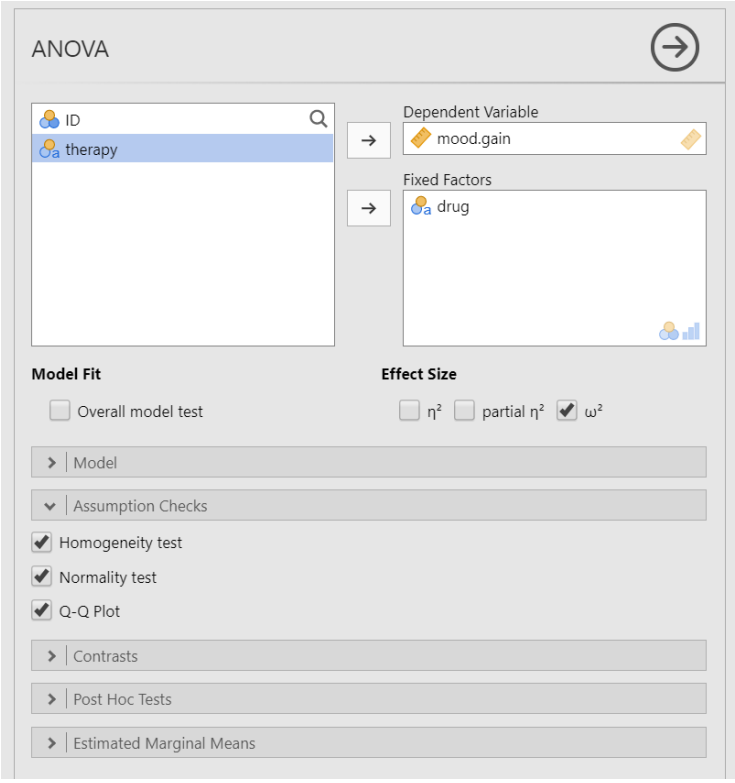


Figure 9.3: One-way ANOVA setup in jamovi

9.1.5 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	1.73	18.61	< .001	0.66
Residuals	1.39	15	0.09			

[3]

Figure 9.4: One-way ANOVA results in jamovi

Our p-value is less than .05, so our results are statistically significant.

9.1.5.1 Write up the results in APA style

We can write up our results in APA something like this:

> There is a significant difference in mood gain across the three drug conditions, F (2, 15) = 18.61, $p < .001$, $\omega^2 = .66$.

Sometimes, people like to put the statistics inside a parentheses. In that case, you need to change the parentheses around the degrees of freedom as brackets. Here's another example write-up of the results in APA style:

> There is a significant difference in mood gain across the three drug conditions (F [2, 15] = 18.61, $p < .001$, $\omega^2 = .66$).

9.1.5.2 Visualize the results

You should visualize the results similarly to how you visualize the results for the independent samples t-test. The default graph in the estimated marginal means output for the ANOVA is not great in my opinion. Presenting the graph of the data in this case (see the graph under Look at the Data) is probably a better option.

9.1.6 In case of violated assumptions

The great news is that jamovi includes the Welch's F-statistic and the Kruskal-Wallis non-parametric test! The bad news is that you lose some functionality in jamovi when you use them. Just like with the Welch's t-statistic (for the independent t-test), it does not have the assumption of equal variances so it's

appropriate to use if your data is normally distributed but does not have homogeneous variances. Similarly, the Kruskal-Wallis test is the non-parametric version of the one-way ANOVA and should be used if you do not satisfy the assumption of normality.

Here's what statistic you should choose based on satisfying assumptions:

	Normality: satisfied	**Normality: not satisfied**
Homogeneity of Variance: satisfied	one-way ANOVA	Kruskal-Wallis
Homogeneity of Variance: not satisfied	Welch's F-test	Kruskal-Wallis

9.1.6.1 Welch's F-test

To conduct this in jamovi, you will need to use the "One-Way ANOVA" test, not the "ANOVA" test. The unfortunate thing about this test is that it strangely does not provide effect sizes.

In jamovi, under Variances select `Don't assume equal (Welch's)`. Move `mood.gain` to the Dependent Variable box and `drug` to your Grouping Variable box. You will interpret the results similarly to the one-way ANOVA:

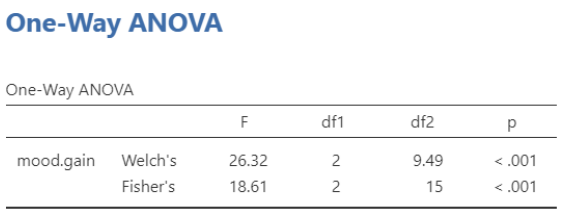


Figure 9.5: One-way ANOVA results in jamovi

> Using a Welch's F-test, there is a significant difference in mood gain across the three drug conditions, F (2, 9.49) = 26.32, $p < .001$.

9.1.6.2 Kruskal-Wallis test

To perform the Kruskal-Wallis test in jamovi, you will need to select under the ANOVA button "One-Way ANOVA, Kruskal Wallis" towards the bottom of the list of options. Move `mood.gain` to the Dependent Variables box and `drug` to the Grouping Variable box. Select Effect size; if you need post hoc comparisons select DSCF pairwise comparisons (see section below on group differences). You will interpret the results similarly to the one-way ANOVA:

> Using a Kruskal-Wallis test, there is a significant difference in mood gain across the three drug conditions, χ^2 (2) = 12.08, p = .002, ϵ^2 = .71.

One-Way ANOVA (Non-parametric)

Kruskal-Wallis

	χ^2	df	p	ε^2
mood.gain	12.08	2	0.002	0.71

Figure 9.6: Kruskal-Wallis results in jamovi

Notice how in this case all three results converge and show there is a statistically significant difference in the results! The problem is… differences in which groups?

9.1.7 Additional information

9.1.7.1 Relationship between ANOVA and t-test

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.

$F = t^2$

$t = \sqrt{F}$

Just a fun little bit of trivia! So if you accidentally do an F-test with two groups, no need to go back and redo the analyses (although you should if you are sharing your code for reproducibility). You can just convert your F to a t statistic easily!

9.1.7.2 A note on one-tailed vs. two-tailed tests

Unlike a t-test, we would not halve the p-value in an F-ratio because it is an omnibus test. Our planned contrasts or post-hoc tests can tell us where differences are, and we can provide directional hypotheses there if we so choose.

9.1.8 Your turn!

Open the `Sample_Dataset_2014.xlsx` file that we will be using for all Your Turn exercises. You can find the dataset here: Sample_Dataset_2014.xlsx Download

Perform one-way ANOVAs based on the following research questions. Check your assumptions and ensure you are using the correct tests.

To get the most out of these exercises, try to first find out the answer on your own and then use the drop-down menus to check your answer.

1. **Does students differ on English scores by rank (i.e., freshmen, sophomore, junior, senior)?**
    - Do you satisfy the assumption of normality? yes no
    - Do you satisfy the assumption of homogeneity of variance? yes no
    - Which statistic should you use? one-way ANOVA Welch's F-test Kruskal-Wallis test
    - Do students differ on English scores by rank? yes no
2. **Does smoking status (Smoking: Nonsmoker = 0, Past smoker = 1, Current smoker = 2) relate to sprint time?**
    - Do you satisfy the assumption of normality? yes no
    - Do you satisfy the assumption of homogeneity of variance? yes no
    - Which statistic should you use? one-way ANOVA Welch's F-test Kruskal-Wallis test
    - Does smoking status relate to sprint time? yes no

9.2 Finding Group Differences

Often, we're not interested in just *whether* there is a difference (which the F-statistic can tell us), but *where* the differences are between groups (which the F-statistic cannot tell us). For that, we use either planned contrasts when you have specific hypotheses you want to test or post-hoc comparisons when you have no specific hypotheses.

Note: You do not perform contrasts or post hoc comparisons if your overall F statistic is not statistically significant. You do not interpret group differences if you fail to reject the null hypothesis that there are no group differences!

9.2.1 Planned Contrasts

If you have before-analysis hypotheses of group differences in your data, you will use planned contrasts. You can find the planned contrasts in the ANOVA (but not the one-way ANOVA) setup as a drop-down menu. Note that while I show all six contrasts that jamovi provides, you do not normally do multiple contrasts. These are just shown for illustrative purposes:

1. **Deviation**: compares the effect of each category (except the first category) to the overall experimental effect. The order of categories is alphabetical or numerical order. Notice how anxifree is considered the first category.

Contrasts

Contrasts - drug

	Estimate	SE	t	p
joyzepam - anxifree, joyzepam, placebo	0.60	0.10	5.91	< .001
placebo - anxifree, joyzepam, placebo	−0.43	0.10	−4.27	< .001

Figure 9.7: Contrasts - Deviation

2. **Simple**: Each category is compared to the first category. The order of categories is alphabetical or numerical order. Notice how anxifree is considered the first category.

Contrasts

Contrasts - drug

	Estimate	SE	t	p
joyzepam - anxifree	0.77	0.18	4.36	< .001
placebo - anxifree	−0.27	0.18	−1.52	0.150

Figure 9.8: Contrasts - Simple

3. **Difference**: Each category (except the first) is compared to the mean effect of all previous categories.

4. **Helmert**: Each category (except the last) is compared to the mean effect of all subsequent categories.

5. **Repeated**: Each category is compared to the last category.

6. **Polynomial**: Tests trends in the data. It will examine the n-1^{th} degree based on the number of groups. In this case, because we have 3 groups it is testing both a linear (^{1}) and quadratic (^{2}) trend. If we had 5 groups, it would test a linear (^{1}), quadratic (^{2}), cubic (^{3}), and quartic (^{4}) trend. Note that your factor levels must be ordinal for a polynomial contrast to make sense.

Test yourself! Which contrast would make most sense to test given that we want to know how our drug compares to the other two drugs? deviation simple difference helmert repeated polynomial

Contrasts

Contrasts - drug

	Estimate	SE	t	p
joyzepam - anxifree	0.77	0.18	4.36	< .001
placebo - anxifree, joyzepam	−0.65	0.15	−4.27	< .001

Figure 9.9: Contrasts - Difference

Contrasts

Contrasts - drug

	Estimate	SE	t	p
anxifree - joyzepam, placebo	−0.25	0.15	−1.64	0.121
joyzepam - placebo	1.03	0.18	5.88	< .001

Figure 9.10: Contrasts - Helmert

Contrasts

Contrasts - drug

	Estimate	SE	t	p
anxifree - joyzepam	−0.77	0.18	−4.36	< .001
joyzepam - placebo	1.03	0.18	5.88	< .001

Figure 9.11: Contrasts - Repeated

Contrasts

Contrasts - drug

	Estimate	SE	t	p
linear	−0.19	0.12	−1.52	0.150
quadratic	−0.73	0.12	−5.91	< .001

Figure 9.12: Contrasts - Polynomial

9.2.1.1 Write up planned contrasts in APA style

Here's some example write-ups of the above results.

> There is a significant difference in mood gain across the three drug conditions, F (2, 15) = 18.61, $p < .001$. Repeated contrasts showed that *Joyzepam* ($M = 1.48$, $SD = .21$) outperformed both *Anxifree* ($M = .72$, $SD = .39$; $p < .001$) and the placebo condition ($M = .45$, $SD = .28$; $p < .001$).

> (Note how this example makes no sense because our data is not ordinal) There is a significant difference in mood gain across the three drug conditions, F (2, 15) = 18.61, $p < .001$. There was not a significant linear trend across the drug conditions ($p = .150$).

9.2.2 Post hoc comparisons

Often, we do not have any *a priori* (or planned) predictions or hypotheses about our group differences. In this case, we use post hoc procedures. These procedures do pairwise comparisons among all of our groups, like t-tests across each of our groups. As we noted on the first page of this handout, this can be highly problematic for our Type I error rate! Therefore, we must perform corrections to control our familywise error rate.

jamovi currently supports five types of post-hoc tests; I generally only use Tukey or Bonferonni:

1. **No correction**: This doesn't correct for a Type I error at all. Don't use this! I won't even show it. It's bad. Never use it. NEVER. You are warned!
2. **Tukey**: This is the post hoc test I use most often. It controls the Type I error rate well, but isn't as conservative of a control as the Bonferonni.
3. **Scheffe**: Honestly, I've never used it. I am not sure how it's calculated.
4. **Bonferroni**: This is the most conservative test. It's good if you only have a small number of comparisons to make or if you *really* want to control your Type I error rate. If you have a lot of them to test , then you should use something else.
5. **Holm**: Honestly, I've never used it. I am not sure how it's calculated.

Games-Howell for when you have unequal variances and Tukey for when you have equal variances. They will each calculate your p-values slightly differently but in a way to control for our Type I error rate as best it can. They are interpreted very similarly, so we will proceed with the Tukey's post hoc comparisons because we satisfied the assumption of equal variances.

To request post hoc tests from the one-way ANOVA, open the collapsed menu at the bottom of the setup menu. Select `Tukey (equal variances` under Post-Hoc Test and select `Mean difference`, `Report significance`, and `Flag`

`significant comparisons` under Statistics. Optionally, you can request the `Test results (t and df)` although this is not necessary.

Below shows the post hoc test results for our one-way ANOVA. Notice the differences in p-values across the four post hoc tests and how all other values are the same. Notice how the Bonferroni is most conservative (i.e., has the largest p-values) and the Holm's is the least conservative (i.e., has the smallest p-values). Keep in mind you do not normally ask for multiple post hoc comparisons. Just pick one! Normally, I just pick Tukey's.

Post Hoc Tests

Post Hoc Comparisons - drug

Comparison			Mean Difference	SE	df	t	Ptukey	Pscheffe	Pbonferroni	Pholm	Cohen's d
drug		drug									
anxifree	-	joyzepam	−0.77	0.18	15.00	−4.36	0.002	0.002	0.002	0.001	−2.52
	-	placebo	0.27	0.18	15.00	1.52	0.312	0.343	0.451	0.150	0.88
joyzepam	-	placebo	1.03	0.18	15.00	5.88	< .001	< .001	< .001	< .001	3.39

Note. Comparisons are based on estimated marginal means

Figure 9.13: Post hoc test results in jamovi

9.2.2.1 Write up post hoc results in APA style

The way we would write up each of the post hoc comparisons is very similar. Given that I usually use Tukey, here is a write-up for those results:

> There is a significant difference in mood gain across the three drug conditions, $F\ (2, 15) = 18.61$, $p < .001$. Post hoc comparisons using Tukey's HSD revealed that our drug *Joyzepam* ($M = 1.48$, $SD = .21$) outperformed both *Anxifree* ($M = .72$, $SD = .39$; $p = .002$) and the placebo condition ($M = .45$, $SD = .28$; $p < .001$); there were no differences between *Anxifree* and the placebo condition ($p = .312$).

Writing up results in APA style is both a science and an art. There's a science to what you need to report. For example, you always report the statistics exactly the same: $F\ (\text{df}_{\text{WG}}, \text{df}_{\text{BG}}) = \text{X.XX}$, $p = .\text{XXX}$. You also always report the group means (M) and standard deviations (SD), although you can report them in-text like I did above or in a descriptives table like you can ask from jamovi.

However, there's also an art to it. Notice how I wrote that up in a way to summarize the findings as succinctly as possible. I could have said there was a difference between *anxifree* and *joyzepam* and a difference between *joyzepam* and the placebo, but that's a lot more words and isn't written in a way to focus on what I'm hoping to see: that my drug *joyzepam* performed better than the competitor *anxifree* and a placebo condition.

