Beginner Statistics for Psychology

Beginner Statistics for Psychology

An unintimidating guide to basic hypothesis testing logic for beginners

VANCOUVER, BC

Contents

Introduction

This text is an informal beginner's guide to the basics of statistics needed to conduct tests of statistical significance for simple experimental, quasi-experimental, and correlational research designs. Many free and open textbooks exist for more theoretical and traditional versions of introductory statistics, with more standard notation. This one is designed to be consistent with the approachable way I aim to teach the use of statistical analysis tools to students of psychology and the social sciences. The text is very informal and conversational, because it spawned from my written-out scripts to record brief video lessons.

Over the years, many students have been surprised by the fact that they actually enjoy learning statistics when presented simply, as a decision-making tool, and supported by engaging examples. If you are a student approaching statistics for the first time, I hope that is your experience with this resource.

In general in this textbook, I choose conceptual formulas with simplified notation. Learning to calculate by hand with less-intimidating formulas is intended to lead you through practice to conceptual inference. Much evidence in the literature of teaching statistics supports this approach. Although some students of statistics are fortunate enough to make mental leaps straight to the abstract, most of us require a more hands-on approach with several examples to make the key cognitive connections.

The material here is of my own creation, with some images and resources borrowed from other open-source materials. Please use this resource if you like. I do appreciate attribution, because it took me a long time to collect and develop some of these explanations, analogies, examples, and flow of logic.

First edition note: This book is bare basics so far. Some day I will hopefully add in the interactive exercises and demonstrations my students use to apply concepts and procedures as we go. My dream is to build exercises and examples for this simplified notation style and also a more traditional style, so students can see they are equivalent. I also hope someday to rework the videos I have created as a companion so they are suitable for embedding in the text. Corrections and suggestions for revisions are welcome.

Dedication: I dedicate this project to Dr. Bryan Hendricks, statistics prof extraordinaire, who worked for years in the University of Wisconsin system. As his teaching assistant, I learned from a master. Rest in peace, Dr. H.

Table of Contents and Learning Objectives

Table of Contents

Learning Objectives

Chapter 1 Why We Need Statistics and Displaying Data Using Tables and Graphs

- articulate the purpose of a course introducing statistical principles and techniques
- supply examples of situations in which data analysis techniques may be necessary
- define descriptive and inferential statistics, variable, value, and score
- distinguish between two levels of measurement and identify the appropriate techniques for summarizing different types of data
- generate frequency tables
- graph a dataset using a histogram, bar graph, or pie chart
- describe a distribution shape in terms of peaks and symmetry

Chapter 2 Central Tendency and Variability

- define and determine mean, median, and mode, as three options to determine central tendency

- distinguish among the measures of central tendency and the circumstances under which each is suitable
- define and determine variance and standard deviation, as two options to determine variability
- interpret standard deviation

Chapter 3 Z-scores and the Normal Curve

- transform scores in any numeric dataset, using any scale, into the standard metric of Z-scores
- interpret Z-scores and apply them for comparison of scores within and between datasets, including data measured on different scales
- define and characterize the normal curve model
- associate Z-scores with areas under the normal curve
- define percentiles and determine Z-scores and raw scores that form the border of percentiles using the normal curve model

Chapter 4 Probability, Inferential Statistics, and Hypothesis Testing

- determine simple probabilities
- appreciate the importance of probability and ubiquity of human failings in the realm of probability
- connect probability to percentiles, areas under the normal curve, and the logic of inferential statistics such as hypothesis testing
- define and distinguish between population and sample
- articulate the central tendency theorem and describe its implications for the normality assumption in inferential statistics
- outline and apply the steps of hypothesis testing

Chapter 5 Single Sample Z-test and t-test

- define and identify Type I and Type II errors
- define and characterize the distribution of means as compared to the distribution of individuals
- determine the mean and standard deviation of the distribution of means based on the characteristics of the distribution of individuals
- conduct a hypothesis test using the single sample Z-test
- define, determine, and interpret a p-value

- articulate a conclusion in plain language from an test of statistical significance
- define and determine degrees of freedom
- articulate the logic behind the sample size correction for sample-based estimates of variance
- describe the difference between t-distribution shapes with varying degrees of freedom
- conduct a hypothesis test using the single sample t-test
- identify scenarios in which a single sample Z-test or t-test is appropriate

Chapter 6 Dependent t-test

- identify and describe repeated measures and matched pairs research designs
- conduct a hypothesis test using the dependent means t-test
- identify scenarios in which a dependent means t-test is appropriate

Chapter 7 Independent Means t-test

- identify and describe classical experimental research designs
- identify the (normal curve and homoscedasticity) assumptions behind the independent means t-test
- conduct a hypothesis test using the independent means t-test
- identify scenarios in which an independent means t-test is appropriate

Chapter 8 Analysis of Variance, Planned Contrasts and Posthoc Tests

- define partitioning of variance and apply the concept to one-way Analysis of Variance
- define and identify factors and levels in research designs
- use graphing techniques to visualize data from a research design using more than 2 levels in a factor
- conduct a hypothesis test using one-way Analysis of Variance
- articulate reasons for conducting planned contrasts or post-hoc tests following ANOVA
- define experimentwise alpha level and articulate ways in which Bonferroni and Scheffé corrections address inflated risk of Type I

error
- outline the procedure for conducting planned contrasts with Bonferroni correction
- outline the procedure for conducting posthoc tests with Scheffé correction
- identify scenarios in which a one-way ANOVA is appropriate

Chapter 9 Factorial ANOVA and Interaction Effects

- apply the concept of partitioning of variance to two-way Analysis of Variance
- describe factorial analysis and articulate its benefits and pitfalls
- describe research designs using ___ X ___ factor and level summaries
- conduct a hypothesis test using two-way Analysis of Variance
- identify scenarios in which a two-way ANOVA is appropriate
- identify and interpret main effects
- identify and interpret interactions

Chapter 10 Correlation and Regression

- define correlation and regression
- detect and describe linear correlation patterns using scatterplots
- define partitioning of covariance
- conduct a hypothesis test using correlation
- find the proportion of variance explained by a correlation
- identify scenarios in which a correlation is appropriate
- create a predictive model using a simple regression line
- articulate limits to accuracy and usefulness of regression models

Chapter 11 Beyond Hypothesis Testing

- define effect size, power, and confidence intervals
- articulate the importance of effect size and power analyses
- find and interpret Cohen's d for a single-sample Z-test scenario
- identify the two major determinants of statistical power
- estimate and interpret power for a single-sample Z-test scenario
- construct confidence intervals for a single-sample Z-test scenario
- articulate similarities and differences between hypothesis testing and confidence interval procedures

PART I

MAIN BODY

1. Why We Need Statistics and Displaying Data Using Tables and Graphs

1a. Why we need statistics

One of the first things I think we need to accomplish in this course is to understand why statistics are important. Our objective in this first part of Chapter 1 is to be able to articulate the purpose of a course introducing statistical principles and techniques, and to be able to supply examples of situations in which the techniques you will learn in such a course may be necessary to use.

First, let us establish that this is *not* a math course. This is a course that is primarily about decision making. Not just any decision making, but decisions that are made after analyzing data in order to make objective decisions that are guided by empirical evidence. Of course, we use some simple calculations in the course in order to process the data into a form that aids our decision making. However, the math is a necessary means to an end, not an end in itself.

In some situations this kind of decision making is not needed. When the decision can be made based on intuition and subjective personal preference, we do not need rigorous data-driven systems. For example, if I am trying to decide whom to date, or what style of clothing I like to wear that suits my personality, I likely am not going to conduct research and a formal data analysis to come to those decisions. Maybe you can think of another situation, in which a good decision can be made without empirical evidence.

On the other hand, sometimes a decision that you need to make is one that affects others, or is so high stakes that you want to make an informed decision that is objective and based on empirical evidence. In this kind of decision making, you should check your intuition at the door, and walk in with an open mind, letting the data be your guide. Examples of situations in which an objective decision making process might be necessary would be when you are trying to decide whether

a medical treatment is safe, or whether a proposed intervention is actually effective. Perhaps you need to find out if a crime prevention program is effective for urban and rural communities alike. Can you think of another kind of decision that should be made objectively based on data? What these scenarios have in common is that they are professional decisions, or are high stakes. In the professional workplace, we are often in situations where, if we just operated based on our intuition, we may make serious mistakes, because we have not considered whether the course of action we decide on is the best choice for all people, all situations, or over time. The techniques you will learn in this course will help you apply data analysis, so that you can set up a decision making framework that is objective and rigorous, and so that the decision you come to will be generalizable, to suit other people, situations, or time frames.

Why does a student in your field of study require statistics? Regardless of your field of study, I bet you are asked to be a critical thinker. If we look at the list of critical thinking guidelines below that make for good science, I bet you can see the value of these guidelines for your own program of study.

Critical thinking guidelines

- Ask Questions: Be Willing to Wonder
- Define Your Terms
- **Examine the Evidence**
- Analyze Assumptions and Biases
- **Avoid Emotional Reasoning**
- Don't Oversimplify
- Consider Other Interpretations
- Tolerate Uncertainty

from Wade, Tavris & Swinkels. (2017). *Psychology*. Boston: Pearson.

Statistics represents a tool for examining evidence and allowing us to use data effectively. However, it is also important to realize that statistics can help us avoid emotional reasoning. Instead of relying on our intuitions about whether a drug is effective, or whether one choice is significantly better than another, statistical analysis allows us to make an objective decision.

In statistics, N stands for sample size. In other words, how many data points did you measure. Very often, in everyday life, we are tempted to make assumptions and derive conclusions from single data points. In the world of statistics, we call these situations, "an N of one". These are situations scientists are extremely wary of, because they are vulnerable to bias.

For example, let's say my friend has a really bad experience in one neighborhood. After that, even if there are no objective reports of comparative neighborhood safety that support this conclusion, I am likely to say to others that that's a bad neighborhood – one to avoid. We are always overly influenced by our own experiences and the experiences of those close to us. In such moments we should always remind ourselves that until we have asked many individuals who have been in that neighborhood what their experiences were, we only have one observation, and it may not be typical or representative. If my friend's experience in the neighborhood were the one bad experience in 1000 experiences, would we still be tempted to consider it a "bad" neighbourhood? Next time you face a situation like this in your daily life, just take a moment to pause and think to yourself... what information should I have to make the right decision?

Fig. 2.2 from *Pulling Together: A Guide for Front-Line Staff, Student Services, and Advisors* by Ian Cull, Robert L. A. Hancock, Stephanie McKeown, Michelle Pidgeon, and Adrienne Vedan

In this course, we will be focusing on only one aspect of one way of knowing. Let us acknowledge the fact that various cultures and systems place particular value on various ways of knowing. For example, if we refer to the indigenous ways of knowing framework shown above, we might see this entire course as being one element of "intellectual" ways of knowing. Its contribution might be to contribute to responsibility and relevance by enhancing the generalizability of decision making as we discussed before. However, no one should mistake statistics for a holistic system of knowing.

I encourage you to think of what you learn in this course as one tool in the toolbox. The reason many academic disciplines require a statistics course is that this is a tool most people do not get in other

areas of their lives. We tend not to learn statistics from our parents or by volunteering in the community. In fact, most of us are very bad at this form of decision making until we learn to use these tools.

By requiring you to learn statistics, disciplines like Psychology are not suggesting it is the only important decision-making tool. It is one we think you need to understand to be a good scientist and to better interpret some types of evidence to which you will have access in your professional life. I encourage you to learn more about holistic ways of knowing and to reflect on the place that formal, data-driven decision-making practices have in your own ways-of-knowing framework. I think we could all gain some insight by looking at such a model with an eye toward acknowledging areas in which we are weaker, because of our own individual experiences or because of the society in which we have grown up.

A video element has been excluded from this version of the text. You can watch it online here: https://pressbooks.bccampus.ca/statspsych/?p=5

1b. Displaying Data Using Tables and Graphs

Have you ever heard the saying, "a picture is worth a thousand words"? That is what the rest of this chapter is all about. First, we need to cover some basic concepts and definitions, including the differences between **descriptive** and **inferential** statistics, and the meaning of the terms **variable**, **value** and **score**. We will then need to learn to distinguish among levels of measurement to be able to choose the appropriate techniques for summarizing different types of data.

Finally, I will demonstrate how to generate **frequency tables** and to graph a dataset, because the first step in data analysis is always to look at it. Just as a picture is worth a thousand words, it is also worth a thousand numbers.

At first we will focus on **descriptive** statistics. These are ways to summarize or organize data from a research study – essentially allowing us to describe what the data are.

A little later in the course, we will move into the realm of **inferential** statistics. These are analytical tools that allow us to draw conclusions based on data from a research study. In other words, we go beyond just saying what the data are, and make a statement about what they mean. Inferential statistics are used in research and policy as a tool to make decisions.

Three basic terms are essential jargon in statistics. A **variable** is a quality or a quantity that is different for different individuals. A **variable** could be a quality, like ethnicity, for which each person might have a different characteristic. Or it could be a quantity, like temperature, that could be different each time you take a reading, and is measured on a number scale. A **value** is just any possible number or category that a **variable** could take on. So for ethnicity you might have 6 categories in which you place individuals. Or for temperature there might be a numeric range from -100 to +100. Those would be the full set of **values** for that **variable**. A score is a particular individual's **value** on the **variable**. For ethnicity, you would identify yourself as one particular category, and that would be your **score**. For temperature, if you check your weather app and see that it is 7 degrees outside, that is the **score** for that time and place.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=5#h5p-1

Measurement is the assignment of a number to the amount of something., or assigning labels for categories. This is often obvious (for example, we might measure time as number of seconds or number of minutes). Sometimes, however, it can be a bit more arbitrary. We might assign numbers to signify a category, for example 1 for male and 2 for female.

Based on how we measure them, there are two major types of

variables in statistics, and these will be important to keep in mind as we go through the semester. The type of variable determines how we can use it.

Type of variable	*Characteristics*	*Examples*
Nominal/ Categorical	·Label and categorize ·If numbered, numbers are arbitrary	·Gender ·Diagnosis ·Experimental or Control
Numeric/ Quantitative	·Numerical data ·Numbers reflect size or amount of something	·Temperature ·IQ ·Golf scores (above/ below par) ·Number of correct answers ·Time to complete task ·Gain in height since last year

The first type of variable is **nominal** or categorical (also called qualitative). **Nominal** variables label or categorize something, and any numbers used to measure these variables are arbitrary and do not indicate quantity or size. For example, if male is scored as 1 and female is scored as 2, that does not indicate that females are twice as good or double the size of males. It is just a code.

Numeric or quantitative variables are ones for which numbers are actually meaningful. They indicate the size or amount of something.

Examples of Numeric variables

- Temperature, in which 10 degrees is warmer than 0
- Golf scores, in which 2 below par means you did well
- IQ, in which 100 is average intelligence

- Number of correct answers, in which 4 correct answers is twice as many as 2 correct answers

When we calculate statistics, we will see that we can calculate an average IQ in a group of people, or an average temperature across several days. But we cannot calculate an average gender. What we will do is use groupings or categories as a basis for comparison of other variables; for example, does the experimental group have a higher number of correct answers than the control group?

Now, a quick side note: If you take a course in research methods, you learn that measurement is a really tricky thing in practice, particularly when you want to measure something internal about a person. The process of operationally defining something so that you can measure it numerically or in discrete categories is a real challenge. This is beyond the scope of this course, but just to give you a sense, try brainstorming a way in which you could measure aggression? Think of at least one way that would create a **nominal variable**, and one way that would create a **numeric variable**.

If you give that example some thought, you will quickly find that a relatively simple variable like aggression can be fiendishly difficult to measure, and in the field of psychology a lot of effort is put into developing good ways to measure mental constructs. In experimental psychology, we often prefer to measure things as numbers, because then we can use statistical methods to summarize and to make inferences about the thing we measured.

We should return to our discussion of experimental research design. A **variable** is something that has different values for different individuals, and that we can measure. As an example, we can measure how fast someone is at completing a puzzle, and get those scores for a bunch of people. This variable would be speed. We can also assign each of those people into categories or conditions: a high-stress vs. low-stress condition, for example. Research is the study of the relationship between **variables**. Therefore, there must be at least two **variables** in a research study (or there is no relationship to study). Typically an experimental study in psychology has one (or more) **independent variables** and one (or more) **dependent variables**.

An **independent variable** is one you manipulate — most often it is categorical, or **nominal** (e.g. experimental group vs. control group). A **dependent variable** is one you measure to detect a difference/change as a result of the manipulation — most often it is **numeric** (e.g. time to complete a puzzle).

Example of Experimental Design

Do members of your experimental group (who were required to give a speech in front of a group of people) solve a puzzle in a shorter or longer amount of time than members of your control group (who were allowed to browse magazines)?

In the example above, the **independent variable** would be the manipulation: whether people are required to give a speech or are allowed to browse magazines. Note that is a **nominal variable**. The **dependent variable** is what you measure after the manipulation: how it takes the participants to solve a puzzle. Note that would be a **numeric variable**.

Now that you have some basic definitions and concepts down regarding types of data and how to measure them, we need to learn how to deal with **numeric** data.

Example of Numeric Dataset

Stress ratings of 10 students: 8,7,4,10,8,6,8,9,9,7

First… what can you say about this data set from the list of numbers above? How would you describe the findings to someone?

Perhaps you want to summarize a dataset in table form, to organize the data and make it easy to get an overview of the dataset quickly. A **frequency table** does just that. To create a **frequency table**, you just ask yourself: for each possible value on this variable, how many individuals have a particular score? That gives you the frequency of each value – or how often it occurred in the dataset. Let's look at an example. We measure the stress levels of 10 students, on a scale of 1 to 10, and above are their scores. Hard to make any sense out of that list, right? By following the steps below, we can create a **frequency table**.

Steps for Making a Frequency Table

- Label the first row: Values, Frequency, and Percentage.
- In the first column, under the heading Values, list all the possible values the variable could take on. In this case, we have 10 possible values, so there should be 10 rows in the data portion of the table.
- Make a list down the page of each score, from lowest to highest, to make it easier to count them.
- Go one by one through the scores, making a mark for each next to its value on the list (e.g., how many 1's are there? 0. … How many 4's are there? 1. … How many 7's are there? 2. Repeat that question for every value from 1 to 10. Write those frequencies, or counts, in the Frequency column.
- Figure the percentage of scores for each value. To calculate a percentage you take the frequency,

divide by how many scores you have in the dataset (here we have 10 students, so 10 scores), and multiply that by 100 to move the decimal to the right two places. So for the value of 7, with frequency of 2, that becomes 2 divided by 10 times 100 or 20%. Calculate and list all the percentages.

Here is what the table should look like once you are done with those steps:

Values	*Frequency*	*Percent*
1	0	0%
2	0	0%
3	0	0%
4	1	10%
5	0	0%
6	1	10%
7	2	20%
8	3	30%
9	2	20%
10	1	10%

Now you can scan down the table and quickly see where most of the **scores** fall within the range of possible **values**. Now that you have an organized summary of the data, you can clearly see that the majority of students are reporting fairly high stress **scores**. By looking at the percentages, you have a quick way to report the proportion of students that are highly stressed. For example, just by doing some quick addition, you can say that 60% of surveyed students report stress levels 8 or higher.

Most people find graphs easier to interpret at first glance than tables. What can you say about the dataset after looking at this graph?

I bet you were able to say that most **scores** pile up at the upper end of the graph, at the higher end of the range of stress score **values**.

The graphical version of a frequency table is a **histogram**. The X axis on a **histogram** should have the values of the variable listed, from lowest to highest. The Y axis should represent the frequencies of each value in the dataset. In other words, the **histogram** is a **frequency table** that has been turned on its side. The added benefit comes from the visual representation of the frequency as the height of the bars in the graph, rather than just a number. You can thus see a clear shape in the dataset.

In some circumstances, a **frequency table** is not an effective way to summarize a dataset. This is the case if the range of **values** is too large. For example, what if you were summarizing temperature **scores**, which can range from 0-100? This would mean more than 100 rows in the table. That is not so helpful. In such a case, a **grouped frequency table** is a much better option.

A **grouped frequency table** defines ranges of values in the first column, and reports the frequency of **scores** that fall within each range. In this example, we have surveyed 30 students' stress levels on a scale of 1-10:

Example of Numeric Dataset

Stress ratings of 30 students:
8,7,4,10,8,6,8,9,9,7,3,7,6,5,1,9,10,7,7,3,6,7,5,2,1,6,7,10,8,8

If we wanted to get the table into the ideal format of 4-8 rows, we could create grouped frequencies, with two **values** in each row.

By following these steps, we can create a **grouped frequency table**:

Steps for Making a Grouped Frequency Table

- Label the first row: Values, Frequency, and Percentage
- Decide on the ranges of values you need. You want to choose ranges that will leave you with 4-8 rows in the table
- In the first column, under the heading Values, list all the possible value ranges the variable cold take on. In this case, we grouping by twos, so there should be 5 rows in the data portion of the table.
- Make a list down the page of each score, from lowest to highest, to make it easier to count them.
- Go one by one through the scores, making a mark for each next to its value range on the list (e.g., how many 1's and 2's are there? 3. … How many 3's and 4's are there? 3. … and so on. Repeat

that question for every value range. Write those frequencies, or counts, in the Frequency column.

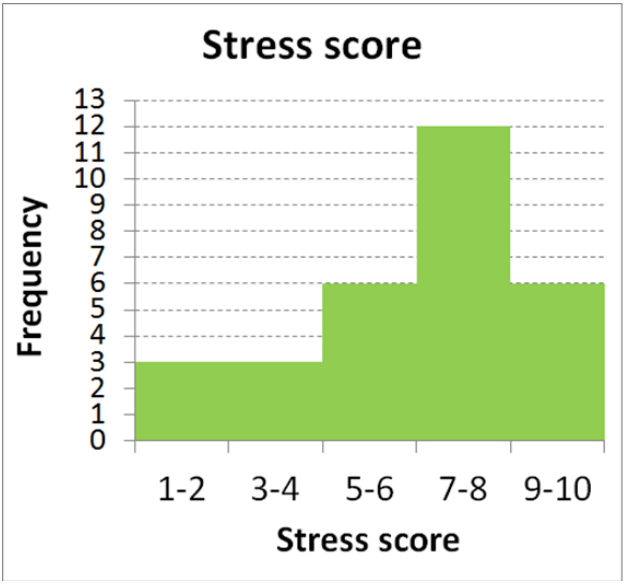
- Figure the percentage of scores for each value range. To calculate a percentage you take the frequency, divide by how many scores you have in the dataset (here we have 30 students, so 30 scores), and multiply that by 100 to move the decimal to the right two places. So for the value range of 1-2, with frequency of 3, that becomes 3 divided by 30 times 100, or 10%. Calculate and list all the percentages.

Here is the completed **grouped frequency table** for the dataset of 30 students:

Values	***Frequency***	***Percent***
1-2	3	10%
3-4	3	10%
5-6	6	20%
7-8	12	40%
9-10	6	20%

Note that we can and should double check our work. Simply add up all the frequencies and make sure the sum is 30. Also add up all the percentages and make sure they add up to 100%.

Here is a **histogram** of the **grouped frequency table** we just generated:



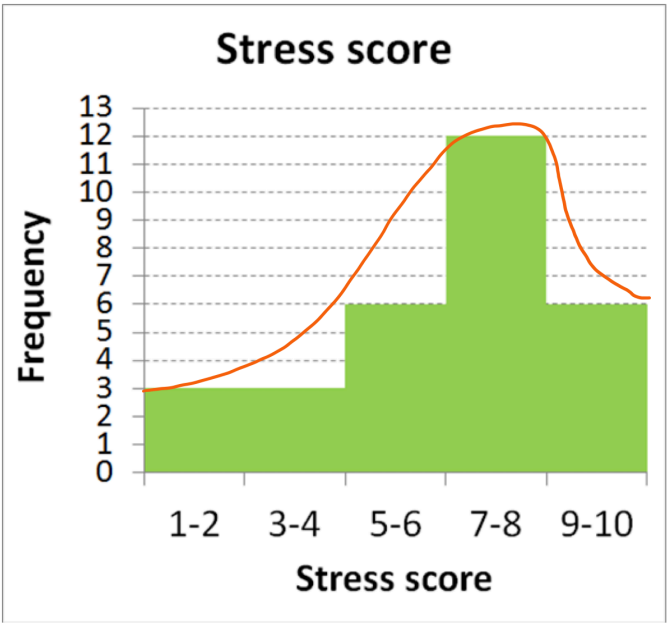
Note that **histograms** are appropriate for plotting **numeric** datasets. The X axis should be labeled with numbers, rather than with categories. That is how a **histogram** differs from your typical bar graph. The other difference is the width of the bar. In **histograms**, because the data are **numeric** or continuous, the bars should appear to touch – with no break in between the bars. This gives a unitary appearance to the shape of the graph. If you were to draw a smooth line over the shape of the distribution, or overall pattern of the data, you would get the impression of a curve.