This is where you need to think creatively and be very critical in checking that what you say makes sense. Read your write-ups carefully! Have someone else

read it. Can they understand what you mean?

9.2.3 In case of violated assumptions

If you are using Welch's F-test using the One-Way ANOVA in jamovi, you should select under Post-Hoc Tests `Games-Howell (unequal variances)`. These will be interpreted similarly to the post hoc comparisons above.

If you are using the Kruskal-Wallis test, you will select the check-box for `DSCF pairwise comparisons`. This stands for the Dwass-Steel-Critchlow-Fligner test. All you need to know is that they, too, are interpreted similarly to the post hoc comparisons above.

Unfortunately, you cannot perform contrasts with either the Welch's F-test or Kruskal-Wallis test.

9.2.4 Your turn!

Open the `Sample_Dataset_2014.xlsx` file that we will be using for all Your Turn exercises. You can find the dataset here: Sample_Dataset_2014.xlsx Download

Perform one-way ANOVAs based on the following research questions. Check your assumptions and ensure you are using the correct tests.

To get the most out of these exercises, try to first find out the answer on your own and then use the drop-down menus to check your answer.

Note: These are the same questions as in the one-way ANOVA chapter, but now you focus on the group differences.

1. **Does students differ on English scores by rank (i.e., freshmen, sophomore, junior, senior)?**
    - Should you perform a planned contrast or post hoc comparison? planned contrast post hoc
    - What are the results of the post hoc comparison? N/A - Don't perform Freshmen had higher English scores than sophomores, juniors, and seniors Freshmen and sophomores had higher English scores than juniors and seniors
2. **Does smoking status (Smoking: Nonsmoker = 0, Past smoker = 1, Current smoker = 2) relate to sprint time?**
    - Should you perform a planned contrast or post hoc comparison? planned contrast post hoc
    - What are the results of the post hoc comparison? N/A - Don't perform Nonsmokers had significantly faster sprint times than current smokers Nonsmokers and past smokers had significantly faster

spring times than current smokers Nonsmokers had significantly faster sprint times than both past and current smokers

9.3 Repeated Measures ANOVA

9.3.1 Overview

The repeated measures analysis of variance (ANOVA) is used to test the difference in our dependent variable between three or more groups of observations in which all participants participate in all groups or levels. Our grouping variable is our independent variable. In other words, we use the one-way ANOVA when we have a research question with a **continuous dependent variable** and a **categorical independent variable with three or more categories in which the same participants are in each category**.

The repeated measures ANOVA is also sometimes called the one-way related ANOVA.

There are two ways we could have the repeated measures ANOVA. Perhaps the same group of participants are measured in the same dependent variable at three or more time points. In this case, our independent variable is time and our dependent variable is whatever is measured at each time point.

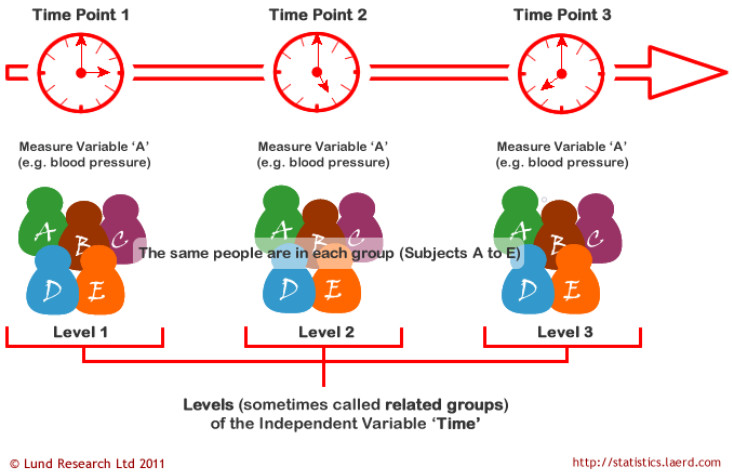


Figure 9.14: Repeated measures ANOVA by Time

The other way we might have the repeated measures ANOVA is if all our participants participate in all conditions of our study. In this case, our independent variable is the treatment or condition and the dependent variable is whatever is measured in each treatment or condition.

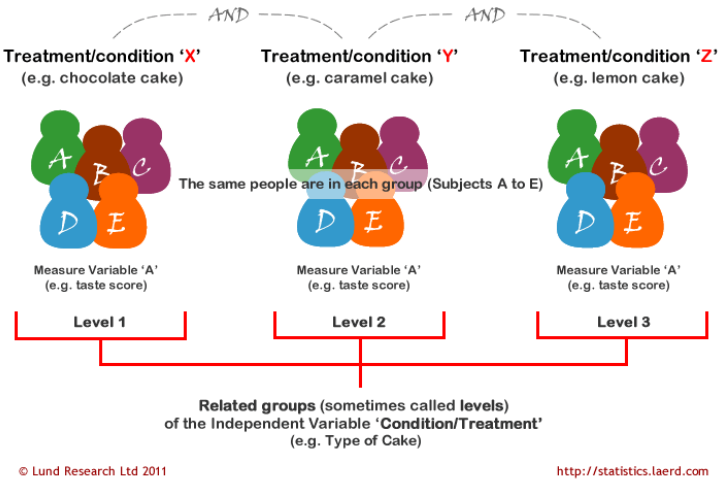


Figure 9.15: Repeated measures ANOVA by Conditions

9.3.2 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".

This dataset is hypothetical data in which six patients suffering from Broca's Aphasia (a language deficit commonly experienced following a stroke) complete 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.

9.3.2.1 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**? Participant Condition Speech Conceptual Syntax TestScore

In the data above, what is your **dependent variable**? Participant Condition Speech Conceptual Syntax TestScore

9.3.2.2 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 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

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

9.3.3 Check Assumptions

9.3.3.1 Assumptions

As a parametric test, the repeated measures 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**); in repeated measures ANOVA this is called the assumption of **sphericity**
3. The dependent variable is **interval or ratio** (i.e., continuous)
4. ~~Scores are **independent** between groups~~ (this assumption is not relevant because all participants participate in all conditions)

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.

9.3.3.1.1 Sphericity Assumption The sphericity assumption is essentially the repeated measures ANOVA equivalent of homogeneity of variances. Spheric-

ity 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.

9.3.3.2 Checking assumptions

You'll notice there are no options to check for normality in the repeated measures ANOVA in jamovi. There's an interesting conversation on the topic in the jamovi forums. Suffice to say, it's complicated and maybe someday they will implement it. For now, we just won't worry about it for the repeated measures ANOVA.

So what we need to worry about is testing our assumption of sphericity. You should have checked 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	0.85	0.720	0.87	1.00

Figure 9.16: Testing sphericity in jamovi

Mauchly's test of sphericity tests the null hypothesis that the variances of the differences between the conditions are equal. Therefore, just like with our previous assumption checks, if Mauchly's test is non-significant (i.e., $p > .05$, as is the case in this analysis) then it is reasonable to conclude that the variances of the differences are not significantly different. This means we satisfy the assumption of sphericity and can conclude that the variances of the differences are roughly equal.

If Mauchly's test had been statistically significant ($p < .05$), then we would conclude that the assumption had *not* been met. In this case, we should apply a correction to the F-value obtained in the repeated measures ANOVA:

- If the Greenhouse-Geisser value in the "Tests of Sphericity" table is $> .75$ then you should use the Huynh-Feldt correction.
- If the Greenhouse-Geisser value is $< .75$, then you should use the Greenhouse-Geisser correction.

You can select these corrections in the Assumption Checks drop-down menu.

9.3.4 Perform the test

1. To perform a repeated measures ANOVA in jamovi, go to the Analyses tab, click the ANOVA button, and choose "Repeated Measures ANOVA".
2. Under "Repeated Measures Factors" name your independent variable. In this case you can name it "Task". Rename the three levels of Task: Speech, Conceptual, and Syntax.
3. Under "Repeated Measures Cells" move the given variable into the correct level. For example, you'll move Speech to the Speech cell.
4. Select Generalised η^2 as your measure of effect size.
5. In the Assumption Checks drop-down menu, select `Sphericity tests`. You'll note that if you violate the assumption of sphericity, there are two corrections provided. These will be discussed later.
6. In the Post Hoc Tests drop-down menu, select `Tukey`. Remember that we only interpret these if the overall F is statistically significant.
7. In the Estimated Marginal Means drop-down menu, move Task to the Marginal Means box, select `Marginal means tables`, and select `Observed scores` . Uncheck `Equal cell weights`.

When you are done, your setup should look like this:

9.3.5 Interpret results

Once we are satisfied we have satisfied the assumptions for the repeated measures, we can interpret our results.

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). Why did it give us a between-subjects table? With the repeated-measures ANOVA (which only has within-subjects IVs), this is just our SS_{BG}. However, because we don't have one, it's not calculating anything. We can ignore it. It is only useful if we are conducting a mixed factorial ANOVA with both between-subjects and within-subjects effects.

Therefore, the Within Subjects Effects table is of most use to us. We can see that the overall effect of Task is statistically significant ($p = .013$). Therefore we can look at our Post Hoc Tests results.

The Tukey post hoc differences show that there was a significant difference between speech and syntax ($p = .011$), but not between conceptual and both speech and syntax. Last, we can look at the Estimated Marginal Means - Task

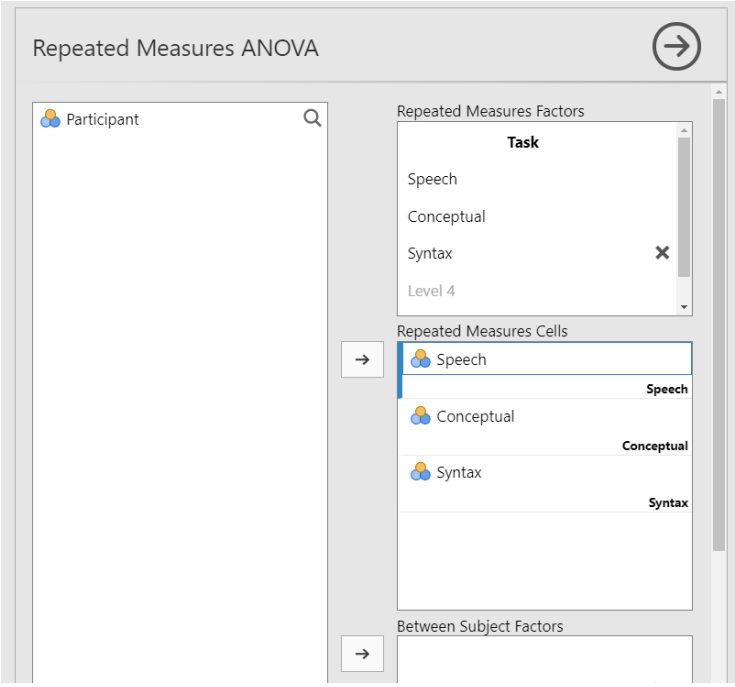


Figure 9.17: Repeated Measures ANOVA setup in jamovi

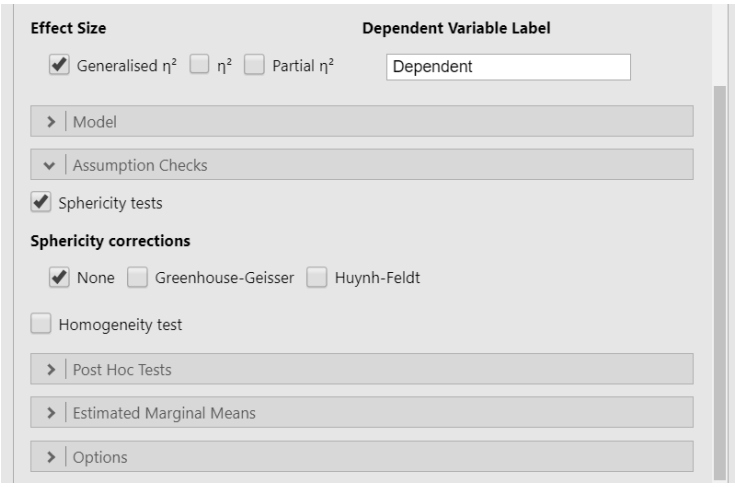


Figure 9.18: Repeated Measures ANOVA setup in jamovi

Repeated Measures ANOVA

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G
Task	24.78	2	12.39	6.93	0.013	0.41
Residual	17.89	10	1.79			

Note. Type 3 Sums of Squares

[3]

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G
Residual	17.11	5	3.42			

Note. Type 3 Sums of Squares

Figure 9.19: One-way ANOVA results in jamovi

Post Hoc Tests

Post Hoc Comparisons - Task

Comparison			Mean Difference	SE	df	t	p_{tukey}
Task		Task					
Speech	-	Conceptual	1.00	0.77	10.00	1.29	0.429
	-	Syntax	2.83	0.77	10.00	3.67	0.011
Conceptual	-	Syntax	1.83	0.77	10.00	2.37	0.091

Figure 9.20: One-way ANOVA results in jamovi

table to see the group means for reporting purposes. This shows us that participants recognized significantly more words in the speech task than in the syntax task.

9.3.5.1 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 significantly affected word recall, F (2, 10) = 6.93, p = .013, η^2_G = .41. Tukey's post hoc difference tests indicated that participants recognized significantly more words in the speech task (M = 7.17, SE = .62) than participants in the syntax task (M = 4.33, SE = .62; p = .011). There were no differences between the conceptual task (M = 6.17, SE = .62) and both the speech and syntax tasks.

9.3.5.2 Visualize the results

Similar to the dependent t-test, I am not a fan of the output for the repeated measures ANOVA in jamovi under "estimated marginal means." You could modify the same syntax for the dependent t-test (you'll need to change line 3 to be sure you're picking the correct lines of data, not just 2:3, and revise the names and such. You could also move the data into Excel and do bar graphs with error bars there instead.

9.3.6 In case of violated assumptions

We have already discussed what to do if you violate the assumption of sphericity above; you select one of the two sphericity corrections.

If you violate the assumption of normality or if the dependent variable is ordinal, then you can use the Friedman test. You can select this using the Repeated Measures ANOVA - Friedman option under the ANOVA analysis.

9.3.6.1 Friedman's test

Friedman's test can only examine one within-subjects variable, so you will move all three conditions (Speech, Conceptual, and Syntax) to the Measures box. Select `Pairwise comparisons (Durbin-Conover` for post hoc comparisons and `Descriptives` for the Means and Medians. Optionally, you can select to plot either the Means or Medians. The Setup is shown below.

Once you've set-up the analysis, you can interpret the results. Overall, we continue to see a statistically significant result and that there is only a significant difference between speech and syntax.

We can write up the results similarly as before:

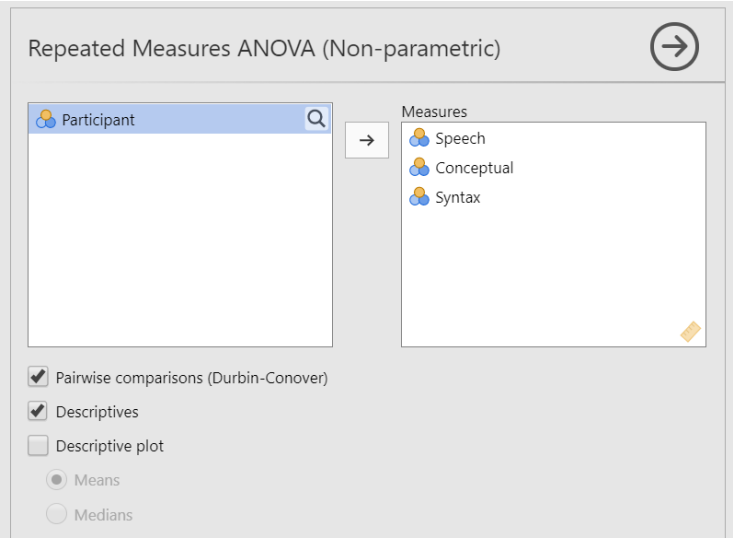


Figure 9.21: Repeated Measures ANOVA setup in jamovi

Friedman's test was performed examining how three tasks affected word recognition in patients suffering from Broca's Aphasia. Task significantly affected word recall, χ^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.