If you drew a smooth line over the shape of the dataset in a **histogram**, you could describe the shape that is generated with two types of descriptors:

Describing a Distribution

- How many peaks there are
- How symmetrical the shape is

Skewness is the term for describing symmetry: is the distribution of data symmetrical (or very close) – meaning a mirror image from left to right, or is it skewed right/positively, or left/negatively? To determine the direction of skew, you need to check the direction of the "tail". If the tail points right, it is **right skewed**. In this case, the tail points left, so it is **left skewed**.

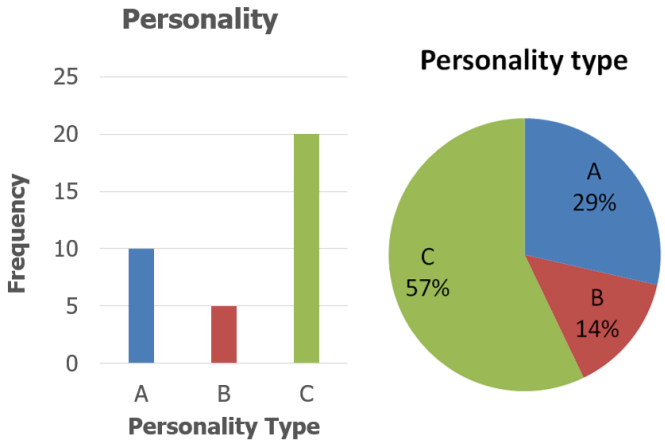


How many peaks the distribution contains is described as **unimodal** or **bimodal**. **Unimodal** distributions show one single collection of scores, whereas **bimodal** distributions look more like a camel's back, with two clear lumps. Do not jump to a **bimodal** description unless

the two peaks are clear and distinct, with some low frequency bins in between. The peaks should also be fairly similar in size to be considered **bimodal**. This **histogram** above clearly displays just one peak, so we would describe it as **unimodal**.

How would the shape of our stress **scores** distribution look if I measured stress **scores** once early in the semester and then again late in the semester? One could speculate that the distribution could become **bimodal**, with the early-in-semester **scores** piling up on the low end of the stress scale, and late-in-semester scores piling up on the high end of the stress scale (as exam and assignment "crunch time" has set in).

Frequency tables and **histograms** are useful for summarizing **numeric** datasets. What about qualitative data, from **nominal variables**? Bar graphs and pie charts are excellent ways to summarize those types of data. Note the gap between bars in a bar graph, as opposed to the touching bars in a **histogram**. This indicates the arbitrariness of the categories. We are still portraying how many of the measured individuals fall into each category, but those categories are not associated with **numeric** values, so no continuity should be implied. Pie charts are excellent for highlighting the relative proportion of **scores** that fall into each category.



A video element has been excluded from this version of the text. You can watch it online here: https://pressbooks.bccampus.ca/statspsych/?p=5

A video element has been excluded from this version of the text. You can watch it online here: https://pressbooks.bccampus.ca/statspsych/?p=5

Chapter Summary

In this chapter, we reviewed why we need statistics. We also introduced some key terms, listed below. We then saw how to summarize data effectively using tables and graphs and describe the patterns the distributions of data make.

Key terms:

descriptive	**nominal**	**independent variable**
inferential	**numeric**	**dependent variable**
variable	**frequency table**	**right skewed**
value	**histogram**	**left skewed**
score	**grouped frequency table**	**unimodal**
		bimodal

Note: concise definitions of all key terms can be found in the Key Terms List at the end of the book.

2. Central Tendency and Variability

2a. Central Tendency

In this chapter we will discuss the three options for measures of central tendency. These measures are all about describing, in one number, an entire dataset.

A measure of central tendency is a statistical measure that defines the centre of a distribution with a single score. The purpose of a central tendency statistic is to find a single number that is most typical of the entire group. It is a number that should represent the entire group as accurately as possible.

Depending on the shape of a distribution, one of these measures may be more accurate than the others. We will see that for symmetrical, unimodal datasets, the **mean** will be the best choice. For asymmetrical (skewed), unimodal datasets, the **median** is likely to be more accurate. For bimodal distributions, the only measure that can capture central tendency accurately is the **mode**.

It is very important to note that two out of three measures of central tendency only apply to numeric data. In order to arrive at a mean or a median, the data need to be measured in number form. It makes sense, for example, to measure the average student height in a class. It does not make sense to determine the average major from a class of students.

Before we can learn to calculate a mean, we need to familiarize ourselves with some statistical notation. In statistics, when we want to denote "taking the sum" of a series of numbers, we use the term **Σ**. This is the Greek capital letter S, known as "sigma".

The tricky thing about **Σ** is learning how to use it within mathematical order of operations. You may remember the mnenonic BEDMAS from school. This indicates that you should first do any operations that are set off in brackets or parentheses. Next you should do exponents, then division or multiplication, and finally addition/

subtraction. But summation fits in just after division/multiplication. So BEDMAS becomes BEDMSAS.

Order of Operations

1. Brackets
2. Exponents
3. Division/Multiplication
4. **Summation**
5. Addition/Subtraction

BEDM**S**AS

Let us try some examples. We will assume my variable X represents a set of scores in a dataset:

X
5
8
9
6
7

If you see the formula:

$$(\sum X)^2$$

What is that telling you to do? Look at order of operations. First do anything in brackets. So first we have to do the **Σ**X part. This tells us to add up all the scores: 5+8+9+6+7 = 35. Next we need to take that result and square it (exponents): 35^2 = 1225.

Let us try:

$$\sum X^2$$

Now there are no brackets, so exponents come first. This formula says to square each score in the dataset, then add together all the results: $5^2+8^2+9^2+6^2+7^2$ = 25+64+81+36+49 = 255. So what seems like a minor difference in the formula really changes the result!

Let us try:

$$\sum (X-1)^2$$

First brackets, so subtract off 1 from each score. Next exponents, so square each result. And finally summation, so add that all together. 42+72+82+52+62 = 16+49+64+25+36 = 190.

And finally, we will try:

$$\sum X - 1^2$$

Now exponents is first, so square the number 1. Next is summation, so add all scores together. Finally, complete the subtraction. 1^2 = 1. 5+8+9+6+7 = 35. 35 – 1 = 34.

So, now you see that order of operations are vital to decode summation notation. Each of these variations have very different solutions. Whenever you see a new formula, try translating it into words after reviewing order of operations.

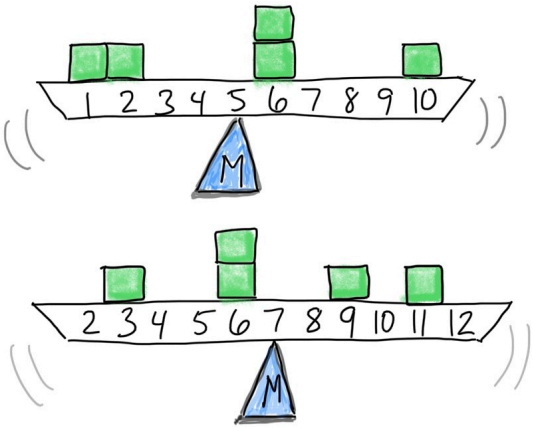
Now we are ready to take a look at calculating a **mean**. The **mean** is the most common measure of central tendency, because it has some powerful applications in statistics. The **mean** is the same thing as an average, something you are very familiar with. You also probably know that to find an average, you add up all the numbers, then divide by how many numbers there were.

$$M = \frac{\sum X}{N}$$

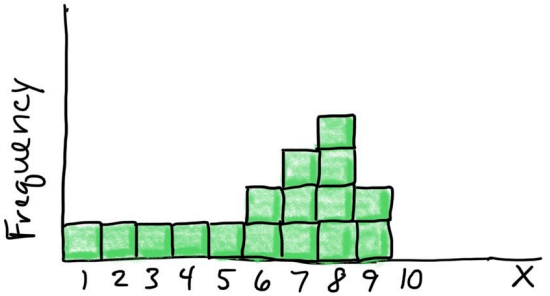
In statistical notation, the formula for the mean is shown above. A **mean** is symbolized as **M**. The number of scores in a dataset is symbolized as **N**.

Conceptually we can think of the **mean** as the balancing point for the

distribution. In a **histogram**, if we were to mentally convert the X-axis into a scale, the **mean** would be the fulcrum, or the point of the scale at which the two sides of the scale balance each other out. Each score is like a weight. Their position along the scale determines where the mean will be. Here are some examples.



In the top **histogram**, the **mean** is (1+2+6+6+10)/5 = 25/5 = 5. In the bottom **histogram**, the **mean** is (3+6+6+9+11)/5 = 35/5 = 7.In the previous examples, the **mean** was pretty close to the middle of the scale – not that surprising, because the data were spread out fairly evenly. However, things can change if the data pile up toward one end of the distribution (i.e. skewed), or if any of the data points are quite extreme (outliers). In the example here, think about where the balance point would be.



If we put it close to the middle of the scale, for example at 5, the scale would tip to the right. So we have to move the balance point rightward. Let's see if that intuition is correct: **M** = (1+2+3+4+5+6+6+7+7+7+8+8+8+8+9+9)/16 = 6.1

The **median**, unlike the **mean**, is a counting-based measure. The values of the the **scores** are not important, just how many of them there are. To get the **median**, you find the midpoint of the scores after placing them in order. The **median** is the point at which half of the scores fall above and half of the scores fall below. Finding the **median** for an odd number of scores is easy. Just find the single middle score, and that is the **median**.

Odd number of scores

1 2 3 4 5

↑

Median = 3

1 2 3 4

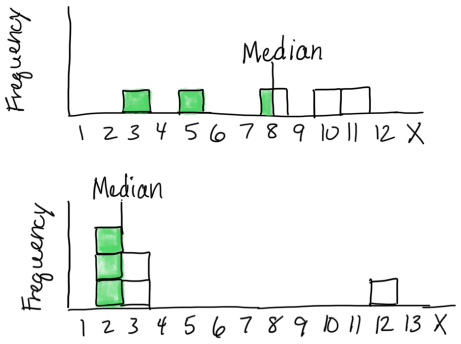
↑

Median = 2.5

Even number of scores

For an even number of scores, it is a little more complicated. You have to find the middle two scores and average those together.

Here are some graphical examples of **medians**.



In the top histogram, 2 scores are below the **median** of 8, and two scores are above. Notice the **median** is not a balancing point (which would be a little to the left to account for the spread of the lower scores). In the bottom histogram, there are 2 scores below and two scores above the middle set of scores (2 and 3, averaged to 2.5). The **median** is not a balancing point (which would need to be further to the right to account for the score weighting down the right end of the scale).

The **mode** is our final option for statistical measures of central tendency. The **mode** is simply the **score** that occurs most often in the dataset, the one with the highest frequency. This is the measure that can be used with **nominal** data too. It is possible to have more than one **mode**, if the dataset is **bimodal**, for example. In fact the term **unimodal** means "one **mode**" and **bimodal** means "two **modes**". Note that the **mode** must correspond to an actual **score** in the dataset, so a **grouped frequency table** or **histogram** will not help you identify it.

Another thing to note is that if your dataset happens to have two **modes**, that does not necessarily mean it is appropriate to describe the distribution as **bimodal**. Remember, true **bimodal** distributions are ones that show two distinct peaks in the smoothed line with some space between, indicating there are two collections of scores that are clustered together. Basically, if the distribution does not look like a camel's back, then it is not truly **bimodal**.

Camel Farm in Mongolia 02 by Alexandr frolov is licensed under CC BY-SA 4.0

Now we should take a look at the difference between a **mean** and a **median**.

On the **histogram** above, you can see that there is a cluster of **scores** at the low end of the scale. However, there is one outlier with an extreme **score** at the upper end. The **median** is concerned with how many scores fall above or below, rather than their values, so it is not affected much by the 12 way out there. It still reflects the center of the distribution accurately. However, the **mean** takes into account the value of each score. Without the outlier, the **mean** would have been (2+2+2+3+3)/5 = 2.4. With the outlier, the **mean** is (2+2+2+3+3+12)/6 = 4.

With an outlier, or a heavily-skewed distribution, the **mean** can be pulled in the direction of the outlier or skew, and is thus not the most accurate measure of central tendency. Under these circumstances, the **median** will better describe the dataset.

2b. Variability

Our objectives in this part of the chapter will be that to explain the concept of variability and why it is important, and to calculate the **descriptive** statistics **variance** and **standard deviation**.

Soon we will be progressing to **inferential** statistics, in which we will often wish to figure out if the central tendency of one group of scores is different from another group of scores. If we want to be able to assert differences in what is typical between one group and another, we need enough uniformity within each of those groups to discern those differences. If there is too much random variability, we will be unable to say much about the data or use them for decision making. There will be too little order in the chaos. For that reason, we need to learn how to measure variability in a data set and take that into account in the process of making inferences about the data.

There are many ways to measure variability. However, we will focus on the two main measures of variability that are commonly use in both descriptive and inferential statistics of the sort we will cover in this course: **variance** and **standard deviation**. It is worth noting that **variance** and **standard deviation** are directly related, but **standard deviation** is easier to interpret and is thus more often reported as a **descriptive** statistic.

In general, measures of variability describe the degree to which **scores** in a data set are spread out or clustered together. They also give us a sense of the width of a distribution. Finally, they help us understand how well any individual statistic (for example the **mean**) can possibly represent the distribution as a whole.

When it comes to **inferential** statistics, smaller variability is better. When comparing two distributions, as we will be doing in inferential statistics, there are two ways to be confident that there is a difference. One is to have dramatically different central tendencies (such as the **means**). The other way is to have small variability, such that an individual statistic represents that distribution well.

The first measure of variability we will learn to calculate is **variance**. **Variance** summarizes the extent to which scores are spread out from the mean. To do that, we calculate the deviation of each score from the mean. Here is a graphical illustration of deviations. As the first step in calculating **variance**, for each score we find its distance from the mean. So here if the distributions mean is 6, a score of 1 is a deviation of 5 away from the mean, and a score of 7 is a deviation of 1 away from the mean.

Here is the formula to calculate variance.

$$SD^2 = \frac{\sum(X - M)^2}{N}$$

I have translated the steps of the formula into words for you here. Keep in mind that we have to do each part if the formula in the order of operations. First brackets, then exponents, then summation, and finally division.

Steps to Calculate Variance

1. Take the distance, or "deviation", of each score from the mean
2. Square each distance to get rid of the sign
3. Add up all the squared deviations
4. Divide by the number of scores

The first part of the formula we need to calculate, then, is to take the distance, or "deviation", of each score from the mean. This is written as X-M.

Next, we square each distance to get rid of the sign (because some deviations will be negative numbers, which we do not want. This is the exponent outside the brackets.

$$SD^2 = \frac{\sum(X - M)^2}{N}$$

Next, we do summation, so we add up all the squared deviations. In fact, the result of this step has its own name: **Sum of Squares,** which we will sometimes abbreviate as SS.

And finally, we divide by the number of scores, to make sure this is an average measure of distances in the dataset. Now, we're doing the division last, because of the notation, because there is a top and a bottom. So that makes it clear that the division is the last thing that we do in the order of operations.

One thing to note is that the purpose of squaring each deviation before taking the sum is the following. All the deviations for scores smaller than the mean will come out negative. All the deviations for scores larger than the mean will come out positive. So if we added up the deviations without squaring them, the negatives would cancel out the positives and the variance would always end up zero. Squaring each deviation is mainly a way to convert them all to positive numbers.

Standard deviation is the other measure of variability we will use in this course. It expresses the variability in terms of a typical deviation in the data set. This will be a single number that gives us the distance of typical scores in the dataset from the mean.

Variance is essentially the average squared deviation; now we want to find the average deviation to get it back into the original units of the data. To find the **standard deviation**, we just need to take the square root of the **variance**.

$$SD = \sqrt{SD^2}$$

Most of the **inferential** statistics we will use in this course will be based on our calculations of the **mean** and either the **variance** or **standard deviation**.

Chapter Summary

In this chapter we examined the purpose and common methods for determining statistics representing the central tendency and the

variability of a dataset. We saw the particular characteristics of each statistic that makes it most appropriate or useful for specific situations. The combination of central tendency and the variability statistics not only provides a very succinct summary of the dataset, but it also will become the basis for making inferences from data.

Key Terms:

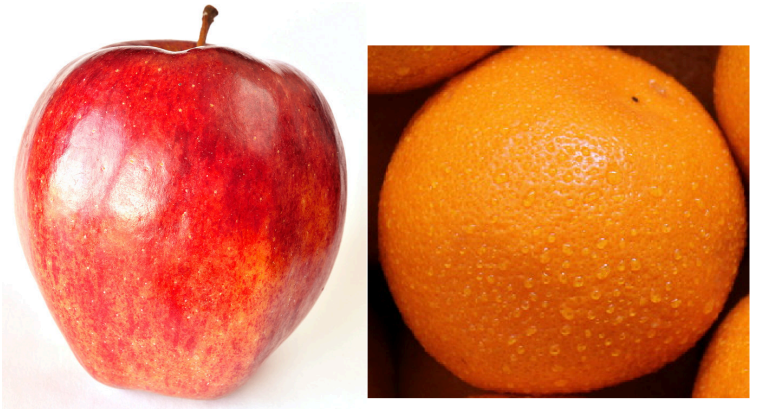
mean	**Σ**	**variance**
median	**M**	**standard deviation**
mode	**N**	**Sum of Squares**

3. Z-scores and the Normal Curve

3a. Z-scores

In this chapter, we will address the topic of **Z-scores**, one type of what are commonly called standard scores.

Before we begin, we will examine a real world example of why standardizing scores is useful and important. Have a look at these images:



"Apple" by Open Grid Scheduler / Grid Engine is marked with CC0 1.0
"Oranges" by Dious is marked with CC PDM 1.0

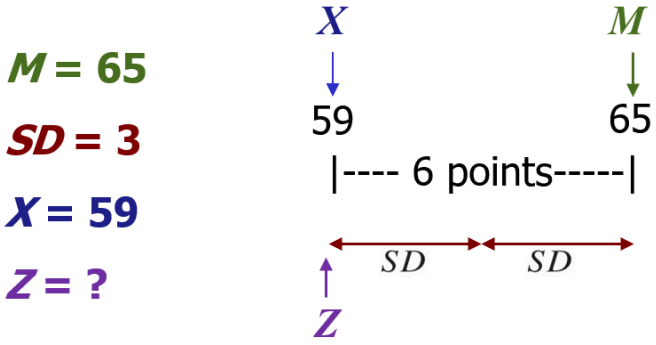
How sweet are these fruits? Is an apple sweeter? Or is an orange sweeter? What do you think? Is it difficult to say? Usually when I survey people with this question, there is a pretty even split, with half saying an apple is sweeter, and the other half saying the orange is sweeter. Why? When we think about it, it is tough to directly compare them in terms of the property of sweetness. One is more tart, the other quite mild in flavour, so it is difficult to compare oranges and apples. In fact, if you are a native speaker of English, you might have heard before "that's

like comparing oranges to apples," meaning it's impossible to compare. The same is true of trying to compare numbers from different datasets.

In this chapter, we will learn how to use the statistics of the mean and standard deviation to generate standard scores, or **Z-scores**. This will allow us to transform scores in any numeric dataset, using any scale, into a standard metric. This allows for precise comparison of a particular score to the rest of the scores in a dataset, and even across different datasets.

A **Z-score** is just a raw score expressed in terms of its position relative to the mean and in terms of standard deviations. A negative **Z-score** means that raw score is below the mean. A positive **Z-score** means that raw score is above the mean. In addition to its position above or below the mean, a **Z-score** also communicates that score's distance from the mean in terms of how many standard deviations away it is.

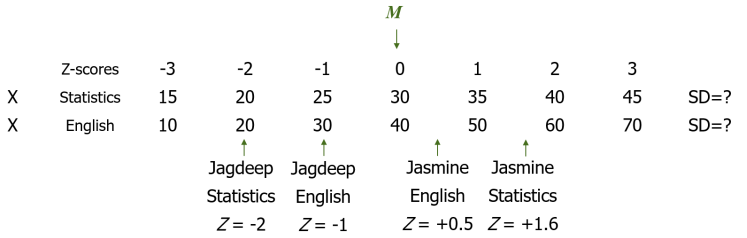
For example, with a mean of 65 and standard deviation of 3, the raw score 59 can be converted into a **Z-score**.



The **Z-score** tells you how many standard deviations away from the mean the raw score is. How many? ... 2. It also tells you if it is above (greater than) or below (less than) the mean. Is it below or above the mean? ... below. Thus the **Z-score** is -2, because it is 2 standard deviations below the mean.

Z-scores can also be useful for comparing scores across two different variables. For example, let's say two students, Jagdeep and Jasmine, are in both a Statistics class and an English class.

Raw scores	Jagdeep	Jasmine
Statistics	20	38
English	30	45



Notice that in statistics, the mean grade is a raw score of 30. From the distance between **Z-scores**, we can surmise that the standard deviation in statistics is 5. In English, however, the mean grade is a raw score of 40, and the standard deviation is 10. It would not make sense to directly compare Jagdeep's test score of 20 in statistics with Jasmine's test score of 45 in English. These were entirely different tests, so some translation to standard scores is needed. With **Z-scores**, we can take their grades from those two different classes and figure out their performance relative to the mean in each class. Then we can compare Jagdeep's performance in statistics to their performance in English, and we can also compare their performance to Jasmine's performance.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=173#h5p-7

So how do we actually calculate a **Z-score**? It's a simple formula:

$$Z = \frac{X - M}{SD}$$

For any given raw score (X) that we want to translate, we subtract off the mean. Then we take that result and divide by the standard deviation. Note that if the result is a negative number, the score is below the mean. If the result is positive, the score is above the mean.

We can also rearrange the **Z-score** formula to go backwards from a **Z-score** to a raw score.

$$X = (Z)(SD) + M$$

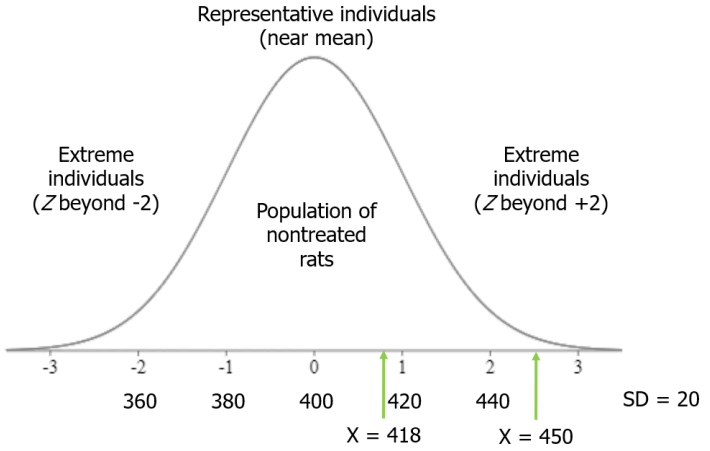
You might wish to do this if you want to figure out a cutoff score, for example in the context of grading. If you are grading a class on a curve, you may decide that anyone who falls 2 or more standard deviations below the mean will be given a failing grade. So you could figure out what the test score is, below which students will be assigned an F. To do this, you take the **Z-score**, in this case -2, and multiply it by the standard deviation. Then add the mean. This will get you to the test score that would serve as the cutoff, below which students would be assigned a grade of F. Note that if the **Z-score** is negative, essentially you are subtracting some amount from the mean.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=173#h5p-8

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=173#h5p-5

Z-scores can be very helpful for figuring out whether a score is extreme relative to a comparison distribution. We'll see that this is what we typically wish to do in inferential statistics: if I give a new medicine to a sick patient, does that patient do better than most of the other patients treated with a standard remedy?

In this example, we will test the hypothesis that rats fed a grain-heavy diet become unusually overweight. We have a group of rats, and their average weight is 400 grams, with a standard deviation of 20 grams. We start feeding the rats a grain-heavy diet, and measure each one after a few weeks on that new diet. The question is, how heavy should they be for us to conclude that they are unusually overweight?



A good rule of thumb is that more than 2 standard deviations from the mean (in either direction) is considered fairly extreme. Beyond 3 standard deviations is very extreme.

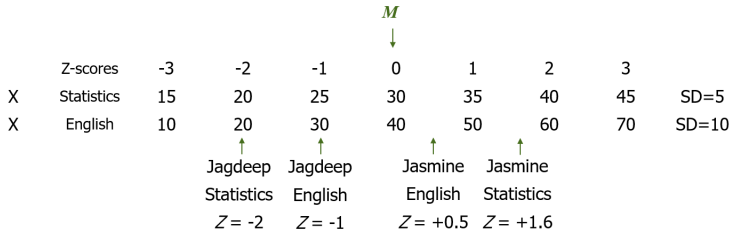
Suppose we measure the first rat after the grain-heavy diet, and it weighs 418 grams. This is less than 1 standard deviation from the mean: Z = (418-400)/20 = 0.9. Therefore, we cannot be sure whether the diet made rats unusually heavy.

Next, though, we measure rat number 2. It weighs 450 grams. That is 2.5 standard deviations from the mean, so more than 2: Z = (450-400)/20 = 2.5. Thus, rat number 2 qualifies as extreme on my distribution.

As we measure rats, if more of them end up in the extreme region on the distribution, then we will have a strong basis for concluding that the grain-heavy diet made for unusually heavy rats.

Now we can take a closer look at our example with Jagdeep and Jasmine and their performance in two different classes. By converting their raw scores, shown here, into **Z-scores**, we can directly compare them to their classmates, each other, and between classes.

Raw scores	Jagdeep	Jasmine
Statistics	20	38
English	30	45



Note that the statistics class has a different mean and standard deviation than does the English class. For statistics, the mean score is 30 and the standard deviation is 5. Given this, with a grade of 20, Jagdeep's **Z-score** in statistics is Z = (20-30)/5 = -2. They are two standard deviations below the mean. In English, the mean score is 40 with a standard deviation of 10. So Jagdeep, with a score of 30, has a **Z-score** of Z = (30-40)/10 = -1, because they are one standard deviation below the mean.

Jasmine, on the other hand, is above the mean in both classes. They are .5 standard deviations above the mean in English: Z = (45-40)/10 = .5. Their **Z-score** in statistics is Z = (38-30)/5 = 1.6.

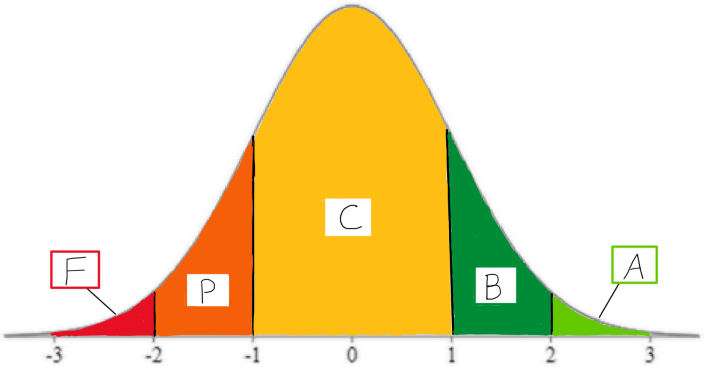
Note that we can compare all these scores against each other only once converted to **Z-scores**. In statistics, if a student had a score of 40, that would be a very good grade! But in English it would be just average. Once converted in to **Z-scores** we can see that Z = +2 in statistics is clearly a higher grade than Z = 0 in English.

3b. Normal Curve

In this part of the chapter, we will be discussing the **normal curve**, as a model distribution, and taking a look at how **Z-scores** relate to the **normal curve**.

First, consider a real-life situation that you can likely relate to. What if I told you this class would be graded on a bell curve? What would your reaction be?

Usually, when I ask that question of students, they have a negative reaction. But why? In highly traditional educational systems, there is a concept of gatekeeping that suggests only a particular proportion of students should receive a particular grade. Typically, the bell curve is used in situations where a program wants to make progression through the program competitive, to "weed out" students who perform less well than their peers on exams. Using the normal curve, or bell curve, as a model upon which to map student scores, the instructor can determine precisely how many students will receive A's, B's, C's, and so on.



If, for example, only about 2% of students should receive the highest grade, then, the instructor will set the minimum standard for that grade as a **Z-score** of +2. Only students who receive a calculated course score that is at least 2 standard deviations above the mean will get the

coveted A grade. Using areas under the normal curve, the instructor can relate that to the percentage of students who will receive a particular grade, and can map that onto which **Z-score** forms the criterion for that grade.

So why is grading on a bell curve unpopular among students? It assumes that, regardless of who is in the class, only the top few students should be awarded a high grade. The bell curve model assumes that most students show a mediocre level of mastery. In effect, this assumption means that students compete directly against one another for their grades. By helping another student in the class, you could be risking losing your place in the rankings to them. This can create a high-stress environment, and it can distort the reality of how many students in the class do actually understand the material quite well.

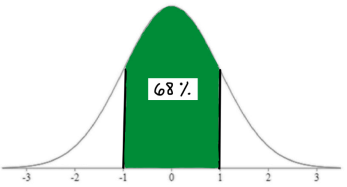
Now that you have a sense of what the normal curve looks like and one of its common uses, let us take a broader view of the basic concepts involved. A **normal curve** is a theoretical distribution that is sometimes called a Z distribution. The normal curve has very distinct set of properties that make it a useful model for data analysis. In real life, few distributions actually match this theoretical model, so when you are describing the shape of a distribution, even if it looks pretty nice and symmetrical like this, you should refrain from describing it as a normal curve.

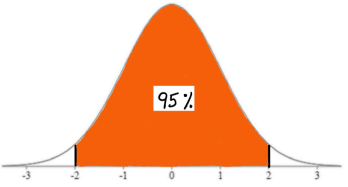
An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=173#h5p-2

Another important concept we will address is that of a **percentile**. A **percentile** is a score on a distribution that corresponds to a certain area under the curve. It's often used as a means to rank scores – for example, if you take a standardized test like the GRE, they will report your **percentile** ranking as well as your actual score.

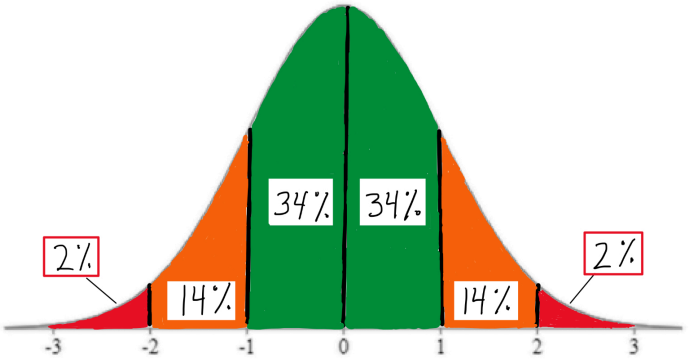
The standard normal distribution is a theoretical distribution that should usually be approximated if a dataset is large enough. For

example, if I were to measure the IQ score of millions of people, or measure the heights of all the people currently alive in the world, or make a histogram of all the scores of all students who could ever take a test, these distributions would likely look pretty close in shape to this theoretical distribution.





For this reason, and because it has very predictable set of properties, it is commonly used as a model for data analysis. As you can see here, using this model you can easily identify what percentage of scores in a standard normal distribution fall between Z score landmarks.

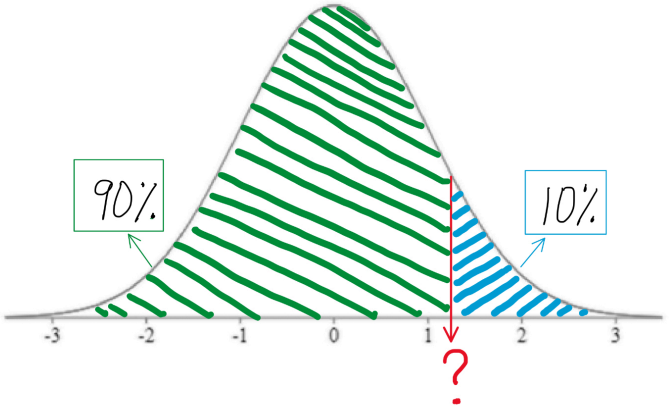


In fact, you should memorize these handy area-under-the-curve landmarks: the 2-14-34 rule. These come in very helpful for quick visual estimation, when we are working with the **normal curve** model.

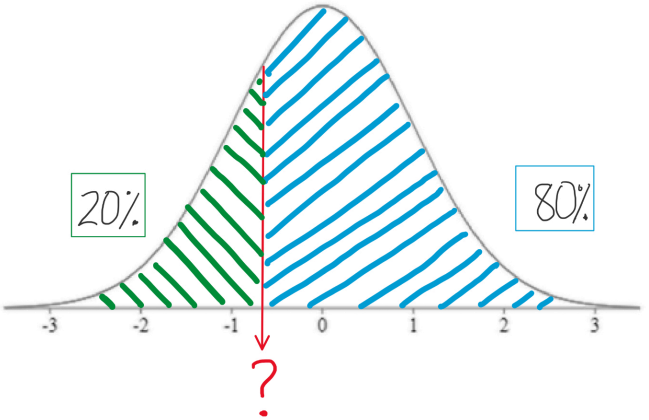
An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=173#h5p-3

In addition to the handy 2-14-34 percentage estimates, the standard **normal curve** also can be used to determine a score's exact percentage probability of occurring within this theoretical distribution. In our interactive exercises, we will see how to work with tables to look up the exact area under the curve associated with any **Z-score**, not just the landmarks shown on these illustrations. We will also be able to work backwards from **percentiles**, which are often used as a measure of relative standing in a distribution.

Percentiles can be a tricky concept at first. Let's take a closer look. A **percentile** is the score at which a given percentage of scores in the distribution fall beneath. To visualize the percentage, we always sketch the **normal curve**, and shade the area under the curve starting from the left end up to the relevant proportion. The 90th **percentile** is the score at which 90% of individuals scored lower than that. So you start from the left end of the curve and shade up until you have shaded in 90% of the area under the curve. Clearly that would be the vast majority of the area.



The 20th **percentile** is the score at which only 20% of individuals scored lower. So you are only shading a pretty small area under the curve, starting from the left end.



Note that the 50th **percentile** would be shading the lower half of the curve, so the 50th **percentile** is always right in the middle of the curve,

and thus it always a **Z-score** of 0, otherwise known as the mean of the distribution.

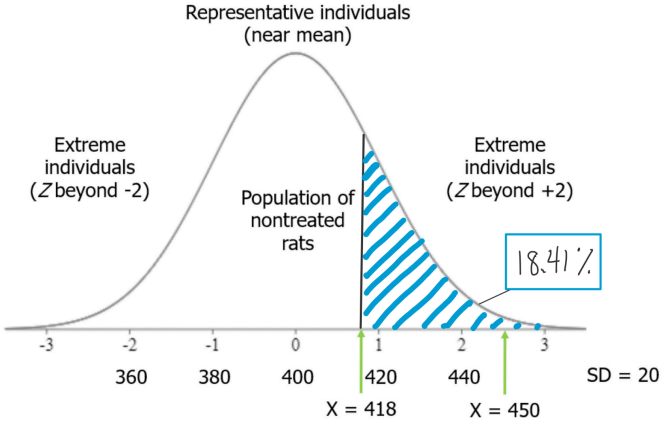
If you ever find yourself getting confused about the concept of the **percentile** and how it works, think about this. New parents frequently report on their baby's achievements in these terms: little Doneil is in the 80th **percentile** for weight, and the 99th **percentile** for height! They are mentioning these things because they are above average. Baby Doneil is heavier and longer than most babies that age! You do not hear people bragging too often about their babies being in the 30th **percentile** for anything. Why? Because that would mean their baby is below average in that characteristic. So with **percentile** rankings, you are always starting from the bottom of the distribution and seeing what percentage of the distribution of scores you are outranking with the score in question.

Recall our experiment with the rats being fed a grain-heavy diet. We saw that the rat who weighed 418 grams after eating the diet was a bit heavier than the average of non-treated rats, but that was not even a full standard deviation higher than the mean. The **Z-score** was 0.9, when we calculated it. If we were to look up in a table the area under the curve to the right of that **Z-score**, we would see that the area in the tail of the distribution associated with that **Z-score** is 18.41%.

Z	% Mean to *Z*	% in Tail
.90	31.59	18.41

So when we ask ourselves, "how unusual is that score?" or "what percentage of scores are more extreme?" we have a precise answer, if we use the **normal curve** as our model. We can say that there is about an 18% probability that any rat drawn at random from the normal rat

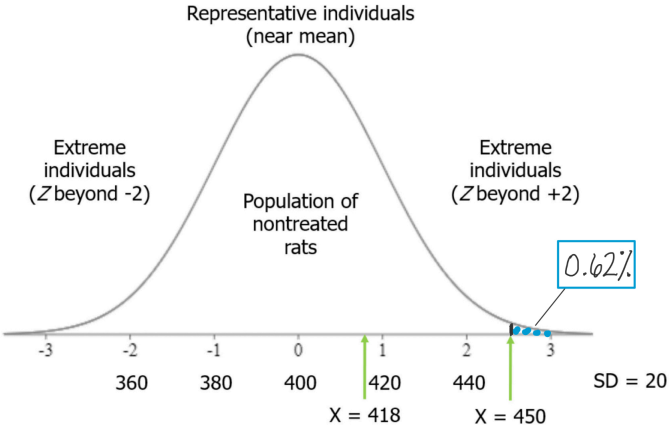
weight distribution will be at least that heavy. That does not seem too unusual.



On the other hand, if you recall, the second rat in our experiment, who ate the grain-heavy diet, weighed quite a bit more. In terms of **Z-scores**, his weight was 2.5, or 2-and-a-half standard deviations above the mean. Such a score is far less probable under our **normal curve** model. If we look up the area under the curve in a table, we will see that the area in the tail of the distribution associated with that **Z-score** is 0.62%.

Z	% Mean to Z	% in Tail
2.50	49.38	.62

There is less than a 1% chance that any rat drawn at random from the normal rat weight distribution will be at least that heavy. That is far more unusual as a rat weight.



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=173#h5p-9

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=173#h5p-10

These are lots of puzzle pieces I have tossed together, just to give you a preview of how they all fit together, and how we might be able to use them for inferential data analysis. A raw score can be translated to a **Z-score**, which can be mapped onto the normal curve. If it's a **normal curve,** then a known proportion of the distribution is associated with that **Z-score**, and that allows us to connect scores to probabilities. That is where we are headed with the next chapters, and soon we will be getting into full-on inferential statistics techniques.

Chapter Summary

In this chapter, we introduced the concepts and utility of standardized scores like **Z-scores** and the **normal curve**. We saw that combining those tools offers the opportunity to rank scores using **percentiles** and to estimate the probability of a particular score occurring within a distribution, from which we can make inferences about how unusual a score is.

Key terms:

Z-score **normal curve** **percentile**

4. Probability, Inferential Statistics, and Hypothesis Testing

4a. Probability and Inferential Statistics

In this chapter, we will focus on connecting concepts of **probability** with the logic of inferential statistics.

These notable quotes represent why **probability** is critical for a basic understanding of scientific reasoning.

"The whole problem with the world is that fools and fanatics are always so certain of themselves, and wiser people so full of doubts."
— Bertrand Russel (1872-1970)

"Medicine is a science of uncertainty and an art of probability."
— William Osler (1849–1919)

In many ways, the process of postsecondary education is all about instilling a sense of doubt and wonder, and the ability to estimate **probabilities**. As a matter of fact, that essentially sums up the entire reason why you are in this course. So let us tackle **probability**.

We will be keeping our coverage of **probability** to a very simple level, because the introductory statistics we will cover rely on only simple **probability**. That said, I encourage you to read further on compound and conditional **probabilities**, because they will certainly make you smarter at real-life decision making. We will briefly touch on examples of how bad people can be at using **probability** in real life, and we will then address what probability has to do with inferential statistics. Finally, I will introduce you to the **central limit theorem**. This is probably

one of the heftiest math concepts in the course, but worry not. Its implications are easy to learn, and the concepts behind it can be demonstrated empirically in the interactive exercises.

First, we need to define **probability**. In a situation where several different outcomes are possible, the probability of any specific outcome is a fraction or proportion of all possible outcomes. Another way of saying that is this. If you wish to answer the question, "What are the chances that outcome would have happened?", you can calculate the **probability** as the ratio of possible successful outcomes to all possible outcomes.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-17

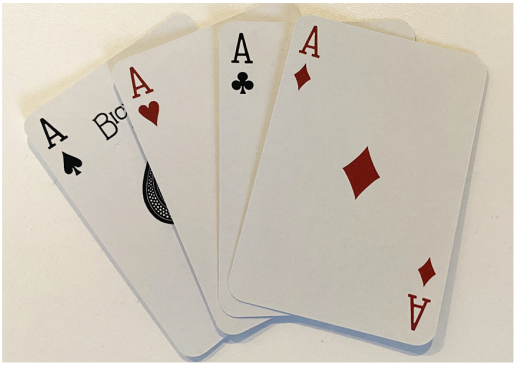
People often use the rolling of dice as examples of simple **probability** problems.

"Dice" by matsuyuki is licensed under CC BY-SA 2.0

If you were to roll one typical die, which has a number on each side from 1 to 6, then the simple probability of rolling a 1 would be 1/6. There are six possible outcomes, but only 1 of them is the successful outcome, that of rolling a 1.

> *An interactive H5P element has been excluded from this version of the text. You can view it online here:*
> *https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-13*

Another common example used to introduce simple **probability** is cards. In a standard deck of casino cards, there are 52 cards. There are 4 aces in such a deck of cards (Aces are the "1" card, and there is 1 in each suit – hearts, spades, diamonds and clubs.)



If you were to ask the question "what is the **probability** that a card drawn at random from a deck of cards will be an ace?", and you know all outcomes are equally likely, the **probability** would be the ratio of the number of times one could draw and ace divided by the number of all possible outcomes. In this example, then, the **probability** would be 4/52. This ratio can be converted into a decimal: 4 divided by 52 is 0.077, or 7.7%. (Remember, to turn a decimal to a percent, you need to move the decimal place twice to the right.)

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-16

Probability seems pretty straightforward, right? But people often misunderstand **probability** in real life. Take the idea of the lucky streak, for example. Let's say someone is rolling dice and they get 4 6's in a row. Lots of people might say that's a lucky streak and they might go as far as to say they should continue, because their luck is so good at

the moment! According to the rules of **probability**, though, the next die roll has a 1/6 chance of being a 6, just like all the others. True, the **probability** of a 4-in-a-row streak occurring is fairly slim: 1/6 x 1/6 x 1/6 x 1/6. But the fact is that this rare event does not predict future events (unless it is an unfair die!). Each time you roll a die, the **probability** of that event remains the same. That is what the human brain seems to have a really hard time accepting.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-15

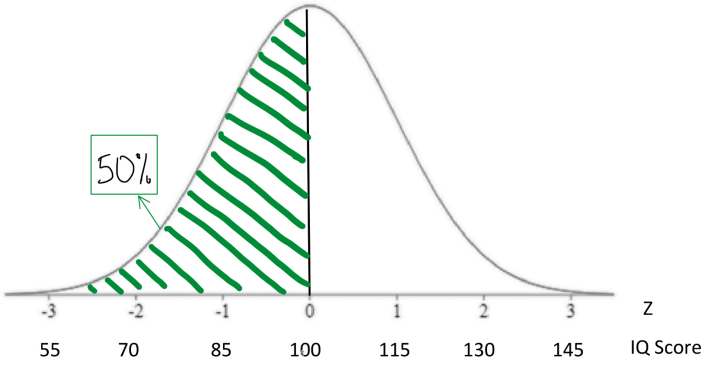
When someone makes a prediction attached to a certain probability (e.g. there is only a 1% chance of an earthquake in the next week), and then that event occurs in spite of that low **probability** estimate (e.g. there is actually an earthquake the day after the prediction was made)... was that person wrong? No, not really, because they allowed for the possibility. Had they said there was a 0% chance, they would have been wrong.

Probabilities are often used to express likelihood of outcomes under conditions of uncertainty. Like Bertrand Russell said, wise people rarely speak in terms of certainties. Because people so often misunderstand **probability**, or find irrational actions so hard to resist despite some understanding of **probability**, decision making in the realm of sciences needs to be designed to combat our natural human tendencies. What we are discussing now in terms of how to think about and calculate **probabilities** will form a core component of our decision-making framework as we move forward in the course.

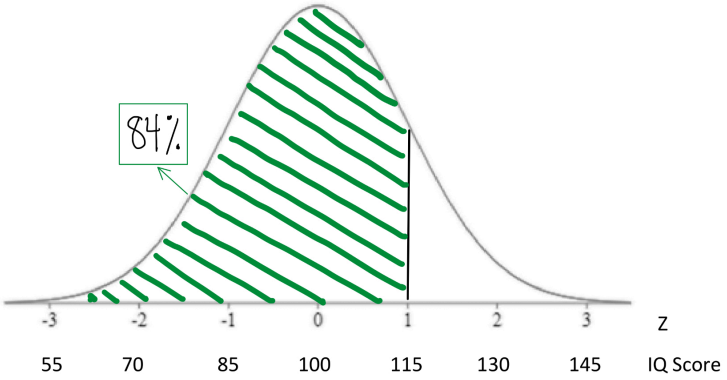
Now, let's take a look at how **probability** is used in statistics.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-18

We saw that percentiles are expressions of area under a normal curve. Areas under the curve can be expressed as **probability**, too. For example, if we say the 50th percentile for IQ is 100, that can be expressed as: If I chose a person at random, there is a 50% chance that they will have an IQ score below 100."



If we find the 84th percentile for IQ is 115 there is another way to say that "If I chose a person at random, there is an 84% chance that they will have an IQ score below 115."



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-19

Any time you are dealing with area under the normal curve, I encourage you to express that percentage in terms of **probabilities**. That will help you think clearly about what that area under the curve means once we get into the exercise of making decisions based on that information.

An interactive H5P element has been excluded from this version of the text. You can view it online here:

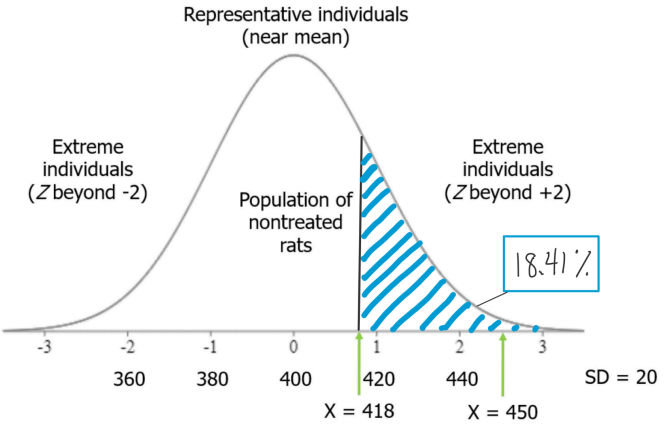
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-21

Probabilities, of course, range from 0 to 1 as proportions or fractions, and from 0% to 100% when expressed in percentage terms. In inferential statistics, we often express in terms of **probability** the likelihood that we would observe a particular score under a given normal curve model.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-20

Although I encourage you to think of **probabilities** as percentages, the convention in statistics is to report to the probability of a score as a proportion, or decimal. The symbol used for "probability of score" is *p*. In statistics, the interpretation of "*p*" is a delicate subject. Generations of researchers have been lazy in our understanding of what "*p*": tells us, and we have tended to over-interpret this statistic. As we begin to work with "*p*", I will ask you to memorize a mantra that will help you report its meaning accurately. For now, just keep in mind that most psychologists and psychology students *still* make mistakes in how they express and understand the meaning of "*p*" values. This will take time and effort to fix, but I am confident that your generation will learn to do better at a precise and careful understanding of what statistics like "*p*" tell us... and what they do not.