9.3.7 Your turn!

Open the `Sample_Dataset_2014.xlsx` file that we will be using for all Your Turn exercises. You can find the dataset here: Sample_Dataset_2014.xlsx Download

Perform a repeated measures ANOVA based on the following research questions. Check your assumptions and ensure you are using the correct tests.

To get the most out of these exercises, try to first find out the answer on your own and then use the drop-down menus to check your answer.

1. **Does students differ on their test scores (English, Reading, Math, Writing)?**
    - Based on your understanding of the nature of the test scores, which statistic should you use? repeated measures ANOVA Friedman's test
    - Should you apply a sphericity correction? If so, which one? N/A - us-

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]

Descriptives

	Mean	Median
Speech	7.17	7.50
Conceptual	6.17	6.50
Syntax	4.33	4.50

Figure 9.22: Repeated Measures ANOVA setup in jamovi

ing Friedman's test no, assumption satisfed yes, Greenhouse-Geisser yes, Huynh-Feldt

- Do students differ on their test scores? yes no
- Should you perform a planned contrast or post hoc comparison? yes no
- What are the results of the post hoc comparison? N/A - Don't perform All test scores were significantly different from one another All test scores were significantly different from one another except for English and Reading

9.4 Factorial ANOVA

9.4.1 Overview

Factorial ANOVA allows us to examine two or more independent variables (IVs) simultaneously. There are several types of factorial designs:

- **Independent factorial design**: several between-group (independent) IVs
- **Repeated measures factorial design**: several within-group (repeated-measures) IVs
- **Mixed factorial design**: some between-group and some within-group IVs

Furthermore, you may read about ANOVAs referred to as "one-way", "two-way", "three-way" or greater. This simply refers to how many independent variables there are. Factorial ANOVAs are sometimes also referenced by how many groups per IV there are; for example, a 2 x 3 ANOVA is a factorial ANOVA in which the first IV has two conditions and the second IV has three conditions. You would also specify which IVs are between-group and which are within-group. For example, you might write that your design is a 2 (between-subjects: gender) x3 (within-subjects: task) mixed factorial.

We won't be going into too much detail on the different factorial ANOVA designs. Instead, I will walk through illustrative cases so that if you want to apply them in the future you can mimic the procedures below.

9.4.2 Independent Factorial ANOVA

The independent factorial ANOVA has multiple between-group IVs. Let's run an example with data from lsj-data. Open data from your Data Library in "lsj-data". Select and open "rtfm". This data has two IVs: attend (whether or not the student turned up to lectures) and reading (whether or not the student actually read the textbook). 1 = they did and 0 = they did not.

Because we do not have a within-group IV, we will select the ANOVA analysis. Move grade to your Dependent Variable box and both attend and reading to your Fixed Factors. Select all the same options as you did for the one-way ANOVA (i.e., ω^2, assumption checks, Tukey's post hoc tests for the two variables attend and reading, estimated marginal means).

Let's go through the output (check that your output matches!) and then discuss how to write up the results in APA format. First, we need to check assumptions. The Levene's test and Shapiro-Wilk's test are shown below. We can see that we meet the assumption of normality but fail to meet the assumption of homogeneity of variance. Unfortunately, we cannot perform a Welch's F-test with more than one independent factorial so we will note this failed assumption and move on.

Assumption Checks

Homogeneity of Variances Test (Levene's)

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

[3]

Normality Test (Shapiro-Wilk)

Statistic	p
0.96	0.851

Figure 9.23: Assumption check results in jamovi

Let's look at the main results next. We got three lines of results in addition to the typical residuals (error). The first two lines 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 of `attend * reading`, which is not statistically significant. This means we do not have an interaction between attend and reading on grades. Interactions will be discussed in more detail in the next section.

Although we could simply look at the means to know whether attending or reading had higher grades than not attending or not reading because there are only two conditions, we can also look at the post hoc tests and definitely need to look at them if we have three or more conditions per IV. These are shown below. Because the mean differences are negative, we can determine that the

ANOVA

ANOVA - grade

	Sum of Squares	df	Mean Square	F	p	ω^2
attend	648.00	1	648.00	18.25	0.013	0.26
reading	1568.00	1	1568.00	44.17	0.003	0.64
attend ✻ reading	8.00	1	8.00	0.23	0.660	−0.01
Residuals	142.00	4	35.50			

Figure 9.24: Independent factorial ANOVA results in jamovi

second group had higher means than the second group. We can confirm that with the estimated marginal means (not shown here).

Post Hoc Tests

Post Hoc Comparisons - reading

Comparison			Mean Difference	SE	df	t	p_{tukey}	Cohen's d
reading		reading						
0	-	1	−28.00	4.21	4.00	−6.65	0.003	−4.70

Note. Comparisons are based on estimated marginal means

Post Hoc Comparisons - attend

Comparison			Mean Difference	SE	df	t	p_{tukey}	Cohen's d
attend		attend						
0	-	1	−18.00	4.21	4.00	−4.27	0.013	−3.02

Note. Comparisons are based on estimated marginal means

Figure 9.25: Post hoc results in jamovi

Last, we can write-up our results!

> We were interested in knowing how attendance and reading affected student grades. An independent factorial ANOVA showed that both attendance (F [1, 4] = 18.25, p = .013, ω^2 = .26) and reading (F [1, 4] = 44.17, p = .003, ω^2 = .64) affected student grades; there was no significant interaction between attendance and reading (F [1, 4] = 8.00, p = .660, ω^2 = -.01). Post hoc comparisons using Tukey's HSD show that students who attended lectures (M = 75.50, SE = 2.98) had higher grades than students who did not (M = 57.50, SE = 2.98; p = .003, d = 4.70); furthermore, students who read (M = 80.50, SE = 2.98) had higher grades than students who did not (M

$= 52.50$, $SE = 2.98$; $p = .013$; $d = 3.02$).

9.4.2.1 Interactions

Interactions occur when the effect of one IV on the DV depends on the lvel of the other IV. If you did not want to test for interaction effects, you could remove them from the Model Terms in the Model drop-down menu.

However, by default they will include them. If you have 2-3 IVs, it may be reasonable to look at these interactions. However, 3-variable interactions (e.g., IV1 * IV2 * IV3) are pushing it and 4-variable interactions are highly implausible. Be critical in which interaction terms you include!

jamovi can provide a plot of your interaction, which can be helpful to help interpret your results. Below is the plot for our interaction of attendance on reading.

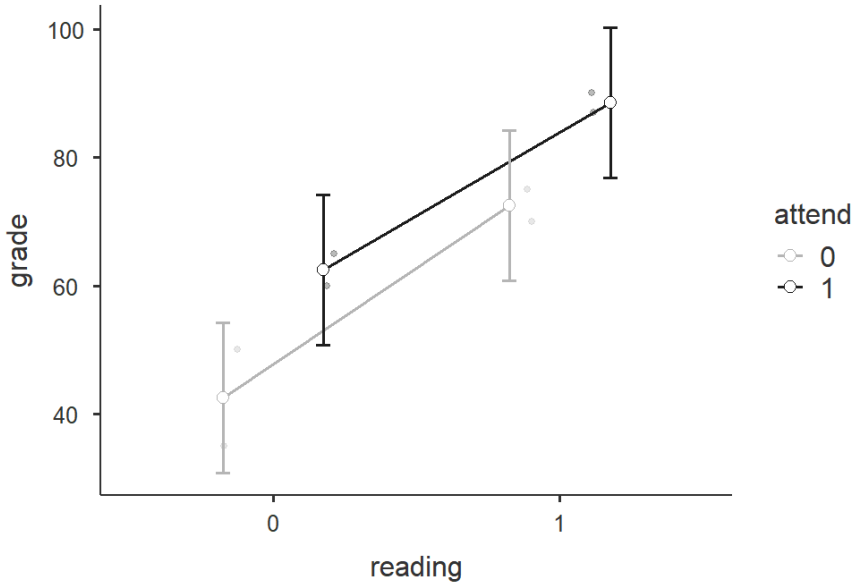


Figure 9.26: Interaction plot in jamovi

The parallel lines that are sloping upward tell me there is a significant main effect for both IVs but no interaction. How do I know that? With two variables, there are only so many interaction shapes possible. This website does a fantastic time showing you all 8 combinations of the three effects (2 main effects and 1 interaction effect). Spend some time looking through it and familiarizing yourself with the plots!

9.4.3 Repeated Measures Factorial ANOVA

This is also sometimes called the two-way (or three-way or n-way, depending on the n of IVs you have) repeated measures ANOVA. Let's go through an example repeated measures factorial ANOVA. The dataset is courtesy of Real Statistics Using Excel.

> A company has created a new training program for their customer service staff. To test the effectiveness of the program they took a sample of 10 employees and assessed their performance in three areas: Product (knowledge of the company's products and services), Client (their ability to relate to the customer with politeness and empathy) and Action (their ability to take action to help the customer). They then had the same 10 employees take the training course and rated their performance after the program in the same three areas. -Real Statistics Using Excel

You can find the dataset here to follow along: Repeated-measures-factorial-ANOVA.xlsx Download

In jamovi, select Repeated Measures ANOVA under the ANOVA analysis option. Here are the general steps:

1. In Repeated Measures Factors, you'll need to name both factors. Rename `RM Factor 1` to Time and `RM Factor 2` to Area. Under Time, specify two levels: Pre and Post. Under Area, specify three levels: Product, Client, and Action.
2. In Repeated Measures Cells, you'll now have six cells with the combinations of the 6 columns you have. Drag the variable from the left into the correct cell on the right. Be careful here! For example, you need to put the Pre-Product variable into the cell "Pre, Product".
3. Name your dependent variable label "Performance" and select 'Generalised η^2.
4. Under Assumption Checks, select Sphericity tests
5. Under Post Hoc Tests, move Area and Time over and select Tukey.
6. Under Estimated Marginal Means, move Time over into Term 1, Area into Term 2, and both Time and Area into Term 3. Select plots and tables, Observed scores, and Equal cell weights.

Now let's go over selected output. First, we need to check our assumption of sphericity. All the Mauchly's W's are not statistically significant so we satisfy the assumption of sphericity and do not need to apply any sphericity corrections.

Next let's look at the within subjects effects table. Remember, we do not need to worry about the between subjects effects table because we do not have one; it will be used in the mixed factorial design below. Overall, we see a significant

Assumptions

Tests of Sphericity

	Mauchly's W	p	Greenhouse-Geisser ε	Huynh-Feldt ε
Area	0.97	0.878	0.97	1.00
Time	1.00	NaN [a]	1.00	1.00
Area ✻ Time	0.62	0.152	0.73	0.83

[a] The repeated measures has only two levels. The assumption of sphericity is always met when the repeated measures has only two levels

Figure 9.27: Assumption testing in jamovi

main effect of area, a significant main effect of time, and a significant interaction effect of both area and time. Neat!

Repeated Measures ANOVA

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G
Area	1365.23	2	682.62	26.96	< .001	0.36
Residual	455.77	18	25.32			
Time	828.82	1	828.82	33.85	< .001	0.25
Residual	220.35	9	24.48			
Area ✻ Time	224.43	2	112.22	12.63	< .001	0.08
Residual	159.90	18	8.88			

Note. Type 3 Sums of Squares

Figure 9.28: Repeated measures ANOVA in jamovi

Next, we can look at post hoc comparisons because the main effects were all statistically significant. For area, we can see that client and action had significantly higher means than product, but there was no difference between client and action. Furthermore, post-intervention performance was significantly higher than pre-intervention.

Last, let's look at the interaction to get a sense of what the interaction looks like. It appears that although there are no differences between pre- and post-intervention for product, there are significant differences from pre- to post-intervention for both client and action. To be more specific on where the statistically significant differences are, you can also ask for post hoc tests for the

Post Hoc Tests

Post Hoc Comparisons - Area

Comparison			Mean Difference	SE	df	t	p_{tukey}
Area		Area					
Product	-	Client	−8.60	1.59	18.00	−5.40	< .001
	-	Action	−11.15	1.59	18.00	−7.01	< .001
Client	-	Action	−2.55	1.59	18.00	−1.60	0.270

Post Hoc Comparisons - Time

Comparison			Mean Difference	SE	df	t	p_{tukey}
Time		Time					
Pre	-	Post	−7.43	1.28	9.00	−5.82	< .001

Figure 9.29: Post hoc tests in jamovi

interaction term. This is where including a plot can be very helpful for your audience!

Now we have everything we need (in addition to the estimated marginal means) and can write-up our results.

> We tested a 2 (time: pre- and post-intervention) x 3 (area: product, client, action) repeated measures factorial design to examine how both time and area affected performance. We satisfied the assumption of sphericity for all effects. There was a significant main effect of time (F [1, 9] = 33.85, $p < .001$, $\eta^2_G = .25$) such that performance at post-intervention ($M = 26.80$, $SE = 1.84$) was higher than at pre-intervention ($M = 19.37$, $SE = 1.84$). There was also a significant main effect of area (F [2, 18] = 26.96, $p < .001$, $\eta^2_G = .36$) such that both client ($M = 25.10$, $SE = 1.95$) and action ($M = 27.65$, $SE = 19.5$) performance was higher than product performance ($M = 16.50$, $SE = 1.95$), but there was no difference between client and action performance. Lastly, there was a significant interaction between time and area such that there were no differences in product performance from pre- to post-intervention but there was for client and action performance (see Figure 1).

9.4.4 Mixed Factorial ANOVA

You can find the dataset here to follow along: mixed-factorial.sav Download

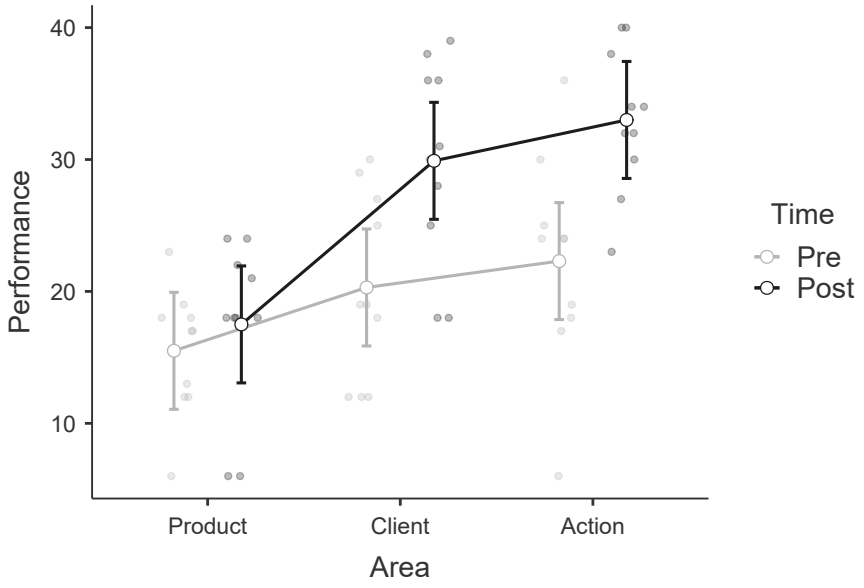


Figure 9.30: Interaction plot in jamovi

This dataset comes from a larger studying examining the effect of a delayed reward preference of three commodities (food, money, and music) on food cravings (as rated by the Food Craving Questionnaire [FCQ]) for each participant. Participants were in one of two conditions: the control condition did not do anything and the experimental condition had participants do the tasks while fasting.