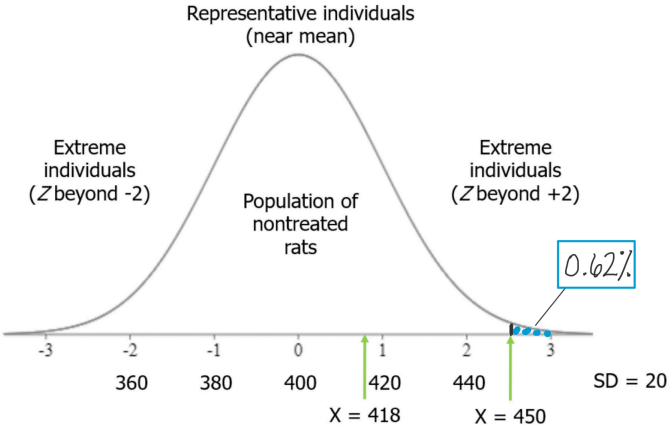
To give you a sense of what a statement of $p < .05$ might mean, let us think back to our rat weights example.



If I were to take a rat from our high-grain food group and place it on the distribution of untreated rat weights, and if it placed at $Z = .9$, we could look at the area under the curve from that point and above. That would tell us how likely it would be to observe such a heavy rat in the general population of nontreated rats — those that eat a normal diet.

Think of it this way. When we select a rat from our treatment group (those that ate the grain-heavy diet), and it is heavier than the average for a nontreated rat, there are two possible explanations for that observation. One is that the diet made him that way. As a scientist whose hypothesis is that a grain-heavy diet will make the rats weigh more, I'm actually motivated to interpret the observation that way. I want to believe this event is meaningful, because it is consistent with my hypothesis! But the other possibility is that, by random chance, we picked a rat that was heavy to begin with. There are plenty of rats in the distribution of nontreated rats that were at least that heavy. So there is always some **probability** that we just randomly selected a heavier rat. In this case, if my treated rat's weight was less than one standard deviation above the mean, we saw in the chapter on normal curves that the **probability** of observing a rat weight that high or higher in the nontreated population was about 18%. That is not so unusual. It would not be terribly surprising if that outcome were simply the result of random chance rather than a result of the diet the rat had been eating.

If, on the other hand, the rat we measured was 2.5 standard deviations above the mean, the tail **probability** beyond that **Z-score** would be vanishingly small.



The **probability** of observing such a rat weight in the nontreated population is very low, so it is far less likely that observation can be accounted for just by random chance alone. As we accumulate more evidence, the **probability** they could have come at random from the nontreated **population** will weigh into our decision making about whether the grain-heavy diet indeed causes rats to become heavier. This is the way **probabilities** are used in the process of **hypothesis testing**, the logic of inferential statistics that we will look at soon.

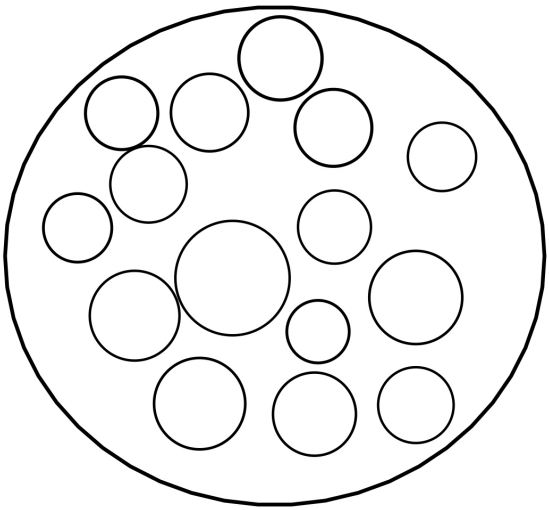
An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-11

Now that you have seen the relevance of **probability** to the decision making process that comprises inferential statistics, we have one more major learning objective: to become familiar with the **central limit theorem**.

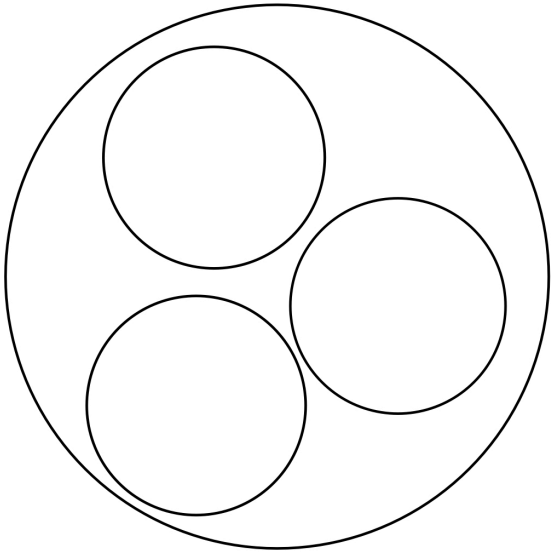
However, before we get to the **central limit theorem**, we need to be clear on the distinction between two concepts: **sample** and **population**. In the world of statistics, the **population** is defined as all possible individuals or scores about which we would ideally draw conclusions. When we refer to the characteristics, or parameters, that describe a **population**, we will use Greek letters. A **sample** is defined as the individuals or scores about which we are actually drawing conclusions. When we refer to the characteristics, or statistics, that describe a **sample**, we will use English letters.

It is important to understand the difference between a **population** and a **sample**, and how they relate to one another, in order to comprehend the **central limit theorem** and its usefulness for statistics. From a **population** we can draw multiple **samples**. The larger **sample**, the more closely our **sample** will represent the **population**.

Think of a Venn diagram. There is a circle that is a **population**. Inside that large circle, you can draw an infinite number of smaller circles, each of which represents a **sample**.



The larger that inner circle, the more of the **population** it contains, and thus the more representative it is.

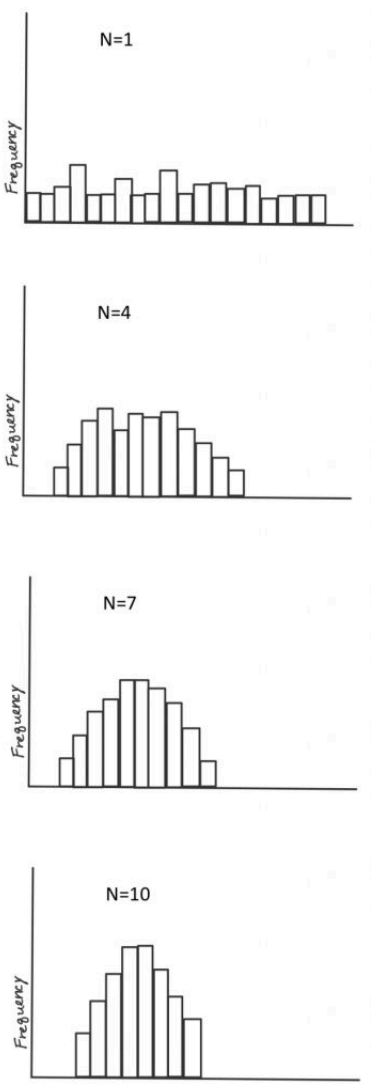


Let us take a concrete example. A **population** might be the depression screening scores for all current postsecondary students in Canada. A sample from that **population** might be depression screening scores for 500 randomly selected postsecondary students from several institutions across Canada. That seems a more reasonable proportion of the two million students in the **population** than a **sample** that contains only 5 students. The 500 student **sample** has a better shot at adequately representing the entire **population** than does the 5 student **sample**, right? You can see that intuitively… and once you learn the **central limit theorem**, you will see the mathematical demonstration of the importance of **sample** size for representing the **population**.

To conduct the inferential statistics we are using in this course, we will be using the normal curve model to estimate **probabilities** associated with particular scores. To do that, we need to assume that data are normally distributed. However, in real life, our data are almost never actually a perfect match for the normal curve.

So how can we reasonably make the normality assumption? Here's the thing. The **central limit theorem** is a mathematical principle that assures us that the normality assumption is a reasonable one as long as we have a decent **sample** size.

According to the theorem, as long as we take a decent-sized **sample**, if we took many **samples** (10,000) of large enough size (30+) and took the mean each time, the distribution of those means will approach a normal distribution, even if the scores from each **sample** are not normally distributed. To see this for yourself, take a look at the histograms shown on the right. The top histogram came from taking from a **population** 10,000 **samples** of just one score each, and plotting them on a histogram. See how it has a flat, or rectangular shape? No way we could call that a shape approximating a normal curve. Next is a histogram that came from taking the means of 10,000 **samples**, if each **sample** included 4 scores. Looks slightly better, but still not very convincing. With a **sample** size of 7, it looks a bit better. Once our **sample** size is 10, we at least have something pretty close. Mathematically speaking, as long as the **sample** size is no smaller than 30, then the assumption of normality holds. The other way we can reasonably make the normality assumption is if we know the population itself follows a normal curve. In that case, even if individual **samples** do not have a nice shaped



histogram, that is okay, because the normality assumption is one apply to the **population** in question, not to the **sample** itself.

Now, you can play around with an online demonstration so you can really convince yourself that the **central limit theorem** works in practice. The goal here is to see what **sample** size is sufficient to generate a histogram that closely approximates a normal curve. And to trust that even if real-life data look wonky, the normal curve may still be a reasonable model for data analysis for purposes of inference.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-36

4b. Hypothesis Testing

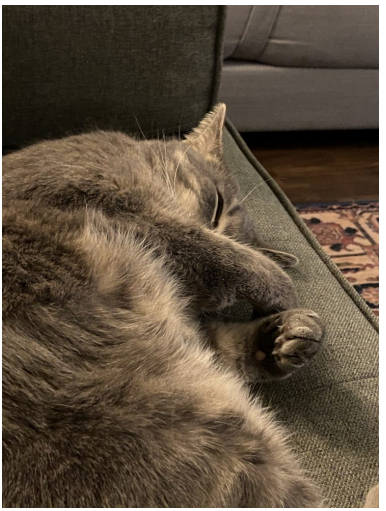
We are finally ready for your first introduction to a formal decision making procedure often used in statistics, known as **hypothesis testing**.

In this course, we started off with descriptive statistics, so that you would become familiar with ways to summarize the important characteristics of datasets. Then we explored the concepts standardizing scores, and relating those to probability as area under the normal curve model. With all those tools, we are now ready to make something!

"(not my) toolbox" by erix! is licensed under CC BY 2.0 "Dovetail Dresser" by Didriks is licensed under CC BY 2.0

Okay, not furniture, exactly, but decisions.

We are now into the portion of the course that deals with inferential statistics. Just to get you thinking in terms of making decisions on the basis of data, let us take a slightly silly example. Suppose I have discovered a pill that cures hangovers!



Well, it greatly lessened symptoms of hangover in 10 of the 15 people I tested it on. I am charging 50 dollars per pill. Will you buy it the next time you go out for a night of drinking? Or recommend it to a friend? … If you said yes, I wonder if you are thinking very critically? Should we think about the cost-benefit ratio here on the basis of what information you have? If you said no, I bet some of the doubts I bring up popped to your mind as well. If 10 out of 15 people saw lessened symptoms, that's 2/3 of people – so some people saw no benefits. Also, what does "greatly

lessened symptoms of hangover" mean? Which symptoms? How much is greatly? Was the reduction by two or more standard deviations from the mean? Or was it less than one standard deviation improvement? Given the cost of 50 dollars per pill, I have to say I would be skeptical about buying it without seeing some statistics!

On this list is a preview of the basic concepts to which you will be introduced as we go through the rest of this chapter.

Hypothesis Testing Basic Concepts

- Hypothesis
- Null Hypothesis
- Research Hypothesis (alternative hypothesis)
- Statistical significance
- Conventional levels of significance
- Cutoff sample score (critical value)
- Directional vs. non-directional hypotheses
- One-tailed and two-tailed tests
- Type I and Type II errors

You can see that there are lots of new concepts to master. In my experience, each concept makes the most sense in context, within its place in the **hypothesis testing** workflow. We will start with defining our **null** and **research hypotheses**, then discuss the **levels of statistical significance** and their conventional usage. Next, we will look at how to find the **cutoff sample score** that will form the critical value for our decision criterion. We will look at how that differs for **directional** vs. **non-directional hypotheses**, which will lend themselves to **one-** or **two-tailed tests**, respectively.

The **hypothesis testing** procedure, or workflow, can be broken down into five discrete steps.

Steps of Hypothesis Testing

1. Restate question as a research hypothesis and a null hypothesis about populations.
2. Determine characteristics of the comparison distribution.
3. Determine the cutoff sample score on the comparison distribution at which the null hypothesis should be rejected.
4. Determine your sample's score on the comparison distribution.
5. Decide whether to reject the null hypothesis.

These steps are something we will be using pretty much the rest of the semester, so it is worth memorizing them now. My favourite approach to that is to create a mnemonic device. I recommend the following key words from which to form your mnemonic device: hypothesis, characteristics, cutoff, score, and decide. Not very memorable? Try association those with more memorable words that start with the same letter or sound. How about "*Happy Chickens Cure Sad Days*." Or you can put the words into a mnemonic device generator on the internet and get something truly bizarre. I just tried one and got "*Hairless Carsick Chewbacca Slapped Demons*". Another good one: "*Hamlet Chose Cranky Sushi Drunkenly*." Anyway, you play around with it or brainstorm until you hit upon one that works for you. Who knew statistics could be this much fun!

The first step in **hypothesis testing** is always to formulate hypotheses. The first rule that will help you do so correctly, is that hypotheses are always about **populations**. We study samples in order to make conclusions about populations, so our predictions should be about the populations themselves. First, we define **population** 1 and **population** 2. **Population** 1 is always defined as people like the ones in

our research study, the ones we are truly interested in. **Population** 2 is the comparison **population**, the status quo to which we are looking to compare our research **population**. Now, remember, when referring to **populations**, we always use Greek letters. So if we formulate our hypotheses in symbols, we need to use Greek letters.

Population	Greek Letter	μ	σ
Sample	English Letter	M	SD

It is a good idea to state our hypotheses both in symbols and in words. We need to make them specific and disprovable. If you follow my tips, you will have it down with just a little practice.

We need to state two hypotheses. First, we state the **research hypothesis**, which is sometimes referred to as the alternative hypothesis. The **research hypothesis** (often called the alternative hypothesis) is a statement of inequality, or that Something happened! This hypothesis makes the prediction that the **population** from which the research sample came is different from the comparison **population**. In other words, there is a really high **probability** that the sample comes from a different distribution than the comparison one.

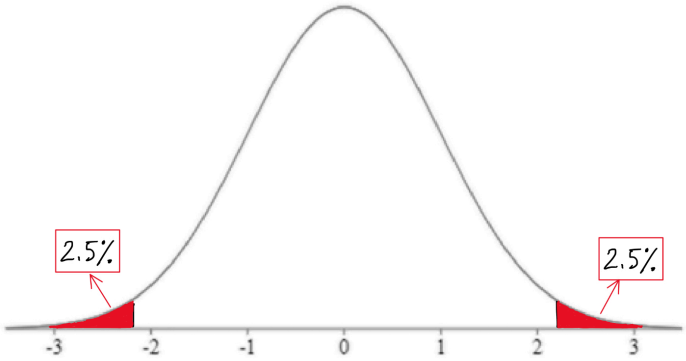
The **null hypothesis**, on the other hand, is a statement of equality, or that nothing happened. This hypothesis makes the prediction that the **population** from which sample came is not different from the comparison **population**. We set up the **null hypothesis** as a so-called straw man, that we hope to tear down. Just remember, null means nothing – that nothing is different between the **populations**.

Step two of hypothesis testing is to determine the characteristics of the comparison distribution. This is where our descriptive statistics, the mean and standard deviation, come in. We need to ensure our normal curve model to which we are comparing our research **sample** is mapped out according to the particular characteristics of the **population** of comparison, which is **population** 2.

Next it is time to set our decision rule. Step 3 is to determine the **cutoff sample score**, which is derived from two pieces of information. The first is the conventional **significance level** that applies. By convention, the **probability** level that we are willing to accept as a

risk that the score from our research sample might occur by random chance within the comparison distribution is set to one of three levels: 10%, 5%, or 1%. The most common choice of **significance level** is 5%. Typically the **significance level** will be provided to you in the problem for your statistics courses, but if it is not, just default to a **significance level** of .05. Sometimes researchers will choose a more conservative **significance level**, like 1%, if they are particularly risk averse. If the researcher chooses a 10% **significance level**, they are likely conducting a more exploratory study, perhaps a pilot study, and are not too worried about the **probability** that the score might be fairly common under the comparison distribution.

The second piece of information we need to know in order to find our **cutoff sample score** is which tail we are looking at. Is this a **directional hypothesis**, and thus **one-tailed test**? Or a **non-directional hypothesis**, and thus a **two-tailed test**? This depends on the **research hypothesis** from step 1. Look for directional keywords in the problem. If the researcher prediction involves words like "greater than" or "larger than", this signals that we should be doing a **one-tailed test** and that our **cutoff sample score** should be in the top tail of the distribution. If the researcher prediction involves words like "lower than" or "smaller than", this signals that we should be doing a **one-tailed test** and that our **cutoff sample score** should be in the bottom tail of the distribution. If the prediction is neutral in directionality, and uses a word like "different", that signals a **non-directional hypothesis**. In that case, we would need to use a two-tailed test, and our cutoff scores would need to be indicated on both tails of the distribution. To do that, we take our area under the curve, which matches our **significance level**, and split it into both tails.



For example, if we have a **two-tailed test** with a .05 **significance level**, then we would split the 5% area under the curve into the two tails, so two and a half percent in each tail.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-26

We can find the Z-score that forms the border of the tail area we have identified based on **significance level** and directionality by looking it up in a table or an online calculator. I always recommend mapping this **cutoff score** onto a drawing of the comparison distribution as shown above. This should help you visualize the setup of the **hypothesis test** clearly and accurately.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-22

The next step in the **hypothesis testing** procedure is to determine your **sample's** score on the comparison distribution. To do this, we calculate a test statistic from the **sample** raw score, mark it on the comparison distribution, and determine whether it falls in the shaded tail or not. In reality, we would always have a sample with more than one score in it. However, for the sake of keeping our test statistic formula a familiar one, we will use a sample size of one. We will use our Z-score formula to translate the **sample's** raw score into a Z-score – in other words, we will figure out how many standard deviations above or below the comparison distribution's mean the **sample** score is.

$$Z = \frac{X - M}{SD}$$

Finally, it's time to decide whether to **reject the null hypothesis**. This decision is based on whether our sample's data point was more extreme than the **cutoff score**, in other words, "did it fall in the shaded tail?" If the **sample** score is more extreme than the **cutoff score**, then we must reject the null hypothesis. Our research hypothesis is supported! (Not proven... remember, there is still some probability that that score could have occurred randomly within the comparison distribution.) But it is sound to say that it appears quite likely that the population from which our sample came is different from the comparison population. Another way to express this decision is to say that the result was **statistically significant**, which is to say that there is less than a 5% chance of this result occurring randomly within the comparison distribution (here I just filled in the blank with the significance level).

What if the research sample score did not fall in the shaded tail? In the case that the sample score is less extreme than the **cutoff score**,

then our **research hypothesis** is not supported. We **do not reject the null hypothesis**. It appears that the **population** from which our **sample** came is not different from the comparison **population**. Note that we do not typically express this result as "accept the null hypothesis" or "we have proved the null hypothesis". From this test, we do not have evidence that the **null hypothesis** is correct, rather we simply did not have enough evidence to reject it. Another way to express this decision is to say that the result was not **statistically significant**, which is to say that there is more than a 5% chance of this result occurring randomly within the comparison distribution (here I just used the most common **significance level**).

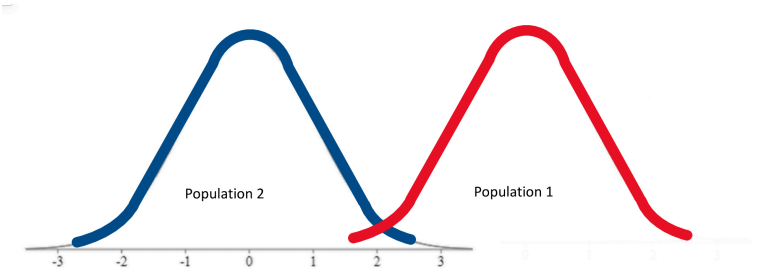
An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-23

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-24

An interactive H5P element has been excluded from this version of the text. You can view it online here:

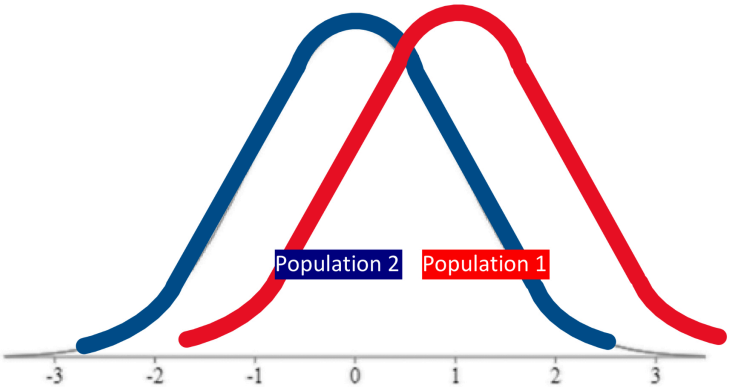
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-25

So we have described the **hypothesis testing** process from beginning to end. The whole process of null **hypothesis testing** can feel like pretty tortured logic at first. So let us zoom out, and look at the whole process another way. Essentially what we are seeking to do in such a **hypothesis test** is to compare two **populations**. We want to find out if the **populations** are distinct enough to confidently state that there is a difference between **population** 1 and **population** 2. In our example, we wanted to know if the **population** of people using a new medication, **population** 1, sleep longer than the **population** of people who are not using that new medication, **population** 2. We ended up finding that the research evidence to hand suggests **population** 1's distribution is distinct enough from **population** 2 that we could **reject the null hypothesis** of similarity.



In other words, we were able to conclude that the difference between the centres of the two distributions was **statistically significant**.

If, on the other hand, the distributions were a bit less distinct, we would not have been able to make that claim of a significant difference.



We would not have **rejected the null hypothesis** if evidence indicated the **populations** were too similar.

Just how different do the two distributions need to be? That criterion is set by the **cutoff score**, which depends on the **significance level**, and whether it is a **one-tailed** or **two-tailed hypothesis test**.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=226#h5p-29

That was a lot of new concepts to take on! As a reward, assuming you enjoy memes, there are a plethora of statistics memes, some of which you may find funny now that you have made it into inferential statistics territory. Welcome to the exclusive club of people who have this rather peculiar sense of humour. Enjoy!

Chapter Summary

In this chapter we examined **probability** and how it can be used to make inferences about data in the framework of **hypothesis testing**. We now have a sense of how two **populations** can be compared and the difference between their means evaluated for **statistical significance**.

Key Terms:

probability	**research hypothesis**	**one-tailed test**
central limit theorem	**null hypothesis**	**non-directional hypothesis**
population	**cutoff sample score**	**statistical significance**
sample	**significance level**	**reject the null hypothesis**
hypothesis testing	**directional hypothesis**	**do not reject the null hypothesis**

5. Single Sample Z-test and t-test

5a. Single Sample Z-test

Now that you have mastered the basic process of hypothesis testing, you are ready for this: your first real statistical test. First, we will examine the types of error that can arise in the context of hypothesis testing.

Which type of error is more serious for a professional? 1. I decide that a new mode of transportation, the driverless car, is much safer than current modes of transportation, and recommend to the state that they force conversion to that new mode. But then it turns out after a year that the rates of accidents actually increased.

"Google Self-Driving Car" by smoothgroover22 is licensed under CC BY-SA 2.0

Or... 2. I decide that a new mode of transportation is safer than current modes, but I do not have enough confidence in that margin of safety and therefore do not recommend a sweeping change. After a year it

becomes clear that driverless cars are indeed much safer, and a switch would have been beneficial.

If you selected the first error as the more serious one, then you chose a **Type I error**. This is the type of error that is considered more serious in the realm of statistical decision making, as well.

In the process of hypothesis testing, the decision you make is all about probabilities. It is an educated guess. However, there is always room for error. In fact, there are two types of errors you can make in hypothesis testing. Their imaginative labels are **Type I** and **Type II error**. As we saw, **Type I error** is considered more serious, so the Significance level is typically set with an eye toward the probability of **Type I error** that is deemed acceptable in that study. α = .05 indicates that the researcher accepts a 5% chance of a **Type I error**. In the decision matrix shown here, you can see how the hypothesis test can play out in four different ways.

		Real Situation	
		Null Hypothesis True	Research Hypothesis True
Decision	Reject Null Hypothesis	**Type I error**: *You went out on a limb and were wrong.* **α**	Correct Decision
	Do not reject null hypothesis	Correct Decision	**Type II error**: *You were too conservative and were wrong.* **β**

Note that the real situation, whether the null hypothesis is true or the research hypothesis is true, is unknown. We can never know which is true – that is why they are hypotheses, after all!

If, hypothetically, the null were true, and if we made the decision to reject the null hypothesis, then we would land in this corner of the decision matrix: **Type I error**. This is what happens if we go out on a limb, reject the null hypothesis, but we are wrong. As researchers we live in fear of making a **Type I error**. No one wants to make a big thing about their research findings and promote change to a new, different,

risky, expensive thing, and then be wrong about it. On the other hand, if the null were true and we did not reject it, then we would be correct.

Hypothetically, let us suppose the research hypothesis were true. If we make the decision to reject the null hypothesis, then we are correct. However, if we decide not to reject the null hypothesis, we would be making a **Type II error**. This is when we are too conservative, decide there is not enough evidence that the research (alternative) hypothesis could be correct, but we are wrong. **Type II error** is also not great, so we do try to minimize that probability by doing something called a power analysis. However, researchers generally would rather make a **Type II** than a **Type I** error. Keeping the status quo may not do much for to make your scientific career exciting, but objectively speaking it is not such a scary proposition.

Now, a key thing to remember when you refer back to this decision matrix to determine for which error you are at risk, you never know which column you are in, because the real situation is always unknown. What determines the possible type of error is the thing you have control over: the decision. So you can switch the row you will place yourself in, depending on the observed data, and thus remove yourself from unreasonable risk of **Type I error**. (The probability of making a **Type I error** is symbolized as $\boldsymbol{\alpha}$, so significance levels are often referred to as $\boldsymbol{\alpha}$.)

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=318#h5p-31

An interactive H5P element has been excluded from this version of the text. You can view it online

here:
https://pressbooks.bccampus.ca/
statspsych/?p=318#h5p-32

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/
statspsych/?p=318#h5p-33

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/
statspsych/?p=318#h5p-34

Okay, we are about to make a connection that is intellectually a bit challenging. So prepare yourself: we will be drawing connections between the central limit theorem, sampling distributions, and the appropriate type of comparison distribution we should use for a hypothesis test where the sample is made up of more than one data point.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=318#h5p-37

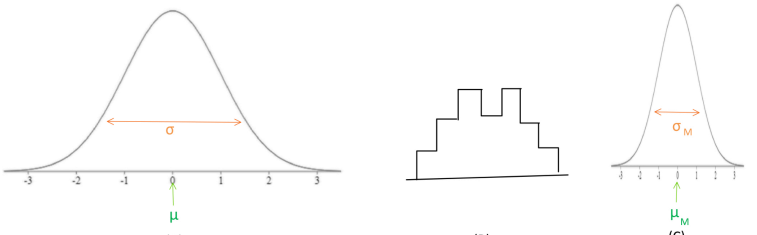
An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=318#h5p-38

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=318#h5p-39

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://pressbooks.bccampus.ca/statspsych/?p=318#h5p-40

In the past, we have been using the normal distribution made of up individual scores as our comparison distribution. In the illustration below, if distribution (A) is the distribution of individual scores, (B) is the histogram of a sample of scores drawn from (A). (C) is the **distribution of means** from many samples such as the one portrayed in (B).



(A) Distribution of individual scores

(B)

(C) Distribution of means

The shape of a **distribution of means** is narrower than the distribution of individual scores from which it came. The variance is smaller, because it is divided by N, the sample size:

$$\sigma_M^2 = \frac{\sigma^2}{N}$$

The standard deviation is smaller, because it is divided by the square root of N:

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

The means are the same for both types of distributions:

$$\mu_M = \mu$$

Up until now we worked with individuals as our samples, so we could use our old Z-score formula and compare the sample score to a comparison distribution of individuals. We did this, so that we could avoid adding new symbols to our formulas, and concentrate on the hypothesis testing logic. Now, we want the more realistic situation, so we will use the **Z-test** to compare a sample mean to a comparison **distribution of sample means**. To rephrase, when we had an individual score, we used the distribution of individuals. When we have a sample mean, we need to use the **distribution of sample means**. The new **Z-test** statistic formula is shown here:

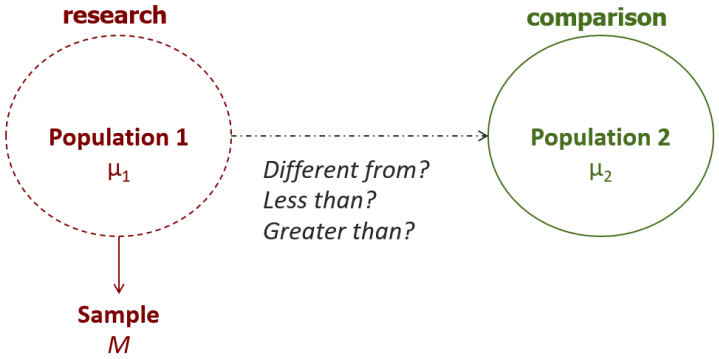
$$Z = \frac{M - \mu_M}{\sigma_M}$$

In this formula, M is the sample mean, μ_M is the comparison distribution mean, and σ_M is the comparison distribution standard deviation.

We will be using the terminology "**distribution of means**," because it's a name that reminds us of how it differs from the distribution of individuals. A term in statistics that means the same thing is "sampling distribution." Why do we need this **distribution of means**? In reality, we do not compare an individual sample score against a population distribution of individuals. For a real statistical test, we collect a sample and calculate a mean, and compare that sample mean against a comparison distribution of means. We will ground that in an example: If I want to know if my introductory psychology class is collectively performing at the expected level on a test, I do not take their average performance and compare that mean against a set of individual scores, I want to compare it to a set of other class averages. How does this impact our hypothesis test? It matters for Step 2, determining the characteristics of the comparison distribution. Well, it does not change anything about the comparison mean, because the mean of a **distribution of means** is equal to the mean of the distribution of individuals. But the standard deviation now needs to be converted into the standard deviation of the **distribution of means**, which is often called the **standard error of the mean**. As these equations above show, the variance of a **distribution of means** is equal to the variance of the distribution of individuals divided by the number of individuals in the sample. And thus, if we take the square root of the variance, the standard deviation of a **distribution of means** is equal to the standard

deviation of the distribution of individuals divided by the square root of the number of individuals in the sample.

We will now examine the way the steps of the hypothesis test procedure play out when using the single-sample **Z-test**. First, we need to formulate the research and null hypotheses. Remember, in such a hypothesis test, we are trying to compare two populations. Population 1, the research population, is the one from which we have drawn a research sample.



We do not have access to the actual research population mean, hence the need to calculate a sample mean. That sample mean is our best proxy for the population 1 mean, which we need to compare to the mean of population 2. The nature of the comparison, in particular its directionality, is determined by the researcher's prediction.

In step 2, we need to determine the characteristics of the comparison distribution. This means that we need the mean and standard deviation of the comparison distribution, which represents Population 2. This is where we need to introduce the new **distribution of means**. The mean is the same as provided, but the standard deviation needs to be converted.

If the variance is provided, you can start with this formula to convert to the distribution of means, then take the square root to get the standard deviation:

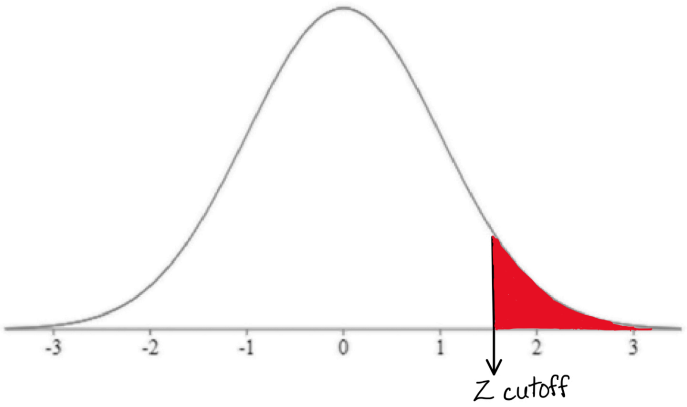
$$\sigma_M^2 = \frac{\sigma^2}{N}$$

If the standard deviation is provided, it is easiest to use this formula:

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

Step 3 is to determine the cutoff sample score, which is derived from two pieces of information. First, we need to know the conventional significance level that applies. Second, we need to know which tail we are looking at. The bottom, the top, or both? This depends on the research hypothesis, so we should always look for directional keywords in problem. If the research hypothesis is directional, we should use a one-tailed test. If the research hypothesis is non-directional, we should use a two-tailed test.

Once we have those two pieces of information, we can find the Z-score that forms the border of the shaded tail by looking it up in a table. We can then map the cutoff score onto a drawing of the comparison distribution.

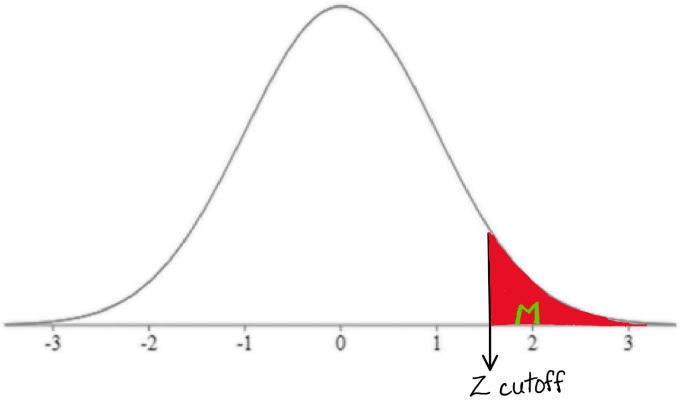


Example of comparison distribution sketch showing Z-score cutoff and shaded tail representing decision rule

In step 4, we need to calculate the Z-test score from the sample mean:

$$Z = \frac{M - \mu_M}{\sigma_M}$$

In this formula, M represent the research sample mean, μ_M represents the comparison population mean, and σ_M represents the standard deviation of the comparison population. Remember, we are using the **distribution of means** as our comparison distribution. Once we have calculated the Z-test score, we will mark where it falls on the comparison distribution, and determine whether it falls in the shaded tail or not.



Example of mapping sample mean (M) onto comparison distribution based on Z-test calculation

Finally, it is time to decide whether or not to reject the null hypothesis. If the sample mean falls within the shaded tail, we reject the null hypothesis. If it does not, we do not reject the null hypothesis. In other words if the sample mean is extreme enough relative to the comparison distribution of means, then we reject the null hypothesis.

Once we make our decision, we will need to take a close look at what kind of error we might have made as a result of our decision. And remember, from our decision matrix, all we have control over whether we're on the top row or the bottom row.

		Real Situation	
		Null Hypothesis True	Research Hypothesis True
Decision	Reject Null Hypothesis	**Type I error**: *You went out on a limb and were wrong.* **α**	Correct Decision
	Do not reject null hypothesis	Correct Decision	**Type II error**: *You were too conservative and were wrong.* **β**

We do not know the real situation. if we do make the conclusion that we should reject the null hypothesis, then we are at some risk of a **Type I error**, but as long as we don't exceed our significance level, we accept that risk. If we do not reject the null hypothesis we could be at a risk for **Type II error**.

Typically, after conducting a hypothesis test, researchers want to obtain the **p-value**, or the probability of the observed sample score or more extreme occurring just by random chance under the comparison distribution. The p-value associated with the sample mean has to be less than the significance level in order to reject the null hypothesis. A common way to find the **p-value** is to use an online calculator.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=318#h5p-41

One last exercise I would like you to try is to pretend you are publishing these results in a scientific journal. So you might write this sentence in your results section based on your findings:

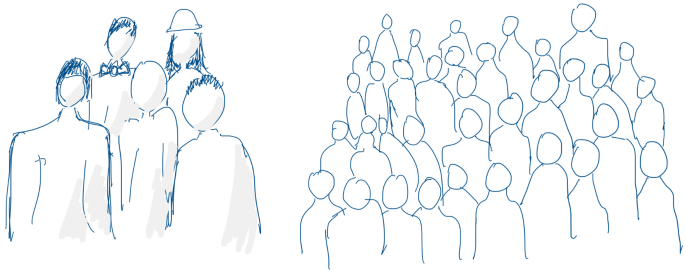
“We found that patients with insomnia who received a new drug slept significantly more than patients who did not receive the new drug (p < .05).”

Notice the term “significantly” is used in this example, because the null hypothesis was rejected at the end of the hypothesis test. Also, the **p-value** is in brackets at the end of this sentence, before the punctuation, just like you would use a citation to back up a factual claim… only here, it is a claim of statistical significance.

5b. Single-sample T-test

Now it is time to introduce your second real statistical test: the single-sample **t-test**. This test is much like the **Z-test**, but more commonly used in real life.

Before we get started on the **t-test**, I am curious… would you have greater confidence in a statistic that comes from a small sample?



Or in one that comes from an entire population?… If you said that you would have greater confidence in a number that comes from an entire population, then the field of statistics would agree with you.

Sometimes, though, we do not have access to all the data from population that we need, and we have to take a best guess based on a sample. In that case, we probably want to be more conservative in our decision making. That’s what a **t-test** is for.

Often, even if we have information regarding the comparison population mean, we do not have access to the population standard deviation (σ). In this situation, what we must do is collect a sample reflecting population 1, calculate its mean and standard deviation, and compare that sample mean against a comparison **distribution of means** using an estimate (S_M) of the population (σ_M). Note the change in symbol (from σ to *S)* to reflect the fact that we are basing our estimate on sample data, rather than using a known population parameter.

$$S^2 = \frac{\sum(X - M)^2}{N - 1}$$

You may have noticed that this new formula for calculating variance in a sample has on the bottom "*N* minus 1" rather than just "*N*" as our old calculation formula had. Why "*N* minus 1"? This is a correction, based on sample size, that derives from the concept of **degrees of freedom**. We will return to that concept later. For now, let us consider "*N* minus 1" as a correction based on sample size that has the following deliberate and intentional effects on our calculation. If we have a small sample size, in which we should not have a great deal of confidence, subtracting 1 really affects the calculation. For example, if we had an *N* of 2, subtracting off 1 from that sample size will mean that we divide by 1 instead of 2. That has a drastic impact on the variance calculation, in effect doubling our estimated variance. Of course in the context of hypothesis testing, more variance is bad – in that we are less likely to be able to make conclusions based on the data.

On the other hand, with a larger sample size, subtracting 1 will have little impact on our calculation... dividing by 100 or dividing by 99 will have nearly identical results. Think of it this way – if we have a small sample size, we are punishing ourselves, like a handicap in a golf game. With a larger sample size, we reward ourselves, boosting our chance of a conclusion statistical decision.

Subtracting off 1 is like a flat tax – a flat income tax of 10% for all Canadians would be harder on those with a low income, who would have even less money to meet basic survival needs, whereas the wealthiest Canadians would still have plenty of money for their needs. That is, of course, why democratic governments typically use progressive tax rates, because those who have more can afford to pay a greater proportion of their income and still have enough to cover

their basic costs of living. With hypothesis testing, though, the flat tax is desirable, because it is our intent to penalize the smallest sample sizes the most, given the fact that we can have little confidence in the estimates we derive from them.

The t-test is designed around the concept of **degrees of freedom**, which is defined as the number of scores in a given calculation that are free to vary. This is an abstract definition that is difficult to grasp, so it is helpful to consider a concrete example.

Degrees of Freedom Practical Example

First, consider the game of baseball. We understand the field-of-play consists of 9 positions. The coach is "free" to assign any of the 9 players to any of the 9 positions. Once the 8th player is assigned to the 8th position, the 9th player-position is pre-determined, so to speak. In other words, the coach is not "free" to pick either the last position or the last player.

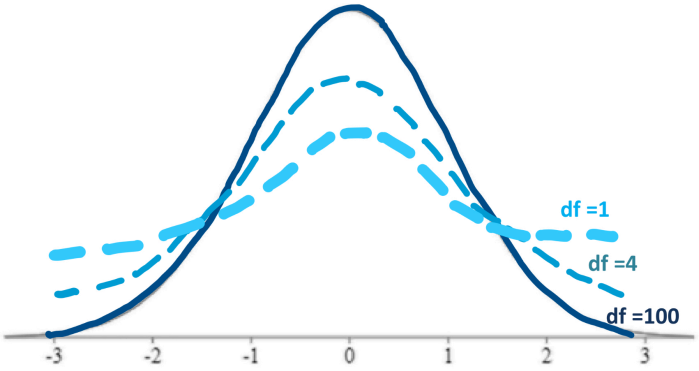
Source: https://www.isixsigma.com/ask-dr-mikel-harry/ask-tools-techniques/can-you-explain-degrees-freedom-and-provide-example/

The same logic holds when doing a calculation involving all of a sample's scores. Up until the last score inputted into the formula, the scores are free to vary; but the last score is pre-determined. Thus **degrees of freedom** in a **t-test** is *N*-1, or 1 less than the total number of scores in the dataset.

In a **t-test** we will be using the **t-distributions**, rather than a normal distribution, as our comparison distributions. The **t-distributions** are a series of distributions, based on the normal distribution, that adjust

their shape according to **degrees of freedom** (which in turn is based on sample size).

As mentioned previously, when we do not have the population standard deviation we must estimate it from our sample. This involves a greater risk of error. The risk is related to how large our sample is. So, the fewer the scores in our sample, the fewer the degrees of freedom, and the more conservative the distribution we must use – one that is wider and shorter (more area in the tails).



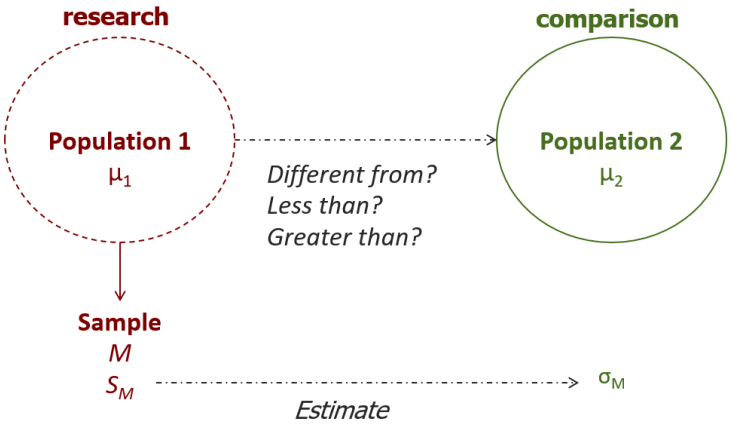
Above you can see examples of how the shape of the t-distributions changes with reduced **degrees of freedom**. With **degrees of freedom** of 100, the shape of the distribution is very close to the normal curve. If we have very few **degrees of freedom**, like 1, then our shading rejection regions, or goal posts, will move farther out into the tails of the distribution, because there is more area under the tails of the curve. In effect, then, with fewer **degrees of freedom**, we will find it more difficult to reject the null hypothesis. This is, of course, by design, in order to penalize the hypothesis test if it is based on a very small sample size.

Inspiration for this creative illustration came from one of my former students, Andi. Blame for artwork quality is entirely on me.

Above is a less technical visual to see the difference in distribution shape when the sample size is small, versus when the sample size is large. As you can see, when you have a large sample size, the goal posts are very close and easy to reach, but when you have a small sample size, they will really be far away, and this will make it difficult to reject our null hypothesis.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=318#h5p-43

Now we are ready to take a look at how a hypothesis test works based on a single-sample **t-test**. As before, the structure of the test is familiar – we are comparing the means of Population 1 and Population 2.



As usual, we do not have access to the mean of population 1, just the mean of a sample from that population. Furthermore, unlike in a **Z-test**, we do not have access to one piece of information about population 2: the population standard deviation. Therefore, we will form hypotheses the same way, but when we do steps 2 and 4, we will need to use the sample to generate an estimate of the comparison population standard deviation.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=318#h5p-44

In step 2, we identify the known comparison population's mean. However, we will need to use the sample to generate an estimate of the comparison population standard deviation. We calculate the sample variance, using the formula with the "*N*-1" correction,

$$S^2 = \frac{\sum(X - M)^2}{N - 1}$$

and then convert the variance to the **distribution of means** by dividing by *N*,

$$S_M^2 = \frac{S^2}{N}$$

and finally we square root to get from variance to standard deviation.

$$S_M = \sqrt{S_M^2}$$

S_M is what we will use to describe our comparison distribution.

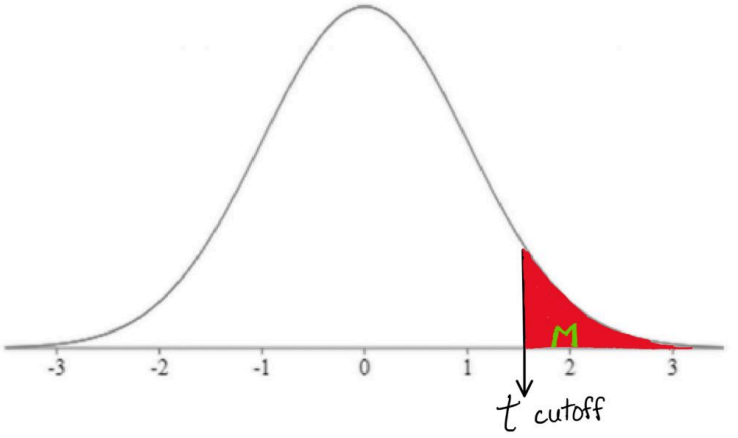
In step 3, as before we need to find our cutoff sample score based on the significance level and directionality of the test. Now, however, we also need to use a third piece of information: **degrees of freedom**. This is found by subtracting one from the sample size. Then we can look in the t-tables, identifying the row of the table with the matching **degrees of freedom**, then looking in the appropriate column based on directionality and significance level. Once we identify the t-score that forms the border of the shaded tail, we can map this onto a drawing of the comparison distribution. For visualization purposes, it is fine to sketch the distribution the same way you did the normal distribution.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=318#h5p-42

Now for step 4. The **t-test** formula is just like the **Z-test** one on top – sample mean minus comparison population mean. Then you divide by S_M, the estimate of standard deviation that you calculated back in step 2.

$$t = \frac{M - \mu_M}{S_M}$$

If your calculated **t-test** score fell into the shaded tail beyond your cutoff score, then you may reject the null hypothesis. In other words, if the sample mean was extreme enough on the comparison distribution of means, reject the null.



After conducting a hypothesis test, it is helpful to also obtain a report the precise **p-value** associated with the result. The **p-value** offers the precise probability of obtaining the **t-test** result (or more extreme) just by random chance under the comparison distribution. It is found by calculating the area under the comparison distribution in the tail beyond the observed **t-test** score calculated in step 4 (or in both tails, if it was a two-tailed test).

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://pressbooks.bccampus.ca/statspsych/?p=318#h5p-45

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=318#h5p-46

Chapter Summary

In this chapter we learned about the two types of logical errors we can make in hypothesis testing: **Type I** and **Type II errors**. We saw that **Type I error** is considered more serious in statistics, because it represents going out on limb, making a strong conclusion, when that is in fact the wrong decision. For that reason, **Type I error** risk is strictly controlled in inferential statistics by setting a particular significance level (***α***). We also introduced our first real inferential statistical tests: the single-sample **Z-test** and the single-sample **t-test**. In each of these tests, we are comparing the means of two populations, using a sample to estimate the mean of the research population. A **Z-test** is used when we know the standard deviation of the comparison population (***σ***); a **t-test** is used when we do not have that information and must estimate the standard deviation from the sample (***S***).