Therefore, this study is a 2 (between-subjects: condition [control or fasting]) x 3 (within-subjects: reward [food, money, and music]) mixed factorial design.

To perform a mixed factorial ANOVA, we use the same procedures as the repeated measures ANOVA but we also need to add a between-subjects factor.

1. To perform a mixed factorial ANOVA in jamovi, go to the Analyses tab, click the ANOVA button, and choose “Repeated Measures ANOVA”.

2. Under “Repeated Measures Factors” name your independent variable. In this case you can name it “Reward”. Rename the three levels of Task: Food, Money, and Music.

3. Under “Repeated Measures Cells” move the given variable into the correct level. For example, you’ll move FQ_1 to Food, FQ_2 to Money, and FQ_3 to Music.

4. Under “Between Subject Factors” add your between-subjects variable ”con-

dition.

5. Select Generalised η^2 as your measure of effect size.
6. In the Assumption Checks drop-down menu, select `Sphericity tests`.
7. In the Post Hoc Tests drop-down menu, move your two independent variables over and select `Tukey`. Remember that we only interpret these if the overall F is statistically significant.
8. In the Estimated Marginal Means drop-down menu, move Reward, Condition, and both reward and condition into the terms under Marginal Means. Select tables and plots, and select `Observed scores`. Uncheck `Equal cell weights`.

Now let's go through the output!

First, as always we check our assumption of sphericity. Mauchly's W is not statistically significant ($p = .073$) so we satisfy the assumption of sphericity. We do not need to apply a sphericity correction.

Assumptions

Tests of Sphericity

	Mauchly's W	p	Greenhouse-Geisser ε	Huynh-Feldt ε
Reward	0.95	0.073	0.95	0.97

Figure 9.31: Assumption checking in jamovi

Next, we interpret out output! This time we interpret both our within subjects effects and between subjects effects tables. In the within subjects effects table, our main effect of reward is statistically significant. In the between subjects effects table, our main effect of condition is statistically significant. However, in the within subjects effects table, there is no statistically significant interaction effect of reward on condition.

To understand where the differences lie between conditions or reward preferences, we look to our post hoc tests. For reward, it looks like there is only a significant difference between food and music ($p = .009$). For condition, it looks like cravings were higher in the fasting group than in the control group. We can look at the estimated marginal means tables to find the actual Means of the conditions and see the plots.

We can look at the interaction plot, but notice that the lines are parallel which is a good indication that there is no significant interaction.

Lastly, we can write up the results in APA style!

Repeated Measures ANOVA

Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G
Reward	0.98	2	0.49	4.45	0.013	0.00
Reward ✻ condition	0.14	2	0.07	0.66	0.519	0.00
Residual	21.59	196	0.11			

Note. Type 3 Sums of Squares

[3]

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G
condition	163.39	1	163.39	82.44	< .001	0.43
Residual	194.24	98	1.98			

Note. Type 3 Sums of Squares

Figure 9.32: Mixed factorial ANOVA in jamovi

Post Hoc Tests

Post Hoc Comparisons - Reward

Comparison			Mean Difference	SE	df	t	p_{tukey}
Reward		Reward					
Food	-	Money	0.07	0.05	196.00	1.41	0.340
	-	Music	0.14	0.05	196.00	2.98	0.009
Money	-	Music	0.07	0.05	196.00	1.58	0.258

Post Hoc Comparisons - condition

Comparison			Mean Difference	SE	df	t	p_{tukey}
condition		condition					
control	-	fasting	−1.48	0.16	98.00	−9.08	< .001

Figure 9.33: Post hoc tests in jamovi

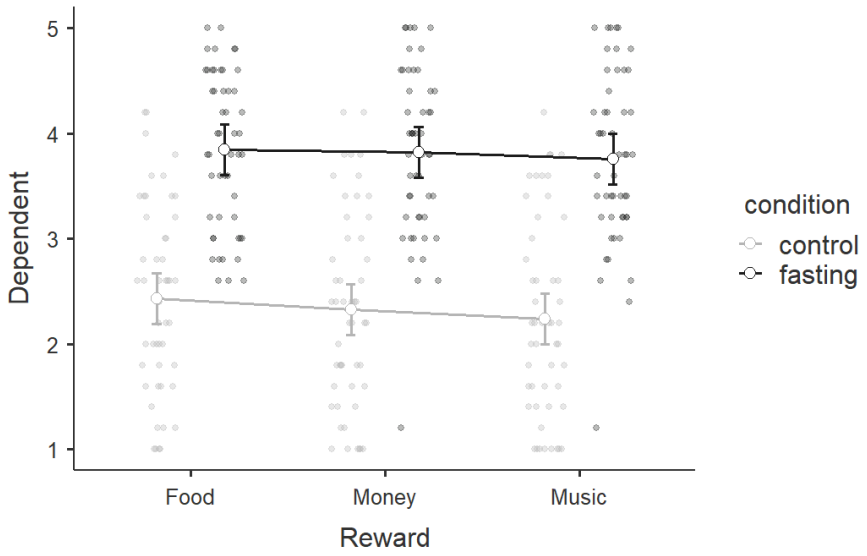


Figure 9.34: Interaction plot in jamovi

To test how both condition (control or fasting) and reward preference (food, money, and music) affected food cravings, we performed a mixed factorial ANOVA. There was a significant main effect of condition (F [2, 196] = 4.45, p = .013, η^2_G = .00) and a significant main effect of reward (F [1, 98] = 82.44, p < .001, η^2_G = .43). However, there was no statistically significant interaction effect of reward on condition (F [2, 196] = .66, p = .519, η^2_G = .00).

For condition, participants who fasted (M = 3.81, SE = .11) reported significantly more food cravings than participants in the control condition (M = 2.33, SE = .11; p_{Tukey} < .001). For reward, the food reward (M = 3.14, SE = .09) led to significantly higher food cravings than the music reward (M = 3.00, SE = .09; p_{Tukey} = .009), but there was no differences between the money reward (M = 3.07, SE = .09) and both food (p_{Tukey} = .340) or music (p_{Tukey} = .258).

9.5 ANCOVA

9.5.1 Overview

ANCOVA (**AN**alysis of **COVA**riance) examines the difference in means between three or more groups, while controlling for or partialling out the effect of one or more continuous confounds or covariates.

Some definitions: A *confounding* variable is a variable that affects or is related to both the independent and dependent variable. A *covariate* variable is a variable that only affects or is only related to the dependent variable.

There are two main reasons for including covariates:

1. **To reduce within-group error variance**: Remember that to get a larger F-statistic, we need to maximize between-groups variance and minimize within-groups variance. Adding covariates can sometimes minimize within-groups variance if that covariate helps *explain* some of the within-group variance.
2. **Elimination of covariates**: Sometimes there are other variables that also explain our outcome variable. We want to look at the effect of another variable on the outcome while removing or eliminating the other variables (confounds) that also explain our outcome variable.

9.5.2 Assumptions

In addition to the same assumptions of the one-way ANOVA (see 9.1.3), the ANCOVA has two additional assumptions:

1. **Independence of the covariate and treatment effect**: When the covariate and treatment effect are related, then we can have incorrect F-statistic values. However, this is only important in experimental designs. In quasi-experimental designs, this is often violated and you just have to interpret results accordingly.
    - If you do have an experimental manipulation with a covariate, you can test this assumption by running a one-way ANOVA but with your experimental manipulation as your IV or group variable and your covariate as your DV. If there is a significant F-ratio, then you have violated this assumption.
2. **Homogeneity of regression slopes**: The relationship between the covariate and the outcomes must be similar across groups.
    - To test this assumption, add an interaction term between the covariate and each independent variable in jamovi under the Model drop-down menu. Add the interactions as model terms.

9.5.3 Perform the test

Let's run an example with data from lsj-data. Open data from your Data Library in "lsj-data". Select and open "ancova". This data is fictional data from a health psychologist who was interested in the effect of routine cycling (1 = driving, 2 = cycling) and stress (1 = high, 2 = low) on happiness levels, with age as a covariate. Notice how this is a 2x2 independent factorial design with a covariate!

1. To perform an ANCOVA in jamovi, select ANCOVA under the ANOVA analysis menu.
2. Move your dependent variable `happiness` to the Dependent Variable box, your independent variables `stress` and `commute` to the Fixed Factors box, and your covariate `age` to the Covariates box.
3. Select ω^2 as your effect size.
4. Under Assumption Checks, select all three assumption checks: `Homogeneity test`, `Normality test`, and `Q-Q Plot`.
5. Under Post Hoc Tests, move both of your independent variables over, select the `Tukey` correction and select `Cohen's d` for your effect size.
6. Under Estimated Marginal Means, move each of your independent variables over into its own term box. Also include combinations of your independent variables if you have an interaction term in your model. Select both plots and tables, select `Observed scores`, and de-select `Equal cell weights`.

First, let's check our assumptions in jamovi. Shapiro-Wilk's test was not statistically significant ($p = .735$) and the Q-Q plot looks good; therefore, we've satisfied the assumption of normality. Levene's test was not statistically significant ($p = .925$); therefore, we've satisfied the assumption of homogeneity of variance.

Assumption Checks

Homogeneity of Variances Test (Levene's)

F	df1	df2	p
0.16	3	16	0.925

[3]

Normality Test (Shapiro-Wilk)

Statistic	p
0.97	0.735

Figure 9.35: Assumption check results in jamovi

However, we have two additional assumptions we need to check. Let's check the assumption of independence of the covariate and treatment effect. For that, we need to perform another ANOVA (not an ANCOVA) with our independent variables predicting age. Our results indicate we violate this assumption: both `stress` and the interaction of `stress * commute` are related to age. This suggests age is in fact a *confounding* variable, not a covariate. We should be performing a mediation, but because we want to illustrate the ANCOVA we will continue.

ANOVA

This is to test the assumption of independence between the covariate and the treatment effect.

ANOVA - age

	Sum of Squares	df	Mean Square	F	p
stress	1155.20	1	1155.20	7.79	0.013
commute	12.80	1	12.80	0.09	0.773
stress ✻ commute	605.00	1	605.00	4.08	0.060
Residuals	2371.20	16	148.20		

Figure 9.36: Assumption check results in jamovi

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). We can test this by adding an interaction term between the covariate and each independent variable in jamovi under the Model drop-down menu. If the interaction effect is not significant it can be removed. If it is significant then a different and more advanced statistical technique might be appropriate (which is beyond the scope of this class). In our case, the interactions between each IV and our covariate are not statistically significant so we can remove the interaction terms and move on.

Now it's time to interpret the results! The ANCOVA table shows that both independent variables (`stress` and `commute`), the interaction term (`stress * commute*`), and the covariate (`age`) are statistically significant. Therefore, we can look at our post hoc tests to find where the differences are.

Technically, we don't need to look at the post hoc table much in this example. Because there are only two groups, we already know one group will have higher means than the other group if the F-test is significant. In fact, check this out: the square root of our F-statistic is equal to the t-statistic in our post hoc table. Neat!

Post hoc tests show that low stress had higher happiness than high stress, and that cycling had higher happiness than driving. We can also look to the esti-

ANCOVA

ANCOVA - happiness

	Sum of Squares	df	Mean Square	F	p	ω^2
stress	136.74	1	136.74	2.40	0.145	0.05
commute	83.75	1	83.75	1.47	0.247	0.02
age	234.02	1	234.02	4.11	0.064	0.11
stress ✻ commute	324.44	1	324.44	5.70	0.033	0.17
stress ✻ age	24.88	1	24.88	0.44	0.520	−0.02
commute ✻ age	5.05	1	5.05	0.09	0.771	−0.03
Residuals	740.10	13	56.93			

Figure 9.37: Assumption check results in jamovi

ANCOVA

ANCOVA - happiness

	Sum of Squares	df	Mean Square	F	p	ω^2
stress	2751.52	1	2751.52	52.61	< .001	0.39
commute	2213.93	1	2213.93	42.33	< .001	0.31
age	334.35	1	334.35	6.39	0.023	0.04
stress ✻ commute	740.12	1	740.12	14.15	0.002	0.10
Residuals	784.45	15	52.30			

Figure 9.38: ANCOVA results in jamovi

mated marginal means tables and plots for information for reporting.

Post Hoc Tests

Post Hoc Comparisons - stress

Comparison			Mean Difference	SE	df	t	p_{tukey}	Cohen's d
stress		stress						
1	-	2	−28.61	3.94	15.00	−7.25	< .001	−3.96

Note. Comparisons are based on estimated marginal means

Post Hoc Comparisons - commute

Comparison			Mean Difference	SE	df	t	p_{tukey}	Cohen's d
commute		commute						
1	-	2	−21.10	3.24	15.00	−6.51	< .001	−2.92

Note. Comparisons are based on estimated marginal means

Figure 9.39: Post hoc results in jamovi

Last, we can write-up our results! Reporting ANCOVA is very similar to reporting an ANOVA test (in this case an independent factorial ANOVA) except that we also report the effect of the covariate, as well. Here's an example write-up:

> We conducted a study examining how stress and commute affect happiness levels in a 2 (stress: high or low) x 2 (commute: cycling or driving) independent factorial design. Furthermore, we collected data on age as a covariate of our study. We satisfied all assumptions of the ANCOVA except that age was in fact a confounding variable in that it relates to our independent variable of stress. Despite failing to meet this assumption, we proceeded with the ANCOVA analysis.
>
> There was a significant main effect of stress on happiness such that participants in the low stress condition (M = 68.45, SE = 2.55) reported significantly greater happiness than participants in the high stress condition (M = 39.85, SE = 2.55), F (1, 15) = 52.61, p < .001, ω^2 = .39. There was also a significant main effect of commute on happiness such that participants who commuted via cycling (M = 64.70, SE = 2.29) reported significantly greater happiness than participants who commuted via driving (M = 43.60, SE = 2.29), F (1, 15) = 42.33, p < .001, ω^2 = .31. There was a significant interaction between stress and commute type such that happiness levels were similar in the low stress condition for both commute types, but happiness was significantly higher for participants who cycled versus those who drove in the high stress condition, F (1, 15) = 14.15, p = .002, ω^2 = .10. Furthermore, age was a significant

covariate of our dependent variable, F (1, 15) = 6.39, p = .023, ω^2 = .04.