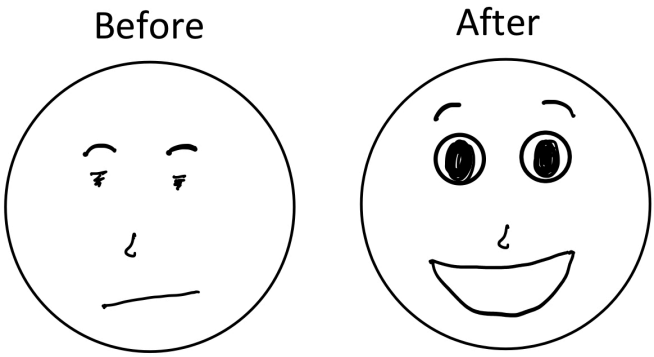
Key terms:

Type I error	**distribution of means**	**p-value**
Type II error	**Z-test**	**t-test**
α	**standard error of the mean**	**degrees of freedom**

6. Dependent t-test

6. Dependent t-test

This chapter will introduce you to the **dependent means t-test**, which is most often used for experiments with a repeated measures design. Repeated measures designs are very common experimental approaches, just as posts on social media often compare before and after images to show off the effects of a diet or a renovation.

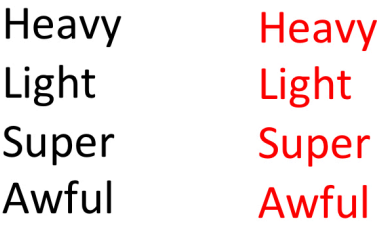


Actual footage of your instructor upon waking in the morning, demonstrating the qualitative effects of coffee

Repeated measures experimental designs are also known as within-subjects designs. Such an experiment involves obtaining two separate scores for each individual in the sample. Instead of having an experimental group and control group, there is just one sample, from which the same participants are used in all treatment conditions. Typically this kind of study uses a pre-test post-test design.

As an example, perhaps I want to see if memory span is affected by the colour in which items are presented. I first test people on black and white items, then test them with red items. I will compare their

performance on the second test with their performance on the first test.



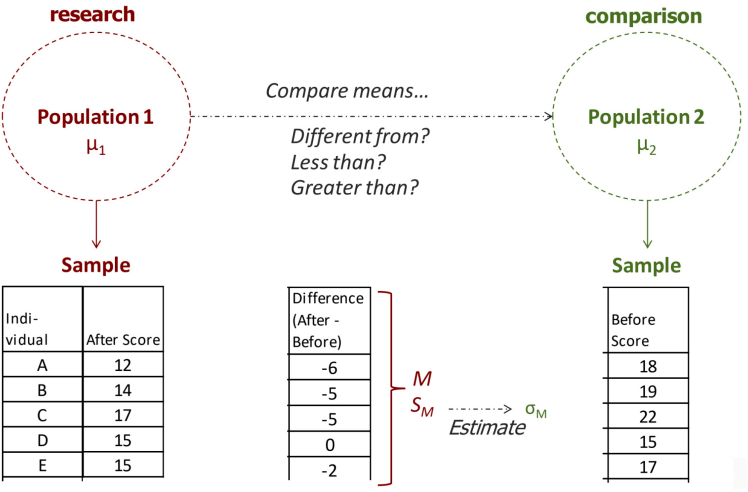
Another less-common type of experimental design that would be analyzed using a **dependent t-test** is the **matched pairs** design. In this type of approach, two separate samples are used, but each individual in a sample is matched one-to-one with an individual in the other sample. This is most commonly used when the researcher is intent on controlling for a possible confounding variable and thus matches participants based on that variable – for example, age or genetic relatedness. Because of this matching, the two samples are not independent, but rather they are related in some way. Hence the name "**dependent t-test**".

This may not be a familiar idea, so we will consider an example of a **matched pairs** design. Perhaps I want to test differences in memory capacity in an experimental group and compare to a control group, but I know age can greatly impact this type of memory. So, I make sure for each person aged 20 in one group there is another person with the same age in the other group, and so on, for each age. That way, I am getting the difference in memory scores for each **matched pair**, and thus age is explicitly controlled for as a possible third variable.

An interactive H5P element has been excluded from this version of the text. You can view it online here:

https://pressbooks.bccampus.ca/statspsych/?p=379#h5p-47

As before, the first step of the hypothesis test is to formulate hypotheses. The goal is to compare the means of two populations. This time, though, we not only have no standard deviation from population 2, we also have no mean. For both populations, we will rely on sample-based estimates for the mean and the standard deviation. In a **repeated measures experiment**, we would have one set of scores measured at baseline, the before scores. These scores will represent population 2, the comparison population. The scores that are measured after the experimental manipulation will represent the research population, population 1.



To conduct the comparison, we will actually take the difference score for each individual in the study, by subtracting the before score from the after score, and then calculate the mean and standard deviation of the difference scores. As an example, I made up some scores on

a mood measure for people after eating chocolate and before eating chocolate, and calculated after-before difference scores for each individual (shown above). Negative difference scores indicate mood scores went down after eating chocolate; positive difference scores indicate mood scores went up after eating chocolate. (Worry not, this is a completely invented dataset — it seems unlikely to me that eating chocolate would actually worsen most people's mood!)

In step 2, we need to approach the mean and standard deviation of the comparison distribution a little differently than before. The comparison distribution will now be the distribution of means for the population of difference scores, which are defined as after-minus-before scores. Under the comparison population, the mean of difference scores should be 0: under the null hypothesis, there is no no difference between those who ate chocolate and those who did not, for example, so there would be no change before to after. Thus we set the mean of the comparison population to 0.

$$\mu_M = \mu = 0$$

Now for the standard deviation: we will need to use the sample of difference scores to generate an estimate of the comparison population standard deviation. Perhaps you are wondering why we calculate difference scores as after-minus-before? This is important for the way we interpret the difference scores, and to fit the directionality of our hypothesis test. In our example, if people's mood scores worsen, this should result in a negative difference score, moving the mean toward the low end of the distribution, right? And if the mood improves that should result in a positive score. That is why we have to set up the difference scores as after-minus-before.

Indi-vidual	Before Score	After Score	Difference (After - Before)
A	18	12	-6
B	19	14	-5
C	22	17	-5
D	15	15	0
E	17	15	-2

If we look at these example data, right away by looking at the difference scores, we can tell the mood went down for almost every one, except for one person who did not change.

The formulas to calculate a sample-based estimate of the comparison distribution standard deviation are exactly the same as for a single sample t-test:

$$S^2 = \frac{(X - M)^2}{N - 1}$$

$$S_M^2 = \frac{S^2}{N}$$

$$S_M = \sqrt{S_M^2}$$

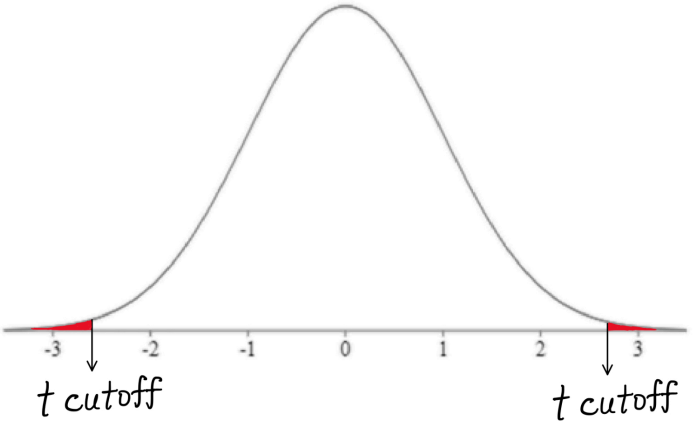
In fact, that is why I wanted to introduce you to the **dependent t-test** before we move on to the independent t-test, which has a different set of formulas.

Before we get to our example, I would like to note that once we calculate the difference scores for each individual in the sample, we will be using those difference scores to calculate the mean and standard deviation. We no longer need the before or after scores for anything. I recommend crossing them out, so you are not confused about which numbers to include as X in the formulas.

Indi-vidual	Before Score	After Score	Difference (After - Before)
A	18	12	-6
B	19	14	-5
C	22	17	-5
D	15	15	0
E	17	15	-2

These become our sample scores!

In step 3, we need to determine the cutoff sample score. As with the single sample t-test, this will be derived from three pieces of information: the significance level, the directionality, and degrees of freedom. We can use t-tables to find the cutoff score and map it onto our drawing of the comparison distribution.



Step 4 is the moment of truth – does the sample mean fall far enough from the comparison population mean to reject the null

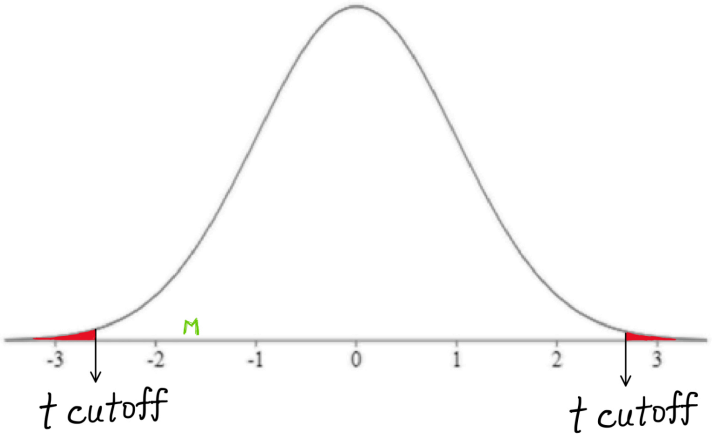
hypothesis? We can use the same t-test formula as for the single-sample t-test.

$$t = \frac{M - \mu_M}{S_M}$$

Remember, the comparison population mean is now set to zero, so we can use that in place of μ.

$$t = \frac{M - 0}{S_M}$$

Once we have calculated the t-test result, we can mark it on comparison distribution to determine whether it falls in shaded tail or not.



Finally, it's decision time. Did the sample mean of difference scores fall within a shaded tail on the comparison distribution? Is it extreme enough to reject the null?

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://pressbooks.bccampus.ca/statspsych/?p=379#h5p-48

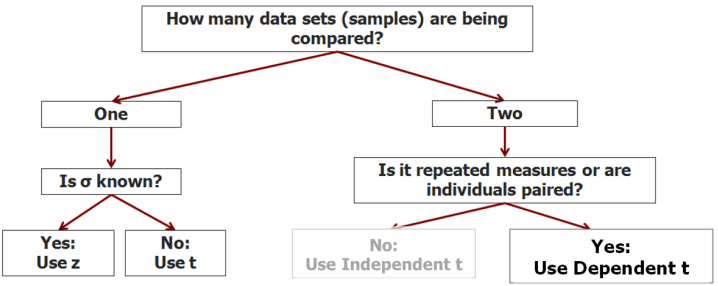
As always, we can also use a calculator find the p-value, the precise probability that this t-test score (or more extreme) would have occurred by random chance alone under the comparison distribution.

If you were writing these results for publication, how would you translate the hypothesis test into a concise sentence? Does the chocolate make people's mood change? In this example, we found that mood scores were *not* significantly different after people consumed chocolate.

We can support that statement with the information that the probability of the sample mean occurring on the comparison distribution was more than 5%. As a result we have an inconclusive hypothesis test.

"We found that mood scores were not significantly different after people consumed chocolate (p = 0.16)."

As we continue to build our decision tree, you can use it to guide your choice of a statistical test appropriate for a particular research design.



If you have two samples, and they are in some way related, like in **repeated measures** or **matched pairs** designs, that's when we should use the **dependent means t-test**. In the next chapter we will add the independent means t-test to our toolbox.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=379#h5p-49

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=379#h5p-50

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=379#h5p-51

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=379#h5p-52

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=379#h5p-53

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=379#h5p-54

Chapter Summary

In this chapter we introduced the use of the **dependent means t-test** in hypothesis tests for research designs such as **repeated measures** and **matched pairs**.

Key terms:

dependent means t-test **repeated measures** **matched pairs**

7. Independent Means t-test

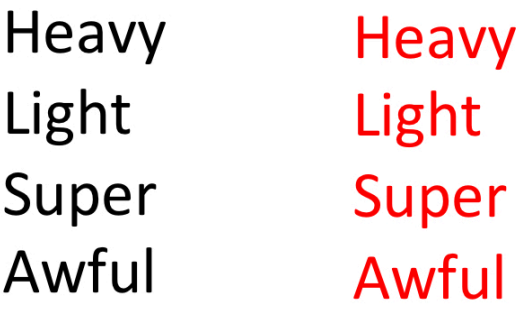
7. Independent Means t-test

In this chapter, I will introduce to you one last t-test variation – the **independent means t-test**. This one is intended for the classic experimental design, in which two independent samples are compared.

In a classic experimental design, we are comparing two samples. The **independent means t-test** is typically used to compare the data from an experimental group to those from a control group. The experimental group is the one that receives the manipulation, or the independent variable, and the control group is the one that receives either no manipulation, or an alternative one that represents the status quo – like a placebo. What makes this different from the repeated measures type of design, is that the scores from the two groups are

independent. They are obtained from different participants who are randomly assigned to one group or the other.

For example, if I want to see if memory span is affected by the colour in which items are presented. I test one group of people with black and white items, and test another group of people with red items.



I will compare one group average with the other group average. There is no relationship or dependency of one group of scores with the other group, so it will be appropriate to analyze the data with an **independent t-test**.

With all statistical tests, we know each sample comes from a population. The question is: are they different populations? To answer this question using statistical tests, we need to make some assumptions about the data. We have been making the **normal curve assumption** all along, and we saw how the central limit theorem can be used to justify this assumption when our samples are large enough. With this new kind of t-test, we are also going to make the **homoscedasticity assumption**: that the two populations we are comparing have the same variance. In an introductory course like this, we will not go into the technicalities of verifying this assumption, but it is possible to do so before we conduct the analysis.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=410#h5p-60

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=410#h5p-58

In the **independent means t-test**, just like with the dependent t-test, we have no direct information about population 1 or 2. We calculate sample means from our research and comparison samples. To find the standard deviation for the comparison population, we will take sample based estimates using all the scores we have to hand, by pooling together the variance of each sample. This makes sense if we are assuming the two populations have equal variance.

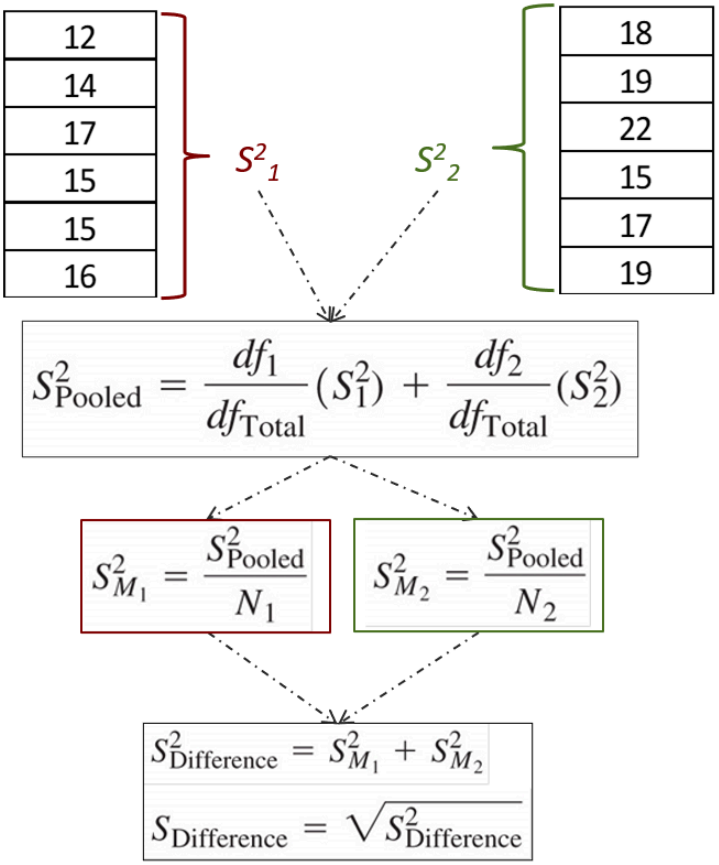
An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=410#h5p-59

In step 2, we will once again set the comparison population mean

to zero, because the comparison reflects the distribution of means under the null hypothesis, in which there is no difference between the populations.

$$\mu_M = \mu = 0$$

The standard deviation will be calculated through a workflow previous students have told me looks like an hourglass shape.



Starting from the top, we first calculate the variance of sample 1 and sample 2 separately.

$$S^2 = \frac{(X - M)^2}{N - 1}$$

We then pool the two variances together using a weighted average formula. This formula just allows us to count one variance more than the other if the sample size is bigger. With two independent samples, it is not uncommon to have an unequal N, or number of scores, in each group.

$$S^2_{Pooled} = \frac{df_1}{df_{Total}}(S^2_1) + \frac{df_2}{df_{Total}}(S^2_2)$$

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://pressbooks.bccampus.ca/statspsych/?p=410#h5p-63

Once we have the pooled variance calculated, we use that to convert to the variance of the distribution of the difference between means, which is the comparison distribution for this test.

$$S^2_M = \frac{S^2_{Pooled}}{N}$$

$$S^2_{Difference} = S^2_{M1} + S^2_{M2}$$

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=410#h5p-64

We then square root to get the estimated standard deviation for the comparison distribution.

$$S_{Difference} = \sqrt{S^2_{Difference}}$$

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=410#h5p-65

Now for step 3: the one new thing here is the degrees of freedom we will use to look up the cutoff sample score in the t tables. Because we have two samples, we will use the pooled, or total, degrees of freedom for lookup. That is the main advantage of the **independent means t-test**. Because we have two samples of scores, we get the benefit of more degrees of freedom.

An interactive H5P element has been excluded from this version of the text. You can view it online here:

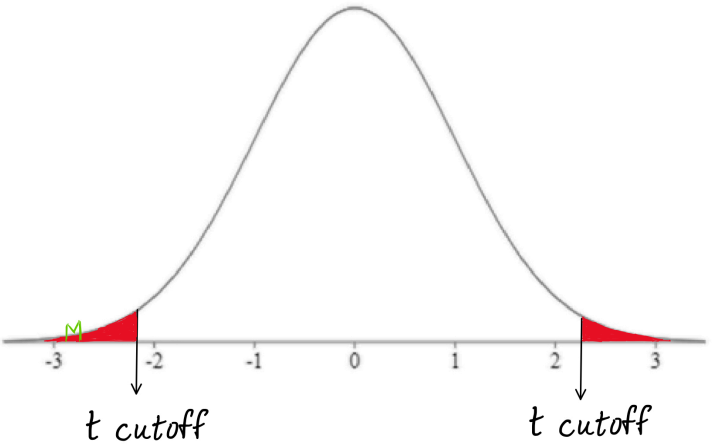
https://pressbooks.bccampus.ca/statspsych/?p=410#h5p-61

For step 4, we have a new t-test formula. We subtract the sample mean of the control group from the sample mean of the experimental group, so that our directionality makes sense when we mark the test score on the comparison distribution and determine whether it falls in the shaded tail.

$$t = \frac{M_1 - M_2}{S_{Difference}}$$

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=410#h5p-66

One thing never changes: In step 5, if the t-test score falls in the shaded tail we reject the null hypothesis.



An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://pressbooks.bccampus.ca/statspsych/?p=410#h5p-62

As we go through the course, we are repeating lots of concepts and procedure enough, that I start to go quickly through those elements. If some aspect of the hypothesis test is still not making sense, that's totally okay, and it's completely normal. But you need to come back to those bits and grapple with them, perhaps by heading back to earlier chapters where those concepts or procedures were first introduced. Do not give up on a concept if it is still fuzzy. By now things should be starting to gel. Are there any aspects that you are still doing by rote rather than through conceptual understanding? I recommend that you persist. It will make sense if you get enough examples and explanations. For most of us it takes quite a bit of repetition and a few

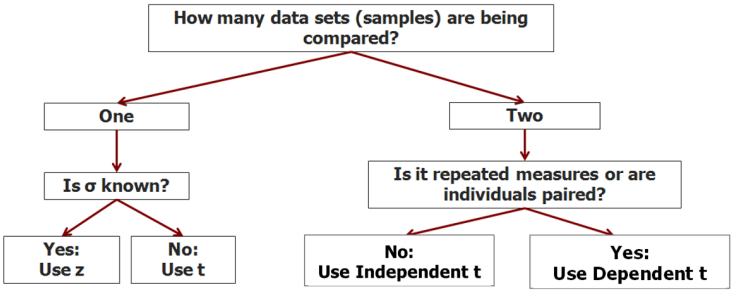
different approaches. Check out another textbook for an alternative look at the same piece.

If we were writing up the hypothesis test outcome from the example illustrated above, we might interpret it this way in the results section:

"We found that people who consumed chocolate had significantly lower mood scores than the control group (p = 0.0145)."

The p-value represents the probability of the test score, or any score that is more extreme than that, occurring under the comparison distribution. To get that, we find the area under the curve beyond the test score, either in one tail or in both tails, depending on the directionality of the test.

We have now completed the decision tree for this section of the course. If we have two samples to compare, and there is no relationship between the individuals in the two samples, we use the **independent t-test**.



An interactive H5P element has been excluded from this version of the text. You can view it online here: https://pressbooks.bccampus.ca/statspsych/?p=410#h5p-57

Chapter Summary

In this chapter, we introduced the use of the **independent means t-test** in the context of hypothesis tests of the difference of two sample means. This test is appropriate for research designs in which two samples are formed through random assignment to groups, for example and experimental group and a control group. Scores from both samples are used to estimate the comparison population distribution, and to contribute to degrees of freedom.

Key terms:

independent means t-test

normal curve assumption

homoscedasticity assumption

8. Analysis of Variance, Planned Contrasts and Posthoc Tests

8a. Analysis of Variance

In this chapter we graduate from teenage statistics to adult statistics! **Analysis of Variance** is a technique that is very widely used in the analysis of data in psychology and many other disciplines. It is a system of analysis that is very flexible, and it is based on a statistical concept called the **general linear model**. Once you learn how to use it, you can adapt it to nearly any situation.

Our tasks for this lesson include grasping the concept of **partitioning variance** into different buckets, like treatment effects vs. error, or between-groups vs. within-groups variance. Next, we will have a look at why **Analysis of Variance** is needed to analyze data from experimental designs with more than 2 groups. In particular we will examine the dangers of inflating the risk of Type I error, or alpha. And finally, we will demystify the **analysis of variance** system by conducting a one-way **ANOVA**. Just to give a little preview, in the following lessons, we will learn how to follow up on **ANOVA** with **planned contrasts** and **post hoc tests**, and then we will progress to a two-way **ANOVA** with factorial analysis.

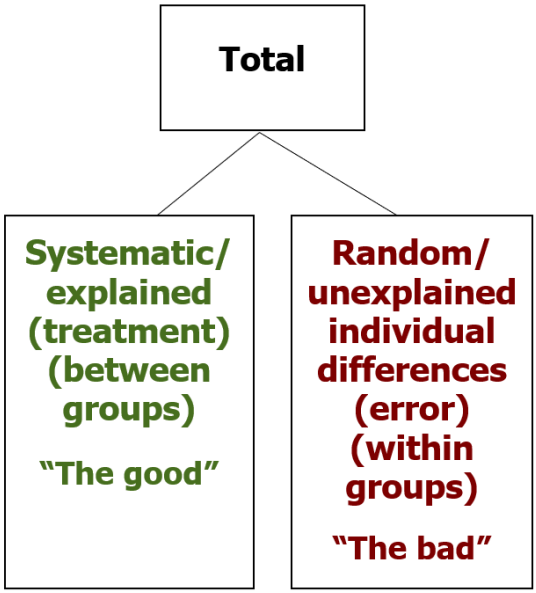
The most important concept to grasp in order to intuitively understand what analysis of variance does, is the **partitioning of variance**. Variance should be a familiar concept by now. Variance is a statistic that summarizes the extent to which the individual scores in a dataset are spread out from the mean. It is calculated by the following steps:

Steps to calculate variance (sample-based estimate for a population)

1. Take the distance ("deviation") of each score from the mean.
2. Next, Square each distance to get rid of the sign (because some deviations will be negative).
3. Add up all the resulting "squared deviations". This number is known as "sum of squares" (SS).
4. Divide the SS by the number of scores minus 1.

This gives us an estimated variance based on a sample, that is appropriate to use in statistical analysis, in which we want to use the differences between sample means to make inferences about the differences between population means.