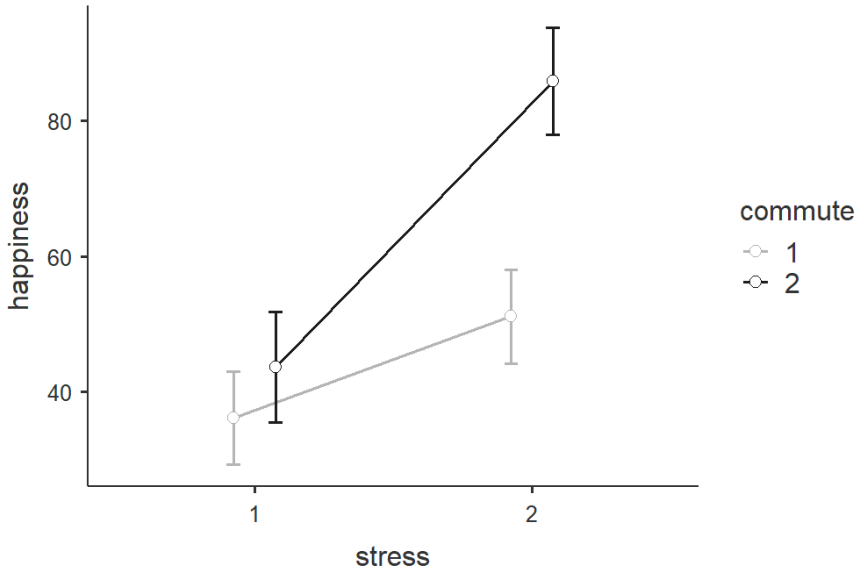


Figure 9.40: Interaction in jamovi

Chapter 10

Correlation and regression

Correlations are the relationships between two (usually) continuous variables.

Regressions have one or more predictor variables (usually at least one is a continuous variable) and a single continuous dependent variable.

10.1 Correlation

10.1.1 Overview

Correlation (r) tests the relationship between two variables, which are usually continuous (i.e., ratio or interval) variables. The relationship between those two variables could be *positively related*, *negatively related*, or *not related at all*.

Note: Remember that correlation does not always equal causation!

We will learn about other types of correlations, but mainly we are interested in the Pearson product moment correlation, which is often just called the Pearson correlation or just correlation. Other correlations are generally referred to by their specific name.

The strength of the correlation is determined by how closely the dots in the scatterplot matrix fit on the regression line. Correlations are really *standardized covariances*. Covariance is the extent to which the deviation of one variable from its mean matches the deviation of the other variable from its mean. We then standardize the covariance into the correlation which ranges from -1 to +1. Because correlations are standardized, they are considered effect sizes! Commonly, we describe values of $\pm.1$ as a small effect size, $\pm.3$ as medium, and $\pm.5$ as large.

Note: Try your hand at guessing the correlation coefficient! Play at www.guessthecorrelation.com

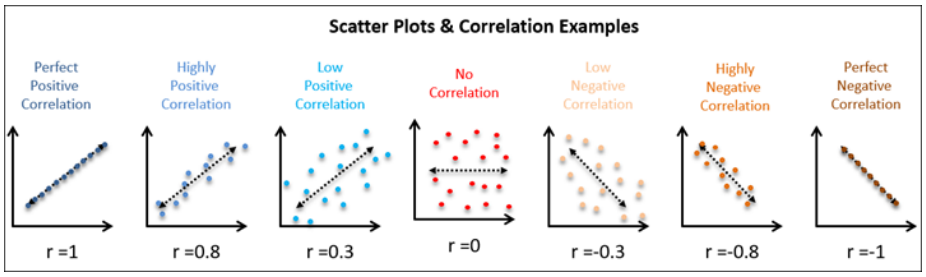


Figure 10.1: Scatter plots and correlation examples

Note: Another fun website to play around with is the interactive visualization website Interpreting Correlations by Kristoffer Magnusson. You can play around with what data looks like at various correlations. https://rpsychologist.com/correlation/

10.1.2 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 "parenthood". This dataset measures a new mother's[1] daily grumpiness very precisely, on a scale from 0 (not at all grumpy) to 100 (extremely grumpy). In addition, I am also tracking her sleeping patterns and her son's sleeping patterns across 100 days.

10.1.2.1 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.

[1] The dataset built into jamovi says the person's name is Dan, but they now go by Danielle. You can follow her on Twitter at @djnavarro.

	ID	dan.sleep	baby.sleep	dan.grump	day
1	1	7.59	10.18	56	1
2	2	7.91	11.66	60	2
3	3	5.14	7.92	82	3
4	4	7.71	9.61	55	4
5	5	6.68	9.75	67	5
6	6	5.99	5.04	72	6
7	7	8.19	10.45	53	7
8	8	7.19	8.27	60	8
9	9	7.40	6.06	60	9
10	10	6.58	7.09	71	10

10.1.2.2 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 can see first 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.

We also see skew and kurtosis. Calculating the z-score for all the skew and kurtosis (remember: skew or kurtosis divided by its standard error) suggests we do not violate the assumption of normality much except for `day`. However, notice what the variable `day` is! It's just the day of the study, from 1-100. If you look at the graph, it has a *uniform distribution* (completely flat and uniform) not a normal distribution (bell curve)!

Descriptives

	dan.sleep	baby.sleep	dan.grump	day
N	100	100	100	100
Missing	0	0	0	0
Mean	6.97	8.05	63.71	50.50
Median	7.03	7.95	62.00	50.50
Standard deviation	1.02	2.07	10.05	29.01
Variance	1.03	4.30	101.00	841.67
Minimum	4.84	3.25	41	1
Maximum	9.00	12.07	91	100
Skewness	−0.30	−0.02	0.45	0.00
Std. error skewness	0.24	0.24	0.24	0.24
Kurtosis	−0.65	−0.61	−0.04	−1.20
Std. error kurtosis	0.48	0.48	0.48	0.48
Shapiro-Wilk W	0.98	0.98	0.98	0.95
Shapiro-Wilk p	0.069	0.256	0.122	0.002

10.1.3 Check assumptions

10.1.3.1 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)
3. The relationship between the two variables is **linear**

We test for normal distribution using the Exploration-Descriptives analysis in jamovi, looking at Shapiro-Wilk's test, the Q-Q plot, a histogram or density plot, and skew and kurtosis z-scores.

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

10.1.3.2 Checking assumptions

10.1.3.2.1 Testing normality By now we've had a lot of practice testing for normality. One of our data points (`day`) is strange because it's just a linear number 1-100, so we can ignore it. The Q-Q plot for `dan.sleep` looks a bit iffy,

but the density plot, skew, kurtosis, and Shapiro-Wilk's tests look fine. We will say we met the assumption of normality. Below, you can see our density plots in the diagonal of our scatterplot matrix.

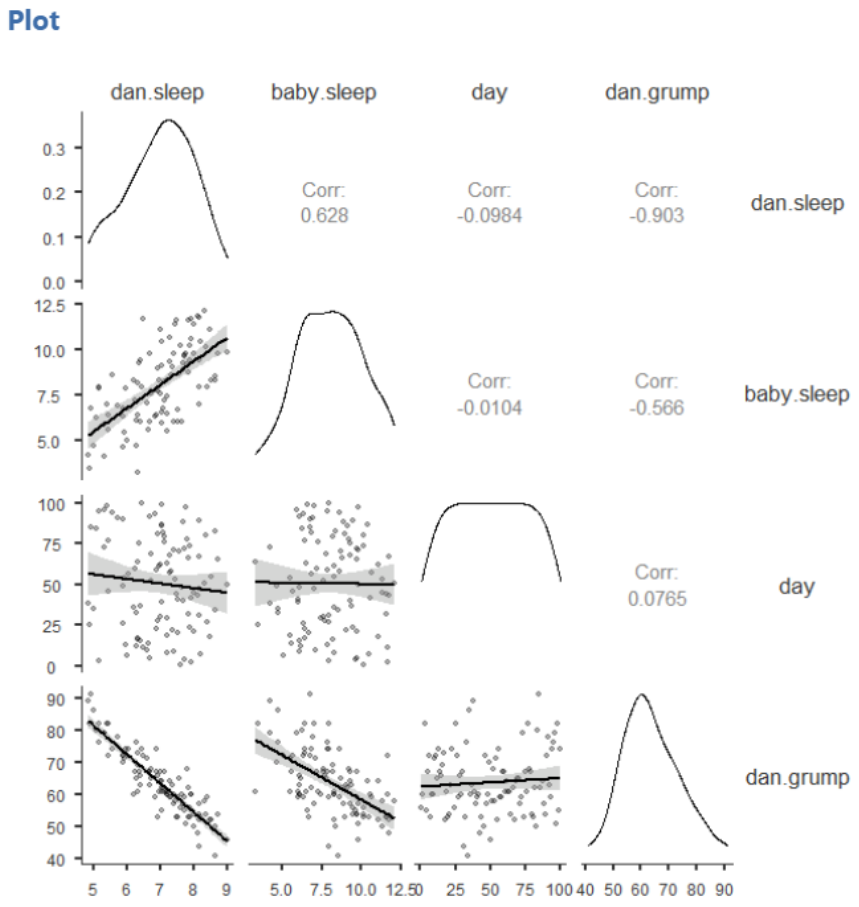


Figure 10.2: Testing linearity in jamovi

10.1.3.2.2 Testing linearity There's nothing we can do here except look out our correlations! Do the underlying data in any of the scatterplots look like there is actually a non-linear (e.g., curvilinear) relationship? If so, you fail to meet this assumption.

The scatterplots above do not suggest a non-linear relationship, so we meet the assumption of linearity.

10.1.4 Perform the test

1. First, you'll need to check your assumption of normality *outside* of the correlations analysis. Go to Explorations and choose Descriptives and check whether you meet the assumption of normality.

2. To perform a correlation, go to Regression and select Correlation Matrix.

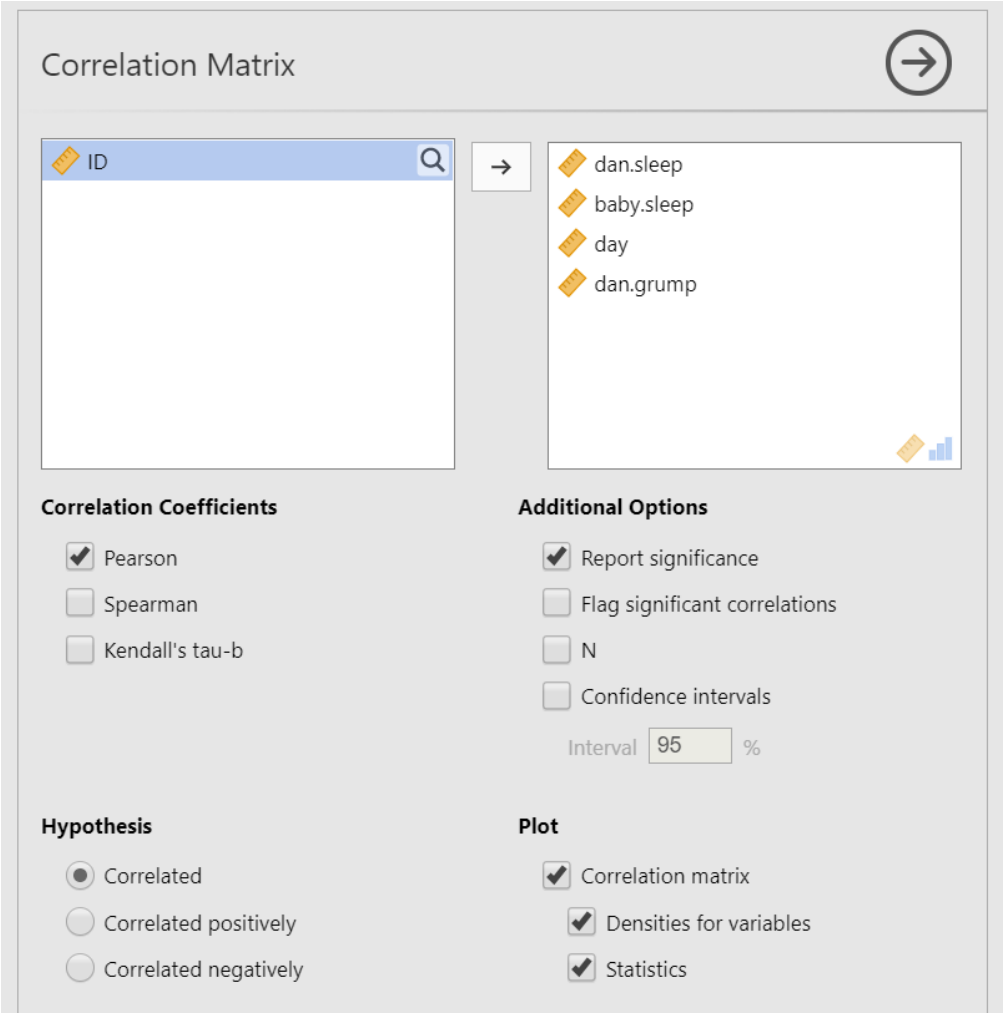
3. Move all four variables into the dialogue box on the right (`dan.sleep`, `baby.sleep`, `day`, and `dan.grump`).

4. Select Pearson under Correlation Coefficients. We'll go over the other two later.

5. Under Additional Options, select `Report significance`, `Flag significant correlations`, and, if you have missing data, `N` (we don't have missing data so we can ignore this).

6. Under Plot, select `Correlation matrix`. Alternatively, you can ask for `Densities for variables` to see the density plots for each variable and `Statistics` to have the correlation coefficient added to the plot.

When you are done, your setup should look like this:



10.1.5 Interpreting results

Once we are satisfied we have satisfied the assumptions for the correlation, we can interpret our results.

Correlation Matrix

Correlation Matrix

		dan.sleep	baby.sleep	day	dan.grump
dan.sleep	Pearson's r	—			
	p-value	—			
baby.sleep	Pearson's r	0.63	—		
	p-value	< .001	—		
day	Pearson's r	−0.10	−0.01	—	
	p-value	0.330	0.918	—	
dan.grump	Pearson's r	−0.90	−0.57	0.08	—
	p-value	< .001	< .001	0.449	—

It looks like three of the variables are significantly ($p < .05$) correlated with each other: `dan.sleep`, `baby.sleep`, and `dan.grump`. `day` does not seem to be significantly correlated with any of the other three variables.

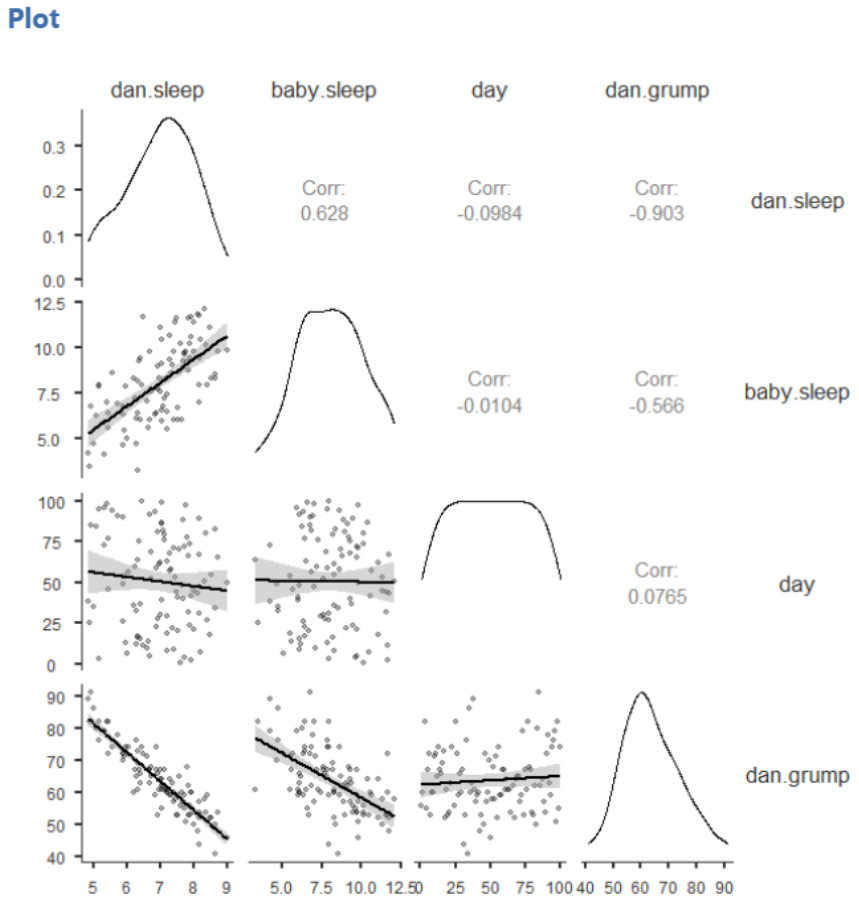
10.1.5.1 Write up the results in APA style

We can write up our results in APA something like this:

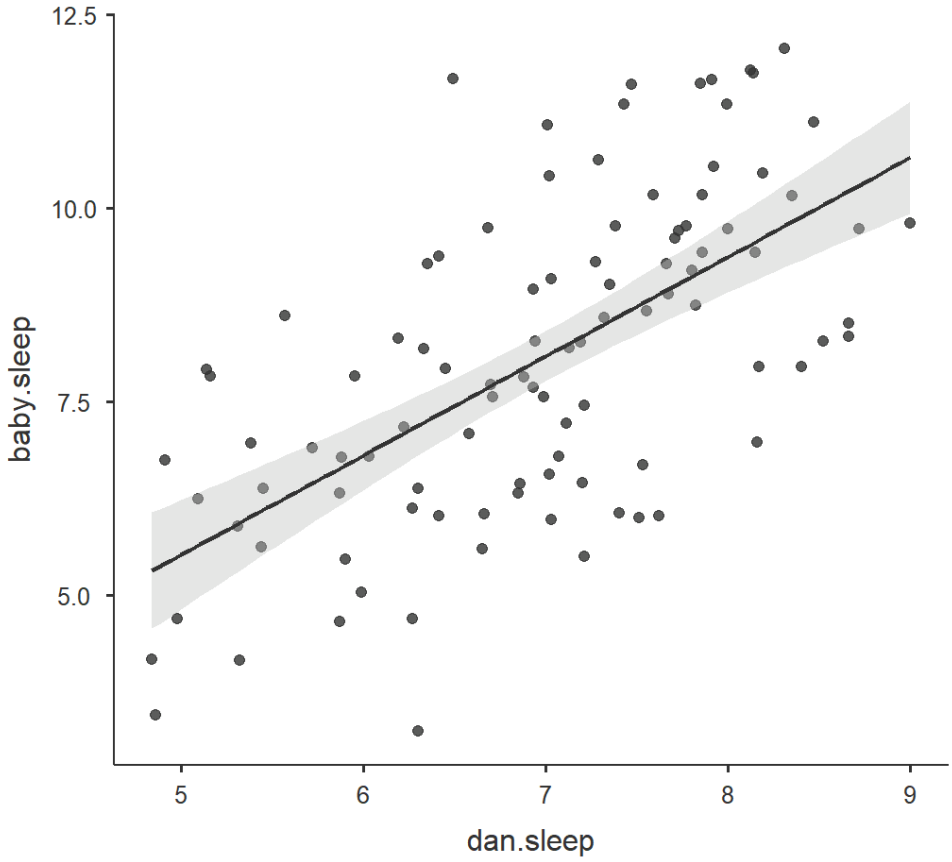
> Dan's grumpiness ($M = 63.71$, $SD = 10.05$) is negatively correlated with both Dan's quality of sleep ($M = 6.97$, $SD = 1.02$; r = -.90, $p < .001$) and the baby's quality of sleep ($M = 8.05$, $SD = 2.07$; r = -.57, $p < .001$). Furthermore, Dan's and the baby's quality of sleep are positively correlated ($r = .63$, $p < .001$).

10.1.5.2 Visualize the results

The default in the Correlation Matrix is to plot the correlation matrix of all the variables, and optionally show the densities for variables and the statistics. This is fine, but I'm not a huge fan of it. You can see it below:



Personally, this is why I like the scatr module in jamovi. You can create high-quality scatterplots of the six graphs above and then stitch them together in a nicer version. For example, here's the correlation between the sleep quality of both Dan and the baby with a linear regression line and standard error:



10.1.6 In case of violated assumptions

If you violate any of the three assumptions, you can choose to perform Spearman's rank correlation instead of a Pearson correlation. Both Spearman's rho and Kendall's tau are non-parametric statistics based on rank order. To perform Spearman's correlation, change the check mark in jamovi from Pearson to Spearman. You will interpret just the same; however, instead of using the letter r you can either use r_s, $r_{spearman}$, or ρ (the Greek letter rho).

What about Kendall's tau? It will likely give you the same results as Spearman's rho but it is interpreted slightly differently. We won't use it in this class.

10.1.7 Additional information

10.1.7.1 R-Squared

The cool thing about the correlation is that we can square r to get r^2, which is the percentage of variance overlap. It is the percentage of variance in one vari-

able that is shared by the other. You simply square the r correlation coefficient to find the r^2. For example, our correlation above between Dan's grumpiness and Dan's quality of sleep is r = -.90; therefore, its r^2 = .81 or 81%. 81% of the variance in Dan's grumpiness can be explained by Dan's quality of sleep!

10.1.7.2 Comparing strengths of correlations

Note: PSYC 290 students you can read the below information but I will not test you on it.

Sometimes you want to compare two correlations to find out if one correlation is significantly stronger than another. You can use this calculator to calculate the p-value: Testing the Significance of Correlations

Note that you use #1 (Independent Samples) when the correlations come from different samples and #2 (Dependent Samples) when the correlations come from the same sample. For example, to compare the correlation between English and Reading to the correlation between English and Writing, you would use #2 (Dependent Samples). But to compare the correlations between English and Reading for men and women, you would use #1 (Independent Samples).

Let's try them both with our Sample_Dataset_2014.xlsx.

1. **Comparison of correlations from independent samples**

We want to test the correlations between English and Reading for men and women. We first need to gather those correlations in jamovi. We can do this through **filters**.

Let's first find the correlation for men. Go to the Data tab in jamovi, click Filters, and enter in the $\int_x$ = Gender == 0. Next, go to the Analyses tab in jamovi, click Regression, and choose Correlation Matrix. Move our two variables over (`English` and `Reading`) and check the box for `N`. You should get r = .36, $p < .001$, n = 181.

Now let's find the correlation for women. Go back to the Data tab in jamovi, click Filters, and change the equation to $\int_x$ = Gender == 1. Your results should automatically update because the filter changed. For women, you should get r = .33, $p < .001$, n = 210.

Now we can compare the correlations in Testing the Significance of Correlations - Independent Samples. In Correlation 1, put 181 in the n column and .36 in the r column. In Correlation 2, put 210 in the n column and .33 in the r column. The results are shown below. The z-score is not statistically significant (p = .369) so there is no significant difference in correlation strength.

2. **Comparison of correlations from dependent samples**

Now let's test whether the correlation between English and Reading differs from the correlation between English and Writing.

	n	r
Correlation 1	181	.36
Correlation 2	210	.33
Test Statistic *z*	0.333	
Probability *p*	0.369	

Figure 10.3: Comparison of correlations from independent samples

Notice how we have three tests we are comparing: (1) English, (2) Reading, and (3) Writing. We can't use this test if we are testing the correlation between variables A and B and the correlation between variables C and D. There needs to be overlap.

If you still have your filter on in your dataset from the previous analysis, turn it off. Go to the Data tab, click Filters, and either select the X to delete it or toggle the active button so it's turned off. Return to your Correlation Matrix results and click on it to edit it. Add Writing to the box.

However, we have a problem! The Testing the Significance of Correlations - Dependent Samples webpage (#2) wants a single N, but our correlation matrix has different Ns because of missing data. What can we do? We need to chain filters! Go back to your Data tab, click Filters, and add three filters like below (note: this is how you can get *listwise deletion* in jamovi):

Our correlation matrix should have automatically updated and all the N's equal 370. Great! We now have all the information we need to input into our Testing the Significance of Correlations webpage, #2. For n we input 370. For r_{12} we enter the correlation between (1) English and (2) Reading. For r_{13} we enter the correlation between (1) English and (3) Writing. For r_{23} we enter the correlation between (2) Reading and (3) Writing. Our z-score is not statistically significant ($p = .213$) so there is no significant difference in the correlation between English and Reading ($r = .32$) with the correlation between English and Writing ($r = .37$).

10.1.8 Your turn!

Open the `Sample_Dataset_2014.xlsx` file that we will be using for all Your Turn exercises. You can find the dataset here: Sample_Dataset_2014.xlsx Download

Perform correlations based on the following research questions.

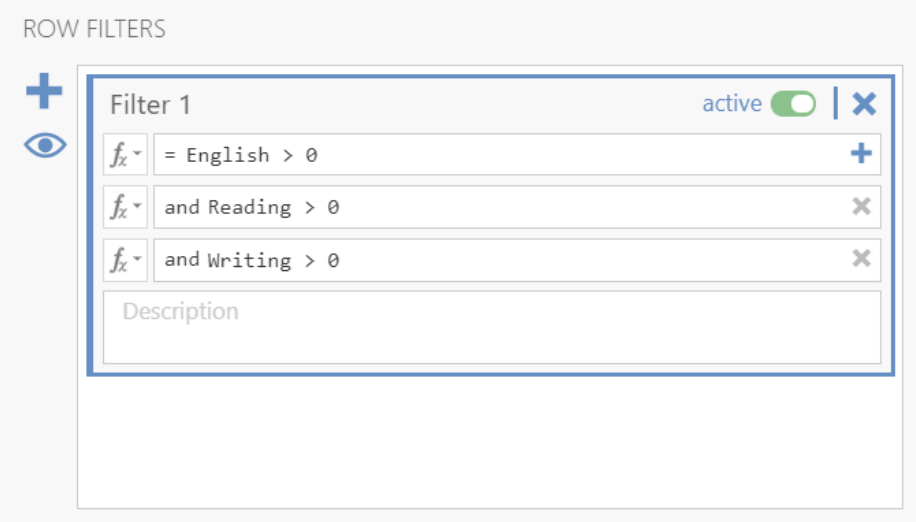


Figure 10.4: Advanced filters

n	r_{12}	r_{13}	r_{23}
370	.32	.37	.13
	Test Statistic *z*	-0.798	
	Propability *p*	0.213	

Figure 10.5: Comparison of correlations from dependent samples

To get the most out of these exercises, try to first find out the answer on your own and then use the drop-down menus to check your answer.

1. **Are there significant correlations among the four tests (English, reading, math, writing)?**
    - Do you meet the assumption of normality for all four tests? yes no yes for all but maybe not writing
    - Do you meet the assumption of linearity for all four tests? yes no
    - Are the four tests significantly correlated among each other? yes no
    - Round your answers to two decimal places:
        - What is the correlation between reading and math?
        - What is the correlation between writing and reading?
        - What is the correlation between writing and English?

10.2 Regression

10.2.1 Overview

Regression can examine multiple predictor variables simultaneously. Whereas the factorial ANOVA can only handle categorical variables (i.e., nominal or ordinal), regression can handle all types of predictor variables including both categorical and continuous.

There are three types of regression in general:

1. Linear regression: this looks at the effect of a single predictor (IV) on a single outcome (DV). This is equivalent to a t-test (dichotomous predictor), one-way ANOVA (ordinal predictor), or correlation (scale predictor).
2. Multiple regression: this looks at the effect of multiple predictors (IVs) on a single outcome (DV).
3. Hierarchical regression: this looks at the effect of multiple predictors (IVs) on a single outcome (DV), but there are multiple "blocks" or "steps" so that you can check the added predictability of new variables.

Note that the linear regression is actually equivalent to a lot of the statistics we've learned. For example, the linear regression will produce the same results as a t-test when we have a dichotomous predictor, a one-way ANOVA when we have an ordinal predictor, and a correlation when we have a continuous predictor. We'll learn more about this at the end of the textbook when we wrap everything up.

10.2.1.1 Understanding regression

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 categorical predictor (whether they are spotted or striped) and a continuous predictor (height) and a continuous dependent variable (weight). 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 two predictor variables (images from Allison Horst).

Figure 10.6: Regression lines and residuals

We determine our line equation from the scatterplot of scores by figuring out the line that fits closest to all data points. The regression line is the line with the *smallest* residuals between the line and data points.

Let's visualize the regression line for how Dan's sleepiness affect Dan's grumpiness. 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.

Let's go back to our dragon example and input one of our dragons into the model to find out how residuals work. On the left, based on our dataset and the fact that our dragon is striped (spotted = 0) and has a height of 5.1 feet, we would expect our dragon to weigh 3.9 tons. However, when we actually weigh him, he weighs 4.2 tons! Therefore, the residual in this case would be .3 tons.

One of our assumption checks is that our residuals are normally distributed, so we would take all our residuals and examine those for normality.

There is more math to regression, which is needed to calculate the *F*-test you get for the overall model test and the t-tests you get for your model coefficients, but we won't get into that detail.

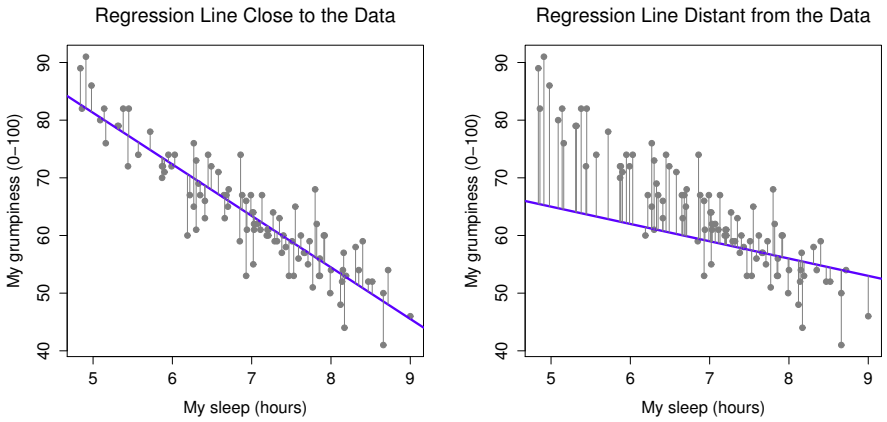


Figure 10.7: Regression lines and residuals

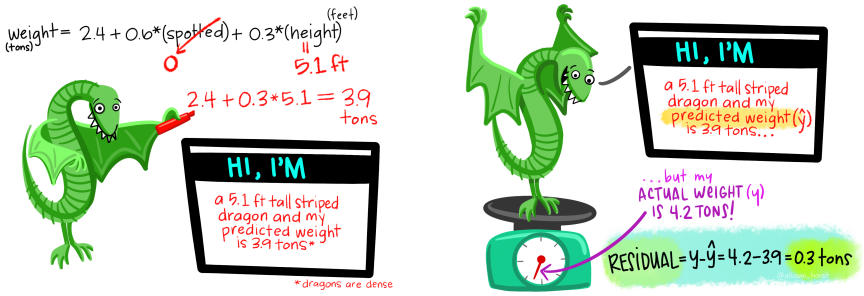


Figure 10.8: Regression lines and residuals

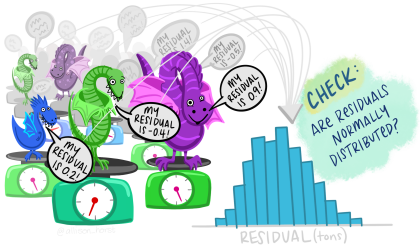


Figure 10.9: Regression lines and residuals

10.2.2 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 `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.