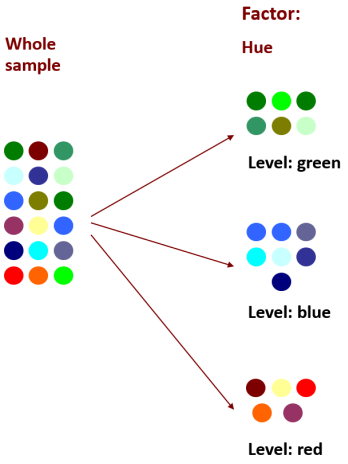
The way the **partitioning of variance** works is this. Differences among scores exist for all sorts of reasons. One of those reasons is the one we are actually interested in. Systematic difference cause by treatments or associated with known characteristics of interest are the differences we are hoping to see in the data. The difference in amount of sleep that can be attributed to the effects of a new drug. The difference in mood specifically caused by chocolate. These are difference between groups, or between samples, that can be explained by the variable of interest.



Sources of variability in data from experimental (or quasi-experimental) research designs

However, there are also difference between scores that are not explained by the variable of interest. These are random, or unsystematic differences. The individual differences among scores within an experimental or control condition count in this category. Error in experimental design or in our measurements also go in this bin. When we want to make an objective decision about data, we need to separate out the systematic, explained differences, which we can label "good variance", from the random, unexplained differences, which we shall label "bad variance". Note that good and bad in this context just means it counts toward ("good") or against ("bad") statistical significance.

Before we get to numeric examples of partitioning variance, maybe a visual example will help. At left we have a whole sample of dots of various colours.



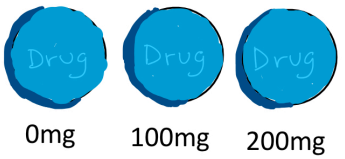
What if we wanted to sort the data by hue, to achieve greater consistency in colour within each group. We can apply the **factor** of hue to the dots, using three levels: green, blue and red. There are still variations of hue within each grouping, but some of the systematic variability has been separated out by grouping into these three levels. Thus we have accounted for (or explained) some proportion of the variance. The more variance we can explain, the more confident we can be in the effect of our **factors**. (In an experimental design, **factors** are independent variables.)

In the next chapter, we will see that applying an additional **factor** can further sort the colours, to account for even more variability among colours. The more variance we can explain, through multiple **factors** and/or multiple **levels**, the better! This is what we will be able to do with two-way **ANOVA** and factorial designs.

Note: a one-way **ANOVA** includes one **factor**, whereas a two-way **ANOVA** includes two **factors**.

Think of data analysis as a game in which the goal is to explain as much of the variability in the scores as possible through known **factors**. It's like imposing order over chaos in order to see patterns more clearly.

Analysis of Variance becomes necessary when we have experimental designs that are more complex than the ones we have used to date. Up until now, we have covered statistical tests that can handle one-sample and two-sample experimental designs. But what if we are comparing three or more samples? For example, what if we have a drug trial in which we are comparing the mean pain levels of patients after receiving placebo, a low dose of the drug, or a high dose of the drug?



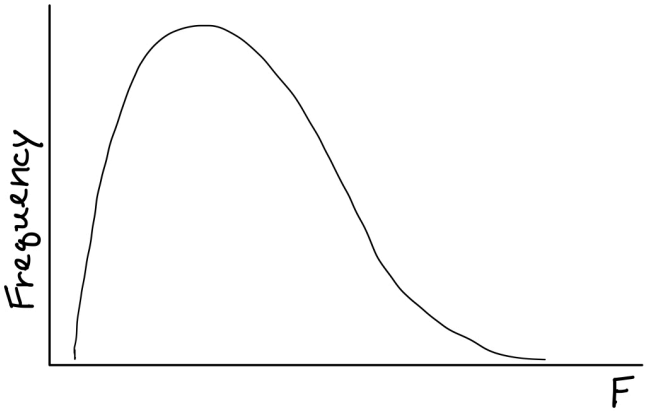
Or what if our memory test using various types of stimuli measures memory for lists of words in black, red, blue or green? **ANOVA** can handle comparisons among 3, 4, or really any number of groups at once.

Heavy	Heavy	Heavy	Heavy
Light	Light	Light	Light
Super	Super	Super	Super
Awful	Awful	Awful	Awful

Let's make sure we have a handle on the jargon that is used in **ANOVA**. First of all, the shortened term **ANOVA** came from making an acronym of sorts from the phrase **Analysis of Variance**. Secondly, the term **factor** is used to designate a nominal variable, or in the case of an experimental design, the independent variable, that designates the groups being compared. If we have a drug trial in which we are comparing the mean pain scores of patients after receiving placebo, a low dose of the drug, or a high dose of the drug, the **factor** would be "drug dose." Finally, the term **levels** refers to the individual conditions or values that make up a factor. In our drug trial example, we have three **levels** of drug dose: placebo, low dose, and high dose.

So how is this **ANOVA** thing different from the t-tests we already learned? Well, in fact, you can think of it as an extension of the t-test to more than 2 groups. If you run an **ANOVA** on just 2 groups, the results are equivalent to the t-test. The only difference is that you get an F-value instead of a t-value. Fun fact – the statistician who invented the t-test published it under a pseudonym "Student". Perhaps he was scared of the angst of the many students who would have to learn to use it. The F-test, however, is named for Fisher. He apparently had no such fear. Maybe he was that confident that students would love learning **ANOVA**. Hopefully he was right… ! Anyway, trust me – if you were to calculate the t-value and the F-value for the exact same two sample, the F-value would be the t-value squared.

There is a nice part of using the F distribution and a not so nice part of it. The F distribution requires two degrees of freedom. Annoying, yes. But the nice part outweighs that annoyance, in my opinion. The F distribution starts from 0 and heads to the right. It has only positive values.



What this means is no more distribution sketches, and no more one-tailed or two-tailed nonsense. So the logistics of the hypothesis test actually get a whole lot simpler.

You might be wondering, okay so the **ANOVA** thing has some advantages, but we do already know the t-test, so could we not just use multiple t-tests to compare each group within the **factor** against each other? The problem is, each comparison includes a risk of a Type I error. The risk of Type I error accumulates with multiple statistical tests on the same data, and that is called the **experimentwise alpha level**. **ANOVA** does one overall, or omnibus, test of treatment effects, to keep our risk of Type I error down. Inflating alpha is dangerous, and any statistical method we can use to keep it under control is a good thing.

The calculation method I will show you differs from more efficient methods you can find on the internet or in many other textbooks. Sorry for that, but the nice thing about the method I will show you here is that it has beautiful symmetry to it and highlights the concept of partitioning of variance. In other words these are conceptual rather than calculational formulas. This is a deliberate choice to help you understand how **ANOVA** works, because if we think about it, you will never need to calculate statistics by hand in the "real world." You will always be able to use a computer instead. However, all the computers in the world cannot help you choose an appropriate statistic for a

particular situation, or to understand/articulate how the selected statistical test works. This is what an introduction to statistics is really about.

There will also be an inherent math double-check opportunity in this method, which I think you will appreciate. In reality, most people use a computer to calculate **ANOVA**. However, I do like to ask you to try calculating things by hand, so you can see how it works. My hope is that you gain a better conceptual understanding of the mechanisms behind these statistical tests by applying them, and seeing how it all fits together like a puzzle, with tangible examples. Given that, I think this calculation system is better than others you can use.

So, how does **ANOVA** work? Essentially it works by calculating different kinds of Sums of squares, which we will continue to abbreviate as SS. As you can see, there are three flavours of SS that can each be calculated using the formulas shown.

SS (sum of squares): sum of squared deviations

$$SS_{T(total)} = SS_{B(between)} + SS_{W(within)}$$

$$\sum (X - M_o)^2 = \sum [N_g(M_g - M_o)^2] + \sum (X - M_g)^2$$

X: individual score, N_g: number of scores in group, M_g: mean of group, M_o: overall mean

The Sum of squares Between-groups (SSB) and Sum of Squares Within-groups (SSW) should add up to the Sum of Squares Total (SST). So here you see the **partitioning of variance** coming in.

There are also three flavours of degrees of freedom, with matching labels. They also should add up.

$$df_{between} = \text{number of groups} - 1$$
$$df_{within} = \text{number of scores} - \text{number of groups}$$
$$df_{total} = \text{number of scores} - 1$$

Notice I used colour coding to help you track the "good" variance in green and the "bad" variance in red.

Once you have the SS and degrees of freedom calculated, you can find the variances.

$$S^2 = SS/df$$

(variance)

(takes into account number of scores in each SS)

The F-test is simple: it is the ratio of explained to unexplained variance, which is represented by the variance between and the variance within. You need more explained than unexplained variance to be able to reject the null.

F = ratio of explained to unexplained variance

$$F = \frac{S^2_B}{S^2_W}$$

How much the ratio needs to be depends on the degrees of freedom. And that's where sample size becomes very important, just as we saw in the t-test.

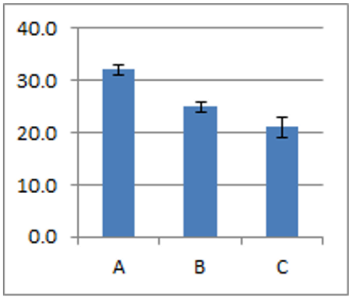
One of the beautiful things about **ANOVA** is the calculation table. This is a way of organizing all the components of the workflow, and also highlighting our two math double-checks. For both SS and degrees of freedom, the Between and Within numbers should add up to the total.

Source	*SS*	*df*	S^2	*F*	*p-value*
Between	$\sum [N_g(M_g - M_o)^2]$	number of groups – 1	$\frac{SS_B}{df_B}$	$\frac{S^2_B}{S^2_W}$	Probability of getting this *F* or greater given these *df*s.
Within	$\sum (X - M_g)^2$	number of scores – number of groups	$\frac{SS_W}{df_W}$		
Total	$\sum (X - M_o)^2$	number of scores – 1	**X: individual score** **N_g: number of scores in group** **M_g: mean of group** **M_o: overall mean**		

This table is a good reference for you to keep to hand, as a reminder of each formula and how the ANOVA puzzle fits together.

Note what each symbol in these formulas means, by referring to the symbols key in the lower right of the table. The one element that tends to be confusing is N_G. This symbol refers to the number of scores in the group – not to the number of groups in the study. This is important to interpret correctly. If you ever find that your SSB and SSW do not add up to your SST, like really not even close, then that is the first thing to double check. Did you use the number of scores in the group when calculating SSB?

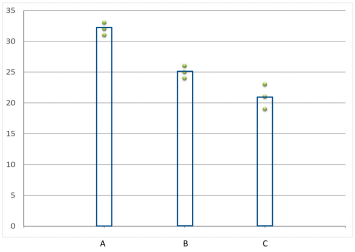
Now that our research designs are getting more complex, our statistical findings will need a little more descriptive statistics and graphical portrayal in order to easily interpret what those hypothesis test results really tell us. I would encourage you to always graph the means and standard deviations of your data before conducting your inferential statistics, so that you get a sense for significance before you begin. Do not go blind into that statistical routine… remember, numbers are never as informative as a picture of the data.



I recommend a bar graph displaying group means, and adding error bars as tall as the standard deviation (or standard error) on top of the mean. You can also show the bars going one standard deviation downward as well, to get the full range of the typical scores in the group.

If the error bars eclipse the difference in group means, that is a bad sign if your goal is to report a significant difference among means. This visual allows you to preview the signal-to-noise ratio in your data, or your between-to-within variance ratio.

Another really great visual is the group scatter plot, shown here. It is not really a standard way to view datasets, but I think it should be.



Step 1 of hypothesis testing for an ANOVA truly becomes a formality. The hypotheses are always the same. Define a population for each group. Set the research hypothesis to be a general statement of difference among population means. Set the null to be a statement of equality among population means. There is no directionality with the F distribution, so we do not need to worry about the predicted direction of differences.

Using our drug-dose example with three levels, the populations and hypotheses would look something like this:

Population 1: People who receive low dose of drug

Population 2: People who receive high dose of drug

Population 3: People who do not receive drug

Research Hypothesis: There exists at least one

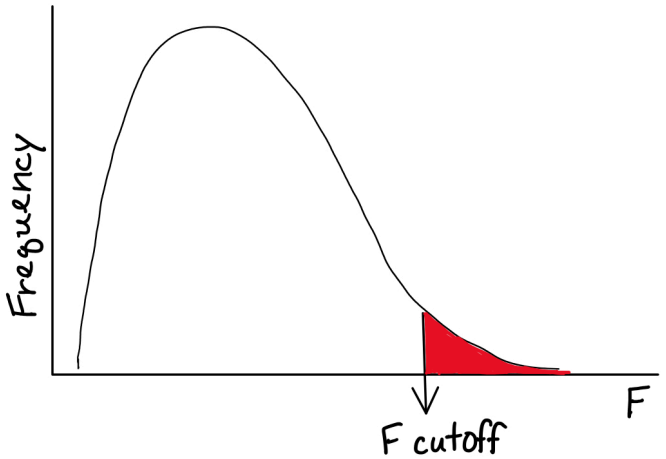
difference among the population means.

Null Hypothesis: $\mu_1 = \mu_2 = \mu_3$ All population means are equal.

Now we can move on to step 2. The F distribution has two degrees of freedom. $df_{Between}$, df_{Within}

We no longer have to worry about the mean or standard deviation of the comparison distribution, we just need to find the degrees of freedom between and within. The "good" variance is the differences between groups, and so the degrees of freedom between is number of groups – 1. The within-groups variance, the "bad" variance, is the individual differences among the scores within each group. The degrees of freedom within, then, is the total number of scores in all groups, minus the number of groups.

For step 3, we can find the cutoff score in the F-tables if we know the significance level, degrees of freedom between and degrees of freedom within.



Step 4 is where things take some getting used to. Here we use this new system of formulas. Start with Sum of squares calculations: Between, Within, and Total, and double check that both they and the degrees of freedom add up.

$$SS_B = \sum [N_g(M_g - M_o)^2]$$

$$SS_W = \sum (X - M_g)^2$$

$$SS_T = \sum (X - M_o)^2$$

Then move across the table, finding the good and the bad variance...

$$S_B^2 = \frac{SS_W}{df_B}$$

$$S_W^2 = \frac{SS_W}{df_W}$$

... and finally getting their ratio for the F-test result.

$$F = \frac{S_B^2}{S_W^2}$$

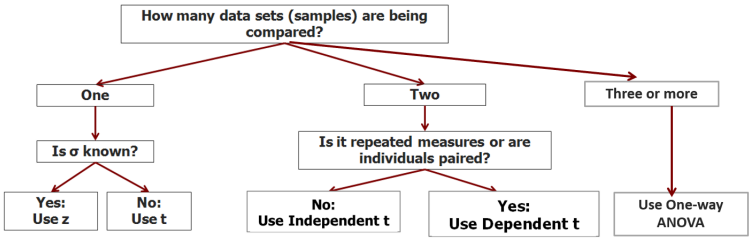
To make our decision in Step 5, we examine the calculated F value (from Step 4) and determine whether it exceeds the cutoff F score (from Step 3). If so, we reject the null hypothesis.

Here is an example of how to express the results – note the phrase "significant difference among the means." If we do not reject the null, we can switch the statement of results to "no significant difference." The test statistic and p-values are expressed here in common formats.

"There is a significant difference among the mean digit memory scores after listening to the three types of music ($f_{2,6} = 27.00$, $p < 0.05$)."

We can continue building a decision tree to help you decide which statistical test to use when you look at a research question. What are

the circumstances in which you would need to use a one-way **ANOVA** test?

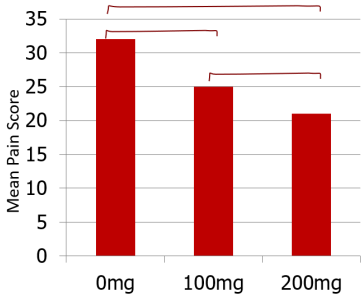


8b. Planned Contrasts and Posthoc Tests

In the second part of this chapter we will have a look at follow-up tests we can conduct after an **ANOVA** hypothesis test, to investigate the findings in greater detail.

Planned contrasts and post-hoc tests are commonly performed following **Analysis of Variance**. This is necessary in many instances, because **ANOVA** compares all individual mean differences simultaneously, in one test (referred to as an omnibus test). If we run an **ANOVA** hypothesis test, and the F-test comes out significant, this indicates that at least one among the mean differences is statistically significant. However, when the **factor** has more than two **levels**, it does not indicate which means differ significantly from each other.

In this example, a significant F-test result from a one-way **ANOVA** with the three drug dose conditions does not tell us where the significant difference lies. Is it between 0 and 100 mg? Or between 100 and 200 mg? Or is it only the biggest difference that is significant – 0 vs. 200 mg?



Planned contrasts and **post hoc tests** are additional tests to determine exactly which mean differences are significant, and which are not. Why is that we cannot just do 3 independent means t-tests here? Each time we conduct a t-test we have a certain risk of a Type I error. If we do 3, we have triple the risk. So first we test for omnibus significance using the overall **ANOVA** as detailed in the first part of this chapter. Then, if a statistically significant difference exists among the means, we do the pairwise comparisons with an adjustment to be more conservative. These follow-up tests are designed specifically to avoid inflating risk of Type I error.

Now, this is very important. We are *only* allowed to conduct these tests *if the F-test result was significant*. This procedural rule also helps protect us from the statistical sin of p-hacking, which is selectively hunting for and reporting significant results in a way that is biased and subjective.

Planned contrasts are used when researchers know in advance which groups they expect to differ. For example, suppose from our worksheet example, we expect the pop group to differ from the classical group on our measure of working memory. We can then conduct a single comparison between these means without worrying about Type I error. Because we hypothesized this difference before we saw the data, perhaps based on prior research studies or a strong intuitive hunch, and because there is only one comparison to be analyzed, we need not be concerned about inflated **experimentwise alpha**. If multiple comparisons are planned, then we will need to adjust the significance level.

Let us take a look at how to conduct a single **planned contrast**. The process is quite simple, as it is just a modified **ANOVA** analysis. First we calculate SSB with just those two groups involved in the planned contrast. We figure out the degrees of freedom between using just

the two groups. Then, we calculate the variance between using the new SSB and degrees of freedom, and we calculate an F-test for the comparison using the new variance between and the original overall variance within. To find out if the F-test result is significant, we can use the new degrees of freedom but the original significance level for the cutoff. (Because there is just one pairwise comparison, we can use original significance level.)

Steps to calculate a planned contrast

1. Calculate $SS_{Between}$ with just those two groups.
2. Find the $df_{Between}$ using just the two groups.
3. Calculate $S^2_{Between}$ using the new $SS_{Between}$ and the new $df_{Between}$.
4. Calculate F using the new $S^2_{Between}$ and the overall S^2_{Within}.

If we were to perform multiple planned contrasts, things change a little. Suppose we had hypothesized in this experiment that each group would differ from the others? The **Bonferroni correction** involves adjusting the significance level to protect from the inflation of risk of Type I error. The procedure for each comparison is the same as for a single planned contrast. The difference is that the cutoff score to determine statistical significance will use a more conservative significance level. When we do multiple pairwise comparisons, the **Bonferroni correction** is to use the original significance level divided by number of planned contrasts. The adjusted significance level is not likely to be in our F-tables, so to find the cutoff for such tests, we would need to use an online calculator in reverse (that is, we enter the p-value and degrees of freedom, and look up the value on the F-distribution corresponding to that area in the tail).

What about **post hoc tests** tests? As the name suggests, these tests come into the picture when we are doing pairwise comparisons

(usually all possible combinations) after the fact to find out where the significant differences were. These are tests that do not require that we had an *a priori* hypothesis ahead of data collection. Essentially, these are an allowable and acceptable form of data-snooping. This is where we must be cautious about doing so many tests – we could end up with huge risk of Type I error. If we use the **Bonferroni correction** that we saw for multiple planned comparisons on more than 3 tests, the significance level would be vanishingly small. This would make it nearly impossible to detect significant differences. For this reason, slightly more forgiving tests like **Scheffe's correction**, Dunn's or Tukey's **post-hoc tests** are more popular. There are many different post-hoc tests out there, and the choice of which one researchers use is often a matter of convention in their area of research.

Now we shall take a look at how to conduct **post hoc tests** using **Scheffé's correction**. In this example, we will test all pairwise comparisons. The **Scheffé** technique involves adjusting the F-test result, rather than adjusting the significance level. The way it works is the same as the **planned contrast** procedure, except for the very end. Before we compare the F-test result to the cutoff score, we divide the F value by the overall degrees of freedom between, or the number of groups minus one. Thus, we keep the significance level at the original level, but divide the calculated F by overall degrees of freedom between from the overall **ANOVA**.

Steps to calculate post-hoc tests with Scheffé's correction

For each pairwise comparison:

1. Calculate $SS_{Between}$ with just those two groups.
2. Find the $df_{Between}$ using just the two groups.
3. Calculate $S^2_{Between}$ using the new $SS_{Between}$ and the new $df_{Between}$.

4. Calculate F using the new $S^2_{Between}$ and the overall S^2_{Within}.
5. Divide F by overall $df_{Between}$.

Chapter Summary

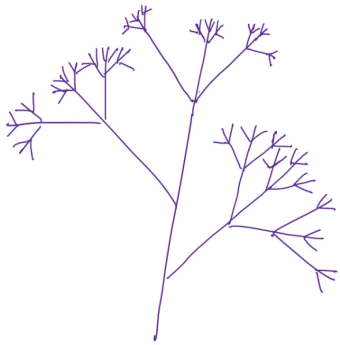
In this chapter we introduced the concepts underlying **Analysis of Variance** and examined how to conduct a hypothesis test using this technique. We also saw how to follow up on a statistically significant F-test result in an **ANOVA** with more than two **levels** in a **factor**, in order to determine which levels were significantly different from each other.

Key terms:

Analysis of Variance	**post hoc tests**	**Bonferroni correction**
general linear model	**factor**	**Scheffé correction**
partitioning of variance	**levels**	
planned contrasts	**experimentwise alpha level**	

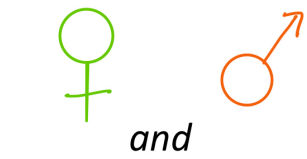
9. Factorial ANOVA and Interaction Effects

9a. Factorial Analysis

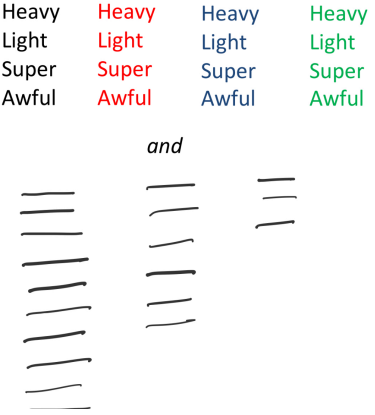


In factorial analysis, just like the fractals we see in nature, we can add multiple branchings to every experimental group, thus exploring combinations of factors and their contribution to the meaningful patterns we see in the data.

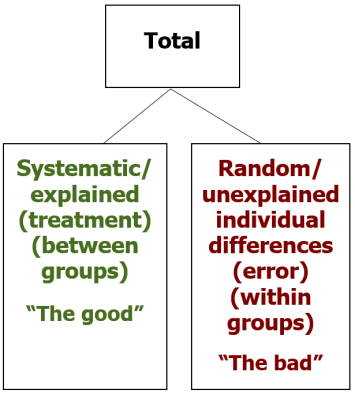
In this chapter we will tackle two-way Analysis of Variance and explore conceptually how factorial analysis works. To understand when you need two-way ANOVA and how to set up the analyses, you need to understand the matching research design terminology. We will also need to define and interpret **main effects** and **interaction** effects, both of which can be analyzed in a factorial research design. Later we will approach the detection and interpretation of **interaction** effects, specifically, which will really help you see the extraordinary complexity of information factorial analyses can offer.



Factorial analyses such as a two-way ANOVA are required when we analyze data from a more complex experimental design than we have seen up until now. Specifically, when an experiment (or quasi-experiment) includes two or more independent variables (or **participant variables**), we need factorial analysis. Examples of designs requiring two-way ANOVA (in which there are two factors) might include the following: a drug trial with three doses as well as the sex of the participant, or a memory test using four different colours of stimuli and also three different lengths of word lists.