10.2.2.1 Data set-up

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).

For this chapter, we're going to return to the `parenthood` dataset from lsj-data. Remember that 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.

	ID	dan.sleep	baby.sleep	dan.grump	day
1	1	7.59	10.18	56	1
2	2	7.91	11.66	60	2
3	3	5.14	7.92	82	3
4	4	7.71	9.61	55	4
5	5	6.68	9.75	67	5
6	6	5.99	5.04	72	6
7	7	8.19	10.45	53	7
8	8	7.19	8.27	60	8
9	9	7.40	6.06	60	9
10	10	6.58	7.09	71	10

10.2.2.2 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 can see first 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.

We also see skew and kurtosis. Calculating the z-score for all the skew and kurtosis (remember: skew or kurtosis divided by its standard error) suggests we do not violate the assumption of normality much except for `day`. However, notice what the variable `day` is! It's just the day of the study, from 1-100. If you look at the graph, it has a *uniform distribution* (completely flat and uniform) not a normal distribution (bell curve)!

Descriptives

	dan.sleep	baby.sleep	dan.grump	day
N	100	100	100	100
Missing	0	0	0	0
Mean	6.97	8.05	63.71	50.50
Median	7.03	7.95	62.00	50.50
Standard deviation	1.02	2.07	10.05	29.01
Variance	1.03	4.30	101.00	841.67
Minimum	4.84	3.25	41	1
Maximum	9.00	12.07	91	100
Skewness	−0.30	−0.02	0.45	0.00
Std. error skewness	0.24	0.24	0.24	0.24
Kurtosis	−0.65	−0.61	−0.04	−1.20
Std. error kurtosis	0.48	0.48	0.48	0.48
Shapiro-Wilk W	0.98	0.98	0.98	0.95
Shapiro-Wilk p	0.069	0.256	0.122	0.002

10.2.3 Check Assumptions

10.2.3.1 Assumptions

The regression has a lot of assumptions.

Some require no testing:

1. **Variable types**: The DV is continuous and the IVs are either continuous, dichotomous, or ordinal.
2. **Independence**: All the outcome variable values are independent (e.g., come from a separate entity).

Other assumptions require testing:

1. **No outliers**: There shouldn't be any data in the dataset that is an outlier which would strongly influence your results.
2. **Normality of the *residuals***: Up to this point, we've examined the normality of the outcome variables. With regression, our variables can be non-normal as long as the residuals (i.e., error) are normally distributed.
3. **Linearity**: The relationship between each IV and DV is linear. Sometimes you may expect a curvilinear relationship between an IV and DV, in which

case we square or cube the IV and use that variable as our predictor variable in the regression.

4. **Homogeneity of variance (homoscedasticity)**: At each level of the predictor variables, the variance of the residual terms should be constant.
5. **Independent residuals**: For any two observations, the residual terms should be uncorrelated (or independent). Our errors must be normally distributed *and* uncorrelated.
6. **No multicollinearity**: There should be no perfect or near-perfect linear relationship between two or more of the predictors in your regression model. For example, you would not include "heigh_cm" and "heigh_in" in your model because they would be perfectly related to one another. We'll learn how to test for this.

10.2.3.2 Checking Assumptions

10.2.3.2.1 Outliers Under Data Summary, you should have a table with Cook's distance. This is one way we can check for *multivariate outliers*. This examines whether any one *line* of data is an outlier, not just one data *point*. In general, Cook's distances greater than 1 indicate a multivariate outlier. Our Cook's distances are very small, so we do not have a problem with outliers.

Data Summary

Cook's Distance

			Range	
Mean	Median	SD	Min	Max
0.01	0.00	0.02	2.62e−5	0.11

Figure 10.10: Checking for multivariate outliers in jamovi

10.2.3.2.2 Normality of the residuals The regression analysis in jamovi allows us to check normality with the Shapiro-Wilk's test and the Q-Q plot of our residuals. We've seen this multiple times, so by now it should be well-ingrained that because Shapiro-Wilk's is not statistically significant and our data points fall along the diagonal line that we satisfy the assumption of normally distributed residuals.

10.2.3.2.3 Linearity & homoscedasticity To examine linearity and homoscedasticity, two of the assumptions of regression, we examine the Residuals Plots. You will get one plot of the overall model (Fitted) and one for each of your variables (DV and IV(s)). I've displayed the residuals plot for the Fitted

Normality Test (Shapiro-Wilk)

Statistic	p
0.99	0.841

Q-Q Plot

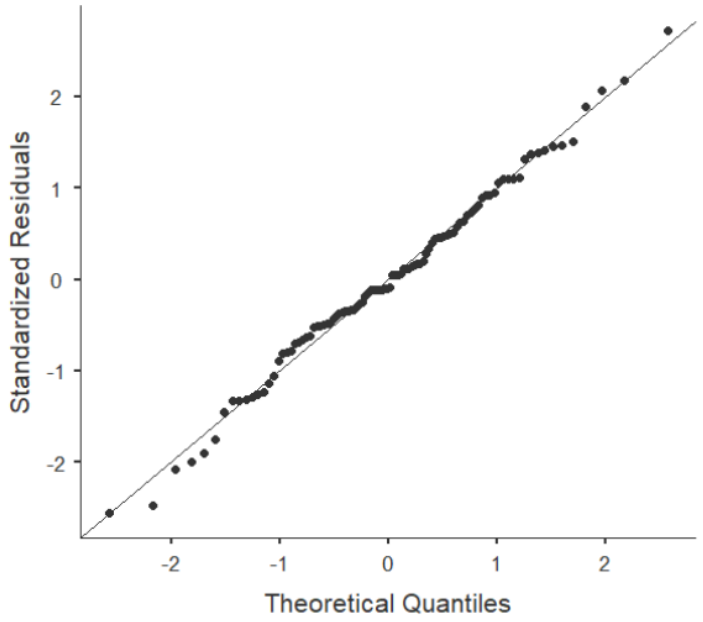


Figure 10.11: Checking the normality of residuals in jamovi

values against the residuals below. In these plots, we want our data to look like a random scattering of dots even dispersed around zero on the y-axis.

Linearity: If the data points seem to have a curve in the graph, then that suggests you have failed the assumption of linearity. Our data doesn't seem to have any curve to it, so we satisfy the assumption of linearity.

Homoscedasticity: If the graph seems to funnel (e.g., widely dispersed on one end of the x-axis and narrowly dispersed on the other end), then that suggests you fail the assumption of homoscedasticity. Our data doesn't seem to be wider at any point, so we satisfy the assumption of homoscedasticity.

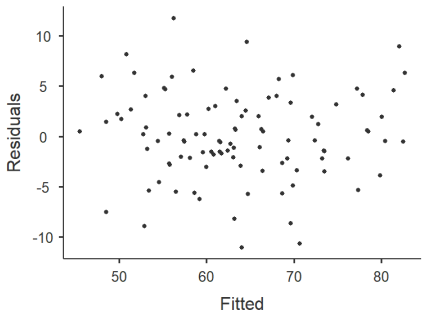


Figure 10.12: Checking linearity and homoscedasticity residuals in jamovi

10.2.3.2.4 Independence of residuals The Durbin-Watson test for autocorrelction tests for independence of residuals. We want the Durbin-Watson value 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, they provide 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
−0.07	2.12	0.518

Figure 10.13: Checking independence of residuals in jamovi

10.2.3.2.5 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 10 or lower, which corresponds to Tolerance values greater than .2.

In our data, our VIF is 1.65 and Tolerance is .61, so we satisfy the assumption of no multicollinearity.

Collinearity Statistics

	VIF	Tolerance
dan.sleep	1.65	0.61
baby.sleep	1.65	0.61

Figure 10.14: Checking multicollinearity in jamovi

Now that we met all the assumptions, we can interpret our results!

10.2.4 Perform the test

1. From the 'Analyses' toolbar select 'Regression' - 'Linear regression'. Note that we select this option regardless of whether we are performing a linear regression, multiple regression, or hierarchical regression.
2. Move your dependent variable `dan.grump` into the Dependent Variable box and all your independent variables into either Covariates (if they are continuous variables) or Factors (if they are categorical variables). In this case, all our variables are continuous so move both `dan.sleep` and `baby.sleep` to the Covariates box.
3. If you are performing a hierarchical regression, you will use the Model Builder drop-down menu. More information on hierarchical regression

will be discussed later.

4. 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 will be discussed later.
5. Under Assumption Checks, check *all* the boxes!
6. Under Model Fit, select `R`, `R-squared`, `Adjusted R-squared`, and `F test`. The other options (AIC, BIC, RMSE) are more useful when we are comparing models and will be discussed later in the Hierarchical regression section.
7. Under Model Coefficients, select `Standardized estimate`.
8. Optionally, you can ask for plots and tables of the estimated marginal means.

I'm not going to show the set-up figure here because there's just too much to show.

10.2.5 Interpret results

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.90	0.82	0.81	215.24	2	97	< .001

Model Coefficients - dan.grump

Predictor	Estimate	SE	t	p	Stand. Estimate
Intercept	125.97	3.04	41.42	< .001	
dan.sleep	−8.95	0.55	−16.17	< .001	−0.90
baby.sleep	0.01	0.27	0.04	0.969	0.00

Figure 10.15: Regression results in jamovi

The first table shows us our overall model results.

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 and the baby's sleep quality predict 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 biased. 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 set of predictors significantly predicts our dependent variable.

Model coefficients: Just like in ANOVA, we first examine if the model is significant (overall model test) and then look at individual factors, in this case being individual variables in our regression model. Each variable–our intercept and both independent variables–have an associated t-test. In this case, Dan's sleep significantly predicts Dan's grumpiness, but the baby's sleep does not.

Standardized coefficients: We also asked for standardized estimates, which we get in the last column of our model coefficients table. These are *standardized* so that we can compare them to other variables. They give us an idea of the *strength* of the relationship between that IV on the DV. Larger values = bigger effects. The standardized estimate is called Beta (β) whereas the unstandardized estimate is just called that or B (the letter B, not Beta). We use the standardized estimates to compare the strength of the estimate to other IVs and we use unstandardized estimates to write our linear equations and predict the DV given values of the IV.

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 is such:

$y = 125.97 - 8.95(dan.sleep) + .01(baby.sleep)$

If Dan's sleep was 5 and baby's sleep was 8, then we'd expect Dan's grumpiness to be:

$y = 125.97 - 8.95(5) + .01(8) = 125.97 - 44.75 + .08 = 81.3$

10.2.5.1 Write up the results in APA style

We can write up our results in APA something like this:

> Dan collected data on how many hours of sleep Dan and Dan's baby got, as well as Dan's grumpiness, for 100 days. Dan tested how the hours of sleep both Dan and the baby got affected Dan's grumpiness using a multiple regression. The combination of predictors were significantly related to Dan's grumpiness, F (2, 97) = 215.24, $p <$

> .001, adjusted R^2 = .81. The number of hours of sleep Dan got significantly predicted Dan's grumpiness, β = -.90, t (97)= -16.17, SE = .55, p < .001. However, the number of hours of sleep the baby got did not significantly relate to Dan's grumpiness, β = 0.00, t (97)= .04, SE = .27, p = .969.

Note that 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.

10.2.6 Additional information

10.2.6.1 Hierarchical regression

Hierarchical regression is exactly the same as multiple regression but now we have multiple models or blocks. You can specify hierarchical regression using the Model Builder drop-down menu in jamovi. Let's try an example where we have `baby.sleep` as Block 1 and `dan.sleep` as Block 2. In addition, using the Model Fit drop-down menu you should check `AIC` and `BIC` in addition to the previously selected options. Your setup should look something like this:

Our model results now change. We now have two lines for the Model Fit Measures and a Model Comparisons table. In addition, under Model Specific Results, we have a drop-down menu to specify which model we want to examine.

Let's interpret. Our first model (with just `baby.sleep` is significant), F (1, 98) = 46.18, p < .001, *adjusted* R^2 = .31. So is our second model (that has both `baby.sleep` and `dan.sleep`), F (2, 97) = 215.24, p < .001, *adjusted* R^2 = .81. There was a significant improvement between model 1 and model 2, F_{change} (1, 97) = 261.52, p < .001, ΔR^2 = .50. The significant improvement means that the predictors added to model 2 significantly predict our DV *above and beyond* the predictors in model 1.

We might write-up these results as such:

> Dan collected data on how many hours of sleep Dan and Dan's baby got, as well as Dan's grumpiness, for 100 days. Dan tested how the hours of sleep both Dan and the baby got affected Dan's grumpiness using hierarchical regression to find out how Dan's sleep predicted Dan's grumpiness above and beyond the baby's sleep.
>
> First, the baby's sleep significantly predicted Dan's grumpiness, F (1, 98) = 46.18, p < .001, *adjusted* R^2 = .31. As the baby's hours of

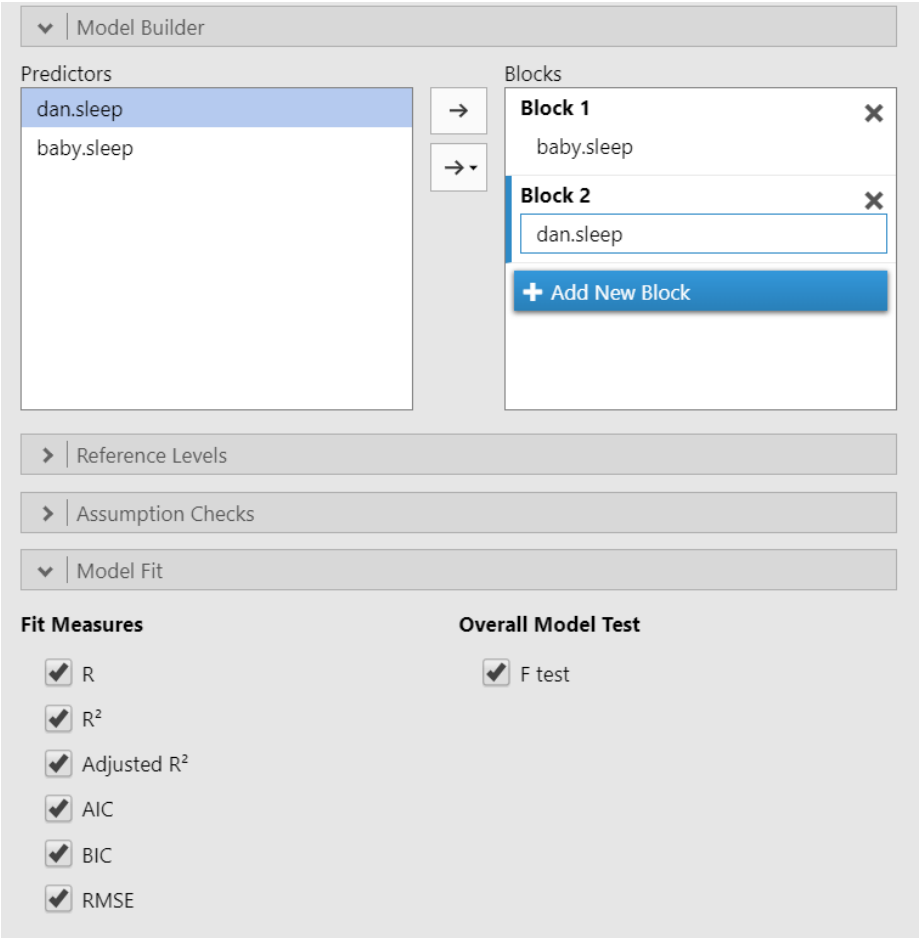


Figure 10.16: Hierarchical regression setup in jamovi

Model Fit Measures

				Overall Model Test			
Model	Adjusted R^2	AIC	BIC	F	df1	df2	p
1	0.31	711.68	719.49	46.18	1	98	< .001
2	0.81	582.95	593.37	215.24	2	97	< .001

Model Comparisons

Comparison							
Model		Model	ΔR^2	F	df1	df2	p
1	-	2	0.50	261.52	1	97	< .001

Figure 10.17: Hierarchical regression results in jamovi

sleep increased, Dan's grumpiness decreased, t (98) = -6.80, SE = .40, $p < .001$, β = -.57.

A second model was tested that added Dan's sleep. This model–comprised of both baby's sleep and Dan's sleep–significantly predicted Dan's grumpiness, F (2, 97) = 215.24, $p < .001$, *adjusted* R^2 = .81. There was a significant improvement between model 1 and model 2, F_{change} (1, 97) = 261.52, $p < .001$, ΔR^2 = .50. In the second model, as Dan's sleep increased, Dan's grumpiness decreased, t (97) = -16.17, SE = .55, $p < .001$, β = -.90. However, the baby's sleep did not significantly relate to Dan's grumpiness when controlling for Dan's sleep, t (97) = .04, SE = .27, p = .969, β = .00.

10.2.6.2 Categorical Predictors

Dummy coded variables (with values 0 or 1) are pretty easy to interpret in regression. If the Beta is positive, then the value of 1 would have a higher mean on the DV than the value of 0. If the Beta is negative, then the value of 0 would have a higher mean on the DV than the value of 1.

However, if we have a nominal variable with more than two categories, then we need to dummy code the data to analyze in a regression. Fortunately, jamovi can do this automatically for us!

The dataset we're using doesn't currently have a categorical variable, so I'm going to manually create one for demonstration purposes. I'm going to transform the `day` variable, which is the day of the data collection from 1 to 100, into a new variable that indicates whether the day is 1-32, 33-65, or 66-100, which is

roughly 3 equal groups. You can see the transformation here:

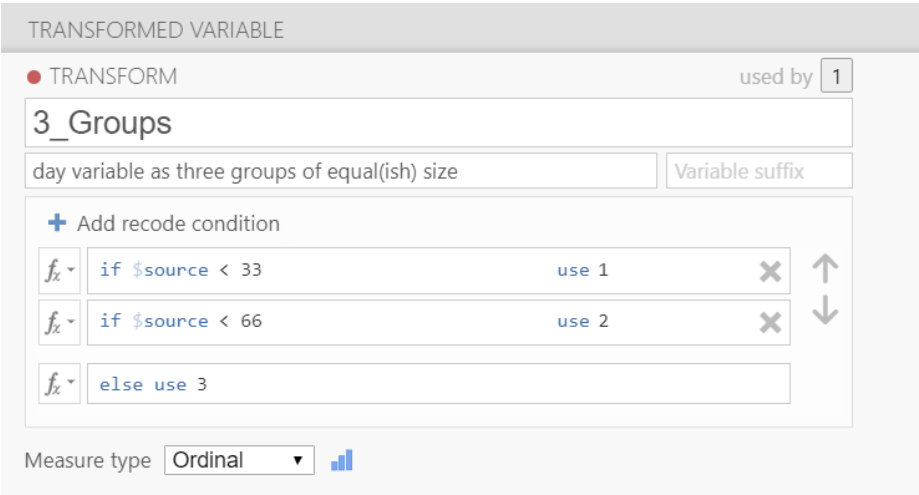


Figure 10.18: Transforming day into a categorical variable in jamovi

Let's add that to our regression model and just make it a simple multiple regression model. Three independent variables (`dan.sleep`, `baby.sleep`, and our new `day_3groups` variable) all in one block.

Now we need to go to the Reference Levels drop-down menu. We have two options:

1. **Reference level (dummy coding)**: We can have our intercept be the mean of our reference level group, meaning that if all other variables were set to 0 this is the mean of our dependent variable for that group. For example, if we set day = 1 to be our reference level, then the intercept is the value of Dan's grumpiness when Dan's sleep is 0 and baby's sleep is 0 for the first 32 days. *This is the option I normally choose.*
2. **Grand mean (simple coding)**: Alternatively, we can have our intercept be the grand mean, or the overall mean when all other variables were set to 0 and we ignored day. *I am not sure when I would use this option, to be honest.*

The other option you have is what is considered your reference level. It will default to your first level in your dataset (in this case, day_3Groups = 1) but you can change to any other level in your variable. I set my reference level to be 1, the default, and I know that because the day variable compares both levels 2 and 3 to 1. Our intercept (126.02) then is the value of grumpiness if Dan and the baby slept 0 hours in the first 32 days of our data collection.

The first line of day - 3_Groups (2 – 1) is then the difference in Dan's grumpiness between the second 1/3 of days (days 33-65) and the first 1/3 of days (days 1-32).

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.90	0.82	0.81	107.29	4	95	< .001

Model Coefficients - dan.grump

Predictor	Estimate	SE	t	p	Stand. Estimate
Intercept [a]	126.02	3.16	39.83	< .001	
dan.sleep	−8.87	0.56	−15.72	< .001	−0.90
baby.sleep	−0.02	0.27	−0.06	0.949	−0.00
day - 3_Groups:					
2 – 1	−1.12	1.08	−1.03	0.305	−0.11
3 – 1	−0.03	1.08	−0.03	0.978	−0.00

[a] Represents reference level

Figure 10.19: Regression with a categorical predictor results in jamovi

It is not statistically significant, so there is no difference in Dan's grumpiness between the first and second 1/3 of days. Because the estimate is negative, that indicates that the first 1/3 of days have a higher estimated mean of Dan's grumpiness, but again it's not statistically significant.

The second line of day - 3_Groups (3 – 1) is the difference in Dan's grumpiness between the third 1/3 of days (days 66-100) and the first 1/3 of days (days 1-32). It is not statistically significant, so there is no difference in Dan's grumpiness between the first and third 1/3 of days, either.

In this case, the Estimated Marginal Means can be very helpful for us to interpret the model coefficients. We can get the estimated marginal means of each group on the DV at the average levels of the other two variables.

10.2.7 Your turn!

Open the `Sample_Dataset_2014.xlsx` file that we will be using for all Your Turn exercises. You can find the dataset here: Sample_Dataset_2014.xlsx Download

To get the most out of these exercises, try to first find out the answer on your own and then use the drop-down menus to check your answer.

1. **Perform a multiple regression examining how `English`, `Reading` and `Writing`, as well as `Gender` relate to the dependent variable `Math`.**

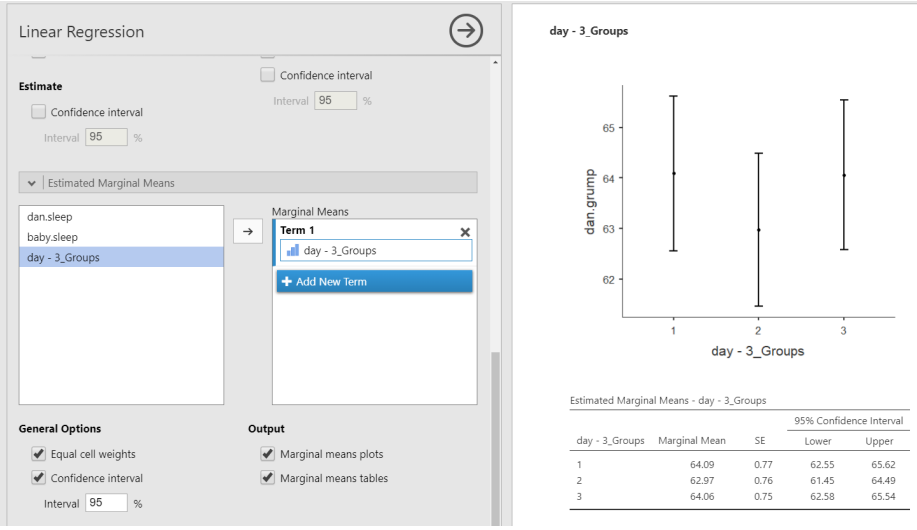


Figure 10.20: Regression with a categorical predictor results in jamovi

- Do you have any significant outliers? yes no
- Are your residuals normally distributed? yes no
- Do you satisfy the assumption of linearity and homoscedasticity of your residuals (just check the Fitted residual plot)? yes no
- Do you meet the assumption of independent residuals? yes no
- Do you meet the assumption of no multicollinearity? yes no
- Can you perform a regression with this data? yes no
- What is your adjusted R-squared, rounded to two decimal places:
- Is the overall model statistically significant? yes no
- Is `English` statistically significant? yes no
- Is `Reading` statistically significant? yes no
- Is `Writing` statistically significant? yes no
- Is `Gender` statistically significant? yes no
- For `Gender`, do male (Gender = 0) or female (Gender = 1) students have higher math scores? male female

10.3 General Linear Model

10.3.1 Overview

The regression is a General Linear Model (GLM). Everything we've learned up to this point is also a general linear model. Pretty much everything we've learned in this class *could* be performed as simple a regression. You may wonder why we have not just taught regression and none of the others. There are some who are indeed proponents of that! However, I believe teaching "simpler" statistics like the t-test and correlation first is easiest to understand. Most of the statistics you will perform from here on out—including in your careers—will be what we have already learned (i.e., t-test, ANOVA, chi-square, correlation). However, I present this information so you may begin to see how all of this is related.

If you'd like to learn more about this, there's a fantastic online book on the subject called Common statistical tests are linear models (or: how to teach stats)

Note that for all these examples I am using the Sample_Dataset_2014.sav dataset and I am only presenting the relevant output. I am also only going to cover the correlation, t-tests, and one-way ANOVA because they are the simplest to compare. You can run these yourself! You can find the dataset here: Sample_Dataset_2014.xlsx Download

10.3.2 Correlation as a regression

Imagine we want to see how English and Reading are related. We would do a Pearson correlation, as seen on the top. However, we could also run a simple regression with English predicting Reading or vice versa, as seen on the bottom.

Notice how the Beta coefficient in the regression output is the *standardized coefficient* and that the correlation is the *standardized covariance*. Therefore, their values (and p-values) match identically!

10.3.3 Independent t-test as a regression

Next let's look at how gender is related to reading scores. First, our t-statistic and p-value match directly from the independent t-test to the linear regression coefficient. Second, our unstandardized estimate in the regression *is exactly the mean difference between the two genders*!

10.3.4 Dependent t-test as a regression

The dependent t-test can't be performed as a regression in the base jamovi, but it can if we run it in the Rj editor using the stats package. The regression formula for a dependent t-test is $y_1 - y_0 = 1$ and jamovi doesn't like it if there are not IVs, just an intercept. However, the stats package doesn't mind! Notice

Correlation Matrix

Correlation Matrix

		English	Reading
English	Pearson's r	—	
	p-value	—	
Reading	Pearson's r	0.33	—
	p-value	< .001	—

Linear Regression

Model Fit Measures

Model	R	R^2
1	0.33	0.11

Model Coefficients - Reading

Predictor	Estimate	SE	t	p	Stand. Estimate
Intercept	51.53	4.39	11.75	< .001	
English	0.37	0.05	7.04	< .001	0.33

Figure 10.21: Correlation as regression

Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p	Mean difference	SE difference
Reading	Student's t	−0.41	414.00	0.681	−0.31	0.75

Linear Regression

Model Fit Measures

Model	R	R^2
1	0.02	4.09e−4

Model Coefficients - Reading

Predictor	Estimate	SE	t	p
Intercept [a]	82.14	0.52	158.42	< .001
Gender:				
0 – 1	−0.31	0.75	−0.41	0.681

[a] Represents reference level

Figure 10.22: Independent t-test as regression

that our t-statistic, df, and p-value are exactly the same, and that the intercept estimate and SE match the mean difference in the paired samples t-test.

Paired Samples T-Test

Paired Samples T-Test

			statistic	df	p	Mean difference	SE difference
English	Reading	Student's t	1.12	398.00	0.262	0.47	0.42

Rj Editor

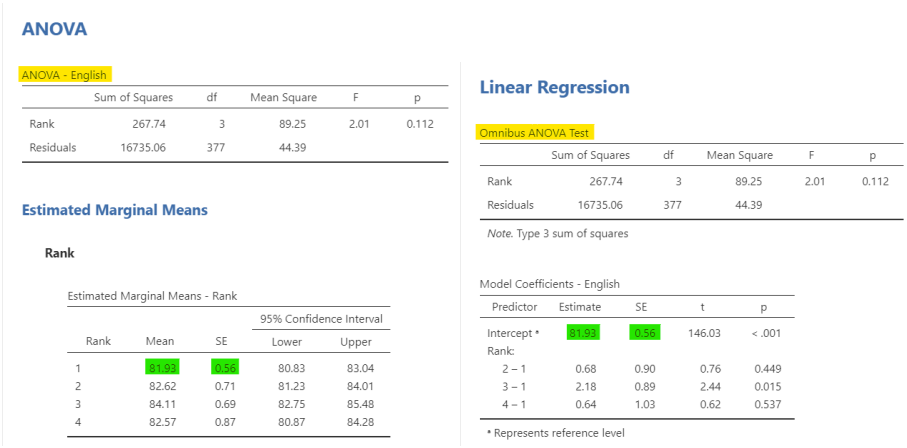
```
> model <- lm(English - Reading ~ 1, data = data)
> summary(model)


Call:
lm(formula = English - Reading ~ 1, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-26.0807  -5.2757  -0.1007   5.5243  27.0293

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4707     0.4187   1.124    0.262

Residual standard error: 8.363 on 398 degrees of freedom
  (36 observations deleted due to missingness)
```

Figure 10.23: Dependent t-test as regression

10.3.5 One-way ANOVA as a regression

The one-way ANOVA is the same as regression, too. Let's examine how rank (an ordinal variable from 1-4) relates to English scores. In the linear regression, we asked for the ANOVA test under Model Coefficients; notice how it directly matches the ANOVA table.

Although I did not directly ask for them, the estimated marginal means for the linear regression match that of the one-way ANOVA. However, I want to call attention to how the coefficients directly match. The intercept is the average English score when Rank = 1 (freshman). However, the estimate for 2 – 1 is .68, but 81.93 + .68 = 82.61; it's not exact due to rounding errors, but it matches the mean for group 2 because in the linear equation if you set that value to 1 you are saying the other values are 0 so the group membership is Rank = 2 (sophomore). You can do this for all the groups and see how they match.



ANOVA

ANOVA - English

	Sum of Squares	df	Mean Square	F	p
Rank	267.74	3	89.25	2.01	0.112
Residuals	16735.06	377	44.39		

Estimated Marginal Means

Rank

Estimated Marginal Means - Rank

			95% Confidence Interval	
Rank	Mean	SE	Lower	Upper
1	81.93	0.56	80.83	83.04
2	82.62	0.71	81.23	84.01
3	84.11	0.69	82.75	85.48
4	82.57	0.87	80.87	84.28

Linear Regression

Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
Rank	267.74	3	89.25	2.01	0.112
Residuals	16735.06	377	44.39		

Note. Type 3 sum of squares

Model Coefficients - English

Predictor	Estimate	SE	t	p
Intercept *	81.93	0.56	146.03	< .001
Rank:				
2 – 1	0.68	0.90	0.76	0.449
3 – 1	2.18	0.89	2.44	0.015
4 – 1	0.64	1.03	0.62	0.537

* Represents reference level

Figure 10.24: One-way ANOVA as a regression

Chapter 11

References