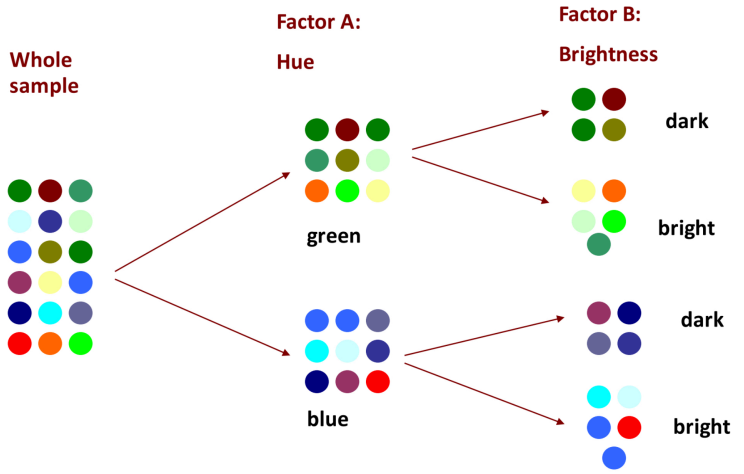


As we saw in the chapter on Analysis of Variance, the total variability among scores in a dataset can be separated out, or partitioned, into two buckets. The first bucket, often called between-groups variance or treatment effect, refers to the systematic differences caused by treatments or associated



with known characteristics. These are the differences among scores we are hoping to see — the explained differences — and thus I casually refer to this as the "good" bucket of variance and colour code it in green. The other bucket, often called within-groups variance or error, refers to the random, unsystematic differences that cannot be explained by the research design. These are the unexplained individual differences that represent the noise in the data, obscuring the signal or pattern we are looking for, and thus I casually refer to it as the "bad" bucket of variance and colour code it in red.

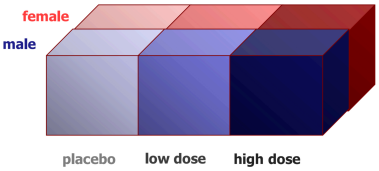
We can revisit our visual example from before, in which the goal is to separate colour swatches according to some factor, such that the colours within each grouping (or level) is more uniform. If we first sort the colours according to the factor of hue, let's say into green or blue hues, then we explain some of the overall variability. But if we add a second factor, brightness, then we can explain even more of the differences among the colour swatches, making each grouping a little more uniform.



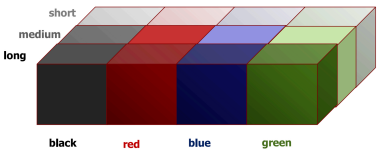
Clearly there is still some work to be done, and if in factor A we could have included a third level of "red", the uniformity would have been much improved. And with factorial analysis, there is technically no limit to the number of factors or the number of levels we can employ to explain away the variability in the data. The more variance we can explain, through multiple factors and/or multiple levels, the better! This is what we will be able to do with two-way ANOVA and factorial designs.

You will recall the jargon of ANOVA, including factors and levels. To grasp factorial research designs, it becomes even more important to develop comfort with these concepts, so that you can identify and describe the design and thus the requisite analysis setup. Let us suppose that we have a research study that measures the effect of a placebo, a low dose and a high dose of the drug, and also takes into account whether the participants were male or female. The first factor could be succinctly identified as "drug dose", and the second factor as "sex". In another example, perhaps we show participants words in black, red, blue or green, and we also take into account whether the word list presented is long, medium, or short. What would you call each of those two factors?

What if, in a drug study, you notice that men seem to react differently than women? If you have that information (male/female), you can use it in your ANOVA and see if you can put more variance in your "good" bucket.



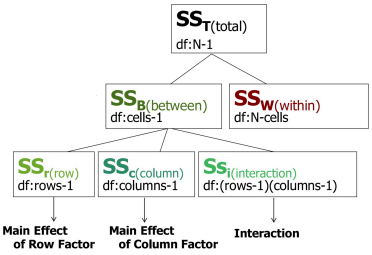
In the design illustrated here, we see that it is a 3 x 2 ANOVA. There are three levels in the first factor (drug dose), and there are two levels in the second factor (sex). This notation, that identifies the number of levels in each factor with a multiplier between, helps us see clearly how many samples are needed to realize the research design. In this example, we would need six samples in total, each of which would need to have a good enough sample size to allow for the central limit theorem to justify the normality assumption (N=30+). That is a lot of participants! However, we could learn much more by including both factors, if indeed the sex of the participant is associated with a different response to the drug.



We can see an example of a 4×3 two-way ANOVA here, with our example of word colour and length of list. Altogether, this design would require 12 samples.

And just for the sake of showing you the potential of factorial analyses, you could also impose a third factor on the design: the age of the participants. In this case, you have a 4x3x2 design, requiring 12 samples. At 30 participants each, that would be 30×12=360 people! You can appreciate how each factor exponentially increases the practical demands (costs) of the research study. For this reason, a cost-benefit analysis must be carefully applied in factorial research design, such that the minimum complexity is used to answer the key research questions sufficiently.

In a two-way ANOVA, just as in a one-way ANOVA, we calculate various flavours of Sums of Squares (SS). The SS total is broken down into SS between and SS within. However, with a two-way ANOVA, the SS between must be further broken down, because there are now two different factors that can have a **main effect** (i.e., can explain some of the total variance). Also, with more than one factor, there can be an **interaction** between the two that itself uniquely accounts for some of the variance. So now, we can SS row (the first factor), SS column (the second factor) and SS **interaction**. For each SS, you can also see the matching degrees of freedom.



Here is the full ANOVA table expanded to accommodate the three subtypes of between-groups variability.

Source	*SS*	*df*	S^2	*F*
(Between) Row	$\sum [N_{row}(M_{row} - M_o)^2]$	rows – 1	$\frac{SS_r}{df_r}$	$\frac{S^2_r}{S^2_W}$
(Between) Column	$\sum [N_{col}(M_{col} - M_o)^2]$	columns – 1	$\frac{SS_c}{df_c}$	$\frac{S^2_c}{S^2_W}$
(Between) Inter-action	$\sum [N_{cell}(M_{cell} - M_o)^2] - SS_{row} - SS_{col}$	(rows – 1) (columns – 1)	$\frac{SS_i}{df_i}$	$\frac{S^2_i}{S^2_W}$
Within	$\sum (X - M_{cell})^2$	N – cells	$\frac{SS_W}{df_W}$	X: individual score N: number of scores M_o: overall mean
Total	$\sum (X - M_o)^2$	N – 1		

Note that all of the Sums of Squares and degrees of freedom still should add up to the total. As you can see, there will now be three F-test results from this one omnibus analysis, one for each of the between-groups terms. Each can be compared to the appropriate degrees of freedom to determine the statistical significance of the degree to which that factor (or interaction) accounts for variance in the dependent variable that was measured in the study.

To help you interpret the formulas as they reference row means, column means, and cell means, I have added a diagram here to help you see how to locate these numbers in a 2×2 two-way ANOVA scenario.

Source	SS	df	S^2	F
(Between) Row	$\sum [N_{row}(M_{row} - M_o)^2]$	rows – 1	$\frac{SS_r}{df_r}$	$\frac{S^2_r}{S^2_W}$
(Between) Column	$\sum [N_{col}(M_{col} - M_o)^2]$	columns – 1	$\frac{SS_c}{df_c}$	$\frac{S^2_c}{S^2_W}$
(Between) Inter-action	$\sum [N_{cell}(M_{cell} - M_o)^2] - SS_{row} - SS_{col}$	(rows – 1) (columns – 1)	$\frac{SS_i}{df_i}$	$\frac{S^2_i}{S^2_W}$
Within	MEANS:			X: individual score N: number of scores M_o: overall mean
Total				

	Col A	Col B	Row Means
Row A	Cell	Cell	
Row B	Cell	Cell	
Column Means			

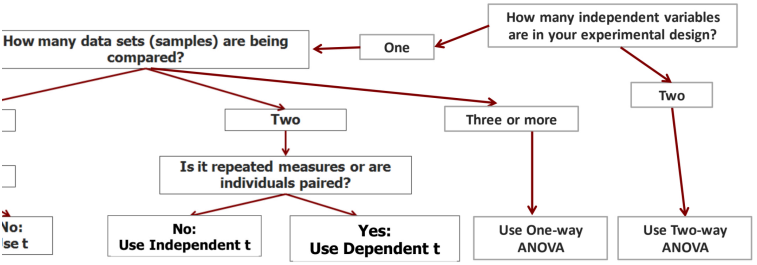
The row and column means, the averages of cell means going across or down this matrix, are often referred to as marginal means (because they are noted at the margins of the data matrix).

When it comes to hypothesis testing, a two-way ANOVA can best be thought of as three hypothesis tests in one. For each factor, and also for the **interaction** of the two, you need to identify populations and hypotheses, cutoffs, calculate the SS between, degrees of freedom, variance between, and F-test results. All three will share the same error terms, the SS, degrees of freedom, and variance within groups. As you can imagine, the complexity of calculating such an analysis could be daunting, but a systematic, organized approach and the use of the ANOVA table keeps it well under control.

As with one-way ANOVA, if any factor has more than two levels, you may need to calculate pairwise contrasts for that factor to determine where exactly a significant difference among group means lies. Even with a 2×2 ANOVA, the **interaction** effect has four possible pairwise comparisons to investigate, and that would require a planned contrast or post-hoc test. The same rules apply to such analyses as before: they may only be conducted if there is a significant overall ANOVA result, and the experimentwise risk of Type I error must be controlled.

We can continue building our statistical decision tree to help us decide which test to use when we examine a research question/design.

If we have two independent variables (factors) in the experimental design, then we need to use a two-way ANOVA to analyze the data.



Before we move on to detecting and interpreting **main effects** and **interactions**, I would like to bring in two cautions about factorial designs. Many researchers new to the trade are keen to include as many factors as possible in their research design, and to include lots of levels just in case it is informative. This is an understandable impulse, given how much effort and expense can go into designing and conducting a research study. We want to gather as much information as possible from that effort! However, as we saw before, the more factors we add in, the more participants we need to ensure a decent sample size in each cell of our data matrix. There is another important element to consider, as well. For each factor we add in, we add **interaction** terms. If we were ambitious enough to include three factors in our research design, we would have the potential for **interaction** effects among each pair of the factors, but we would also potentially see a three-way **interaction** effect.

Interactions for a three-way ANOVA

In a three-way ANOVA involving factors A, B, and C, one must analyze the following interactions:

- A x B
- B x C
- A x C
- A x B x C

The interpretation of all these **interactions** becomes very challenging. For this reason, solid advice to researchers is to limit ourselves to two factors for any given analysis, unless there is a very strong hypothesis regarding a three-way **interaction**.

9b. Interaction Effects

In this part of the chapter, we will dig into **interaction** effects and how to detect and interpret them alongside **main effects** in factorial analyses. We will see that **main effects** can be detected using group means tables, and **interactions** can be detected using the tools of bar graphs and **interaction** plots.

We will also look at how to interpret three major scenarios: when we have significant **main effects** but no significant **interaction**; when we have a significant **interaction**, but no **main effects** and when we have both **interactions** and **main effects** that turn out significant.

A **main effect** means that one of the factors explains a significant amount of variability in the data when taken on its own, independent of the other factor. You can tell (roughly) whether a **main effect** is likely to exist by looking at the data tables. Specifically, you want to look at the marginal means, or what we called the row and column means in the context of a two-way ANOVA above.

Let us look at the first example.

Data Example 1

	Male	Female	**Row means**
Low dose of drug	20	10	15
High dose of drug	10	20	15
Column means	15	15	

Going across the data table, you can see the mean pain score measured in people who received a low dose of a drug, and those who received a high dose. The marginal means are 15 vs. 15. This indicates there is clearly no difference between the two, so there is no **main effect** of drug dose. Now look top to bottom to find the comparison between male and female participants on average. 15 vs. 15 again, so no **main effect** of education level.

Now look at the second example.

Data Example 2

	Male	Female	**Row means**
Low dose of drug	40	20	30
High dose of drug	30	10	20
Column means	35	15	

Going across, we can see a difference in the row means. People who receive the low dose have less pain that those who receive the high dose: this could be a significant **main effect**. Going down, we can see a different in the column means as well. Males report more pain than females. Another likely **main effect**. So in this example there is an apparent **main effect** of each factor, independent of the other factor.

Now, detecting **interaction** effects in a data table like this is trickier. But if you can see a clear X-pattern in the group means table (the four cell means), such that similar numbers connect in an "X", then that is a sign that there is probably an interaction. If not, there may not be.

In the first example, it is clear that there is an X pattern if you connect similar numbers (20 with 20 and 10 with 10). Probably an interaction.

Data Example 1

	Male	Female	**Row means**
Low dose of drug	20	10	15
High dose of drug	10	20	15
Column means	15	15	

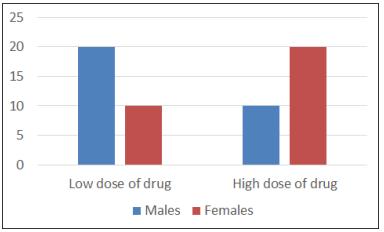
In the second example, it is not so clear. Ask yourself: if you take one row at a time, is there a different pattern for each or a similar one?

Data Example 2

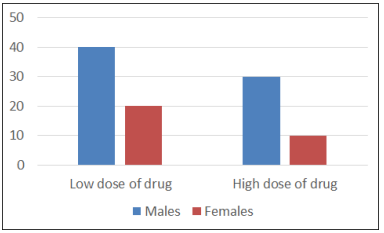
	Male	Female	**Row means**
Low dose of drug	40	20	30
High dose of drug	30	10	20
Column means	35	15	

People with a low dose have lower pain scores if they are female. A similar pattern exists for the high dose as well. This similarity in pattern suggests there is no interaction. You can do the same test with the columns and reach the same conclusion.

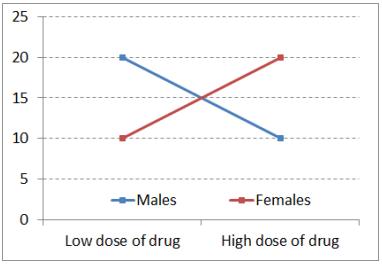
It is far easier to tell at a glance whether an **interaction** exists if you graph the data. In a bar graph, look for a U- or inverted-U-shaped pattern across side-by-side bar graphs as an indication of an **interaction**.

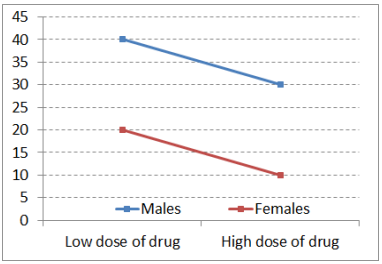


In the top graph, there is clearly an **interaction**: look at the U shape the graphs form.



In the bottom graph, there is no such U shape. When you look at each set of bars in turn, the pattern displayed is similar – just a little higher overall for the older people. Clearly, there is no hint of an **interaction**.





Interaction plots make it even easier to see if an **interaction** exists in a dataset. If you were to connect the tops of like-coloured bars of the graphs on the previous bar graphs, you would get line plots like those shown here.

If the two resulting lines are non-parallel, then there is an **interaction**. On the other hand, if the lines are parallel or close to parallel, there is no **interaction**.

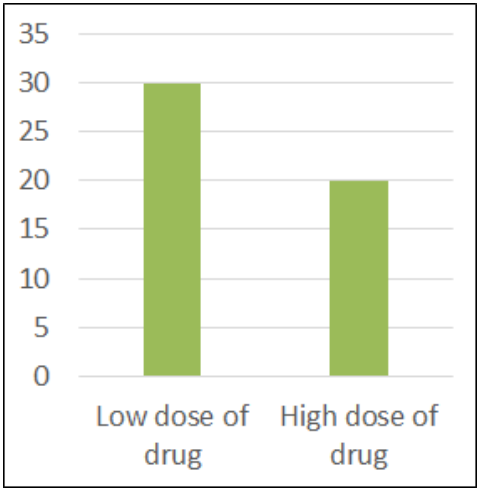
Now you have seen the same example datasets displayed in three different ways, each making it easy to see particular aspects of the patterns made by the data.

More challenging than the detection of **main effects** and **interactions** is determining their meaning. Learning to interpret **main effects** and **interactions** is the most challenging aspect of factorial analyses, at least for most of us. Now we will take a look systematically at the three basic possible scenarios.

The first possible scenario is that **main effects** exist with no **interaction**. This can be interpreted as the following: each factor independently influenced the dependent variable (or at least accounted for a sizeable share of variance).

Data Example 2

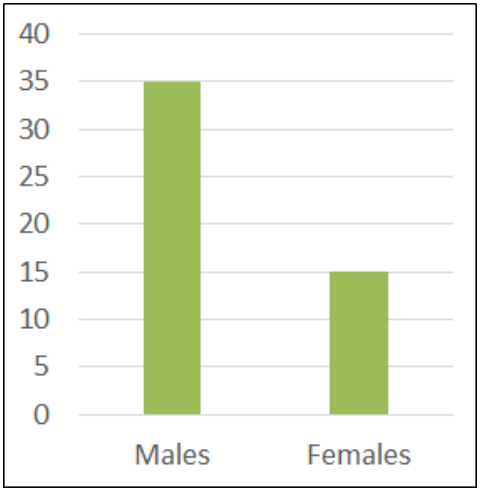
		Row means
Low dose of drug		**30**
High dose of drug		**20**
Column means		



In other words, if you were to look at one factor at a time, ignoring the other factor entirely, you would see that there was a difference in the dependent variable you were measuring, between the levels of that factor.

Data Example 2

	Male	Female	Row means
Column means	**35**	**15**	

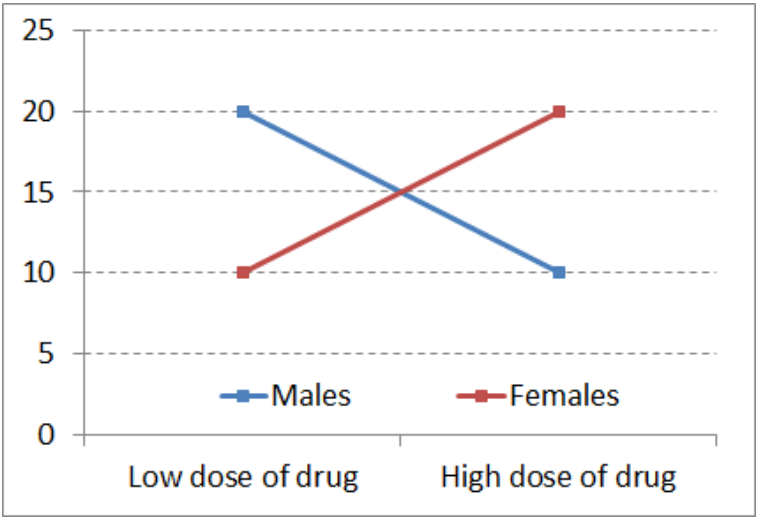


The second possible scenario is that an **interaction** exists without **main effects**. We can interpret this as follows: each factor did not, in and of itself, influence the dependent variable.

Data Example 1

	Male	Female	**Row means**
Low dose of drug	20	10	15
High dose of drug	10	20	15
Column means	15	15	

Here you can see that neither dose nor sex marginal means differ – no **main effects**. But the non-parallel lines in the graph of cell means indicate an **interaction**.



The best way to interpret an **interaction** is to start describing the patterns for each level of one of the factors. First we will examine the low dose group. They have lower pain scores only if they are female. Now look at the high dose group: they have a lower pain scores only if they are male – the opposite pattern. This means that the effect of the drug on pain depends on (or interacts with) sex.

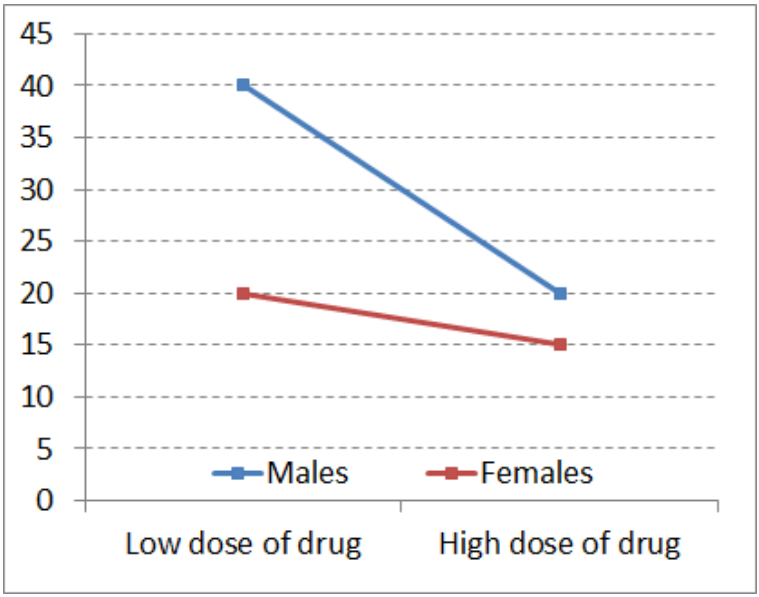
The third possible basic scenario in a dataset is that **main effects** *and* **interactions** exist. This means each factor independently accounted for variability in the dependent variable in its own right. But also, they interacted synergistically to explain variance in the dependent variable. Together, the two factors do something else beyond their separate, independent **main effects**.

In this example, at both low dose and high dose of the drug, pain levels are higher for males.

Data Example 3

	Male	Female	**Row means**
Low dose of drug	40	20	30
High dose of drug	20	15	17.5
Column means	30	17.5	

For both sexes, the higher dose is more effective at reducing pain than the lower dose. But there is also an **interaction**, in that the difference between drug dose is much more accentuated in males. Just look at the difference in the slope of the lines in the **interaction** plot.



The lines are certainly non-parallel. So drug dose and sex matter, each in their own right, but also in their particular combination.

You can probably imagine how such a pattern could arise. Perhaps males are more sensitive to pain, and thus require a high dose to achieve relief. Or perhaps the higher body mass in males means a higher dose of drug is required to be effective. For females, both doses are similar in their efficacy.

Chapter Summary

In this chapter we introduced the concept of factorial analysis and took a look at how to conduct a two-way ANOVA. We further examined ways to detect and interpret **main effects** and **interactions**.

Key terms:

main effects **interactions** **participant variables**

10. Correlation and Regression

10a. Correlation

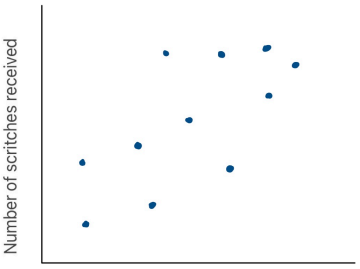
This chapter marks a big shift from the inferential techniques we have learned to date. Here we will be looking at relationships between two numeric variables, rather than analyzing the differences between the means of two or more experimental groups.

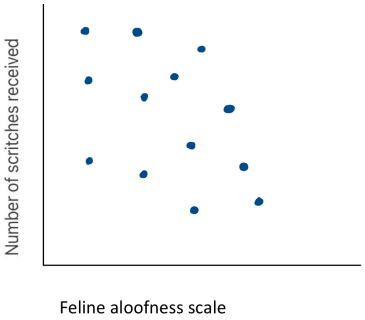
Correlation is used to test the direction and strength of the relationships between two numerical variables. We will see how scatterplots can be used to plot variable X against variable Y to detect linear relationships. The slop of the linear relationship can be positive or negative, which reveals systematic patterns in how the two variables co-relate. We will also look at the theory of **correlational** analysis, including some cautions around interpreting the results of **correlational** analyses. Thanks to the third variable problem, **correlation** does NOT equal causation, a mantra that should be familiar from your introductory psychology courses. And finally, we will try calculating correlation by partitioning **covariance**, and put it all into practice in a hypothesis test. Later in the chapter, we will build in **regression**, which allows us to predict the future from the past.

Just like a bar graph is helpful to examine visually the differences among means, a scatterplot allows us to visualize the pattern that represents the relationship between two numeric variables, X vs. Y.

If the trend line that best indicates the linear pattern in the scatter plot has an upward slope, we consider that a positive directionality.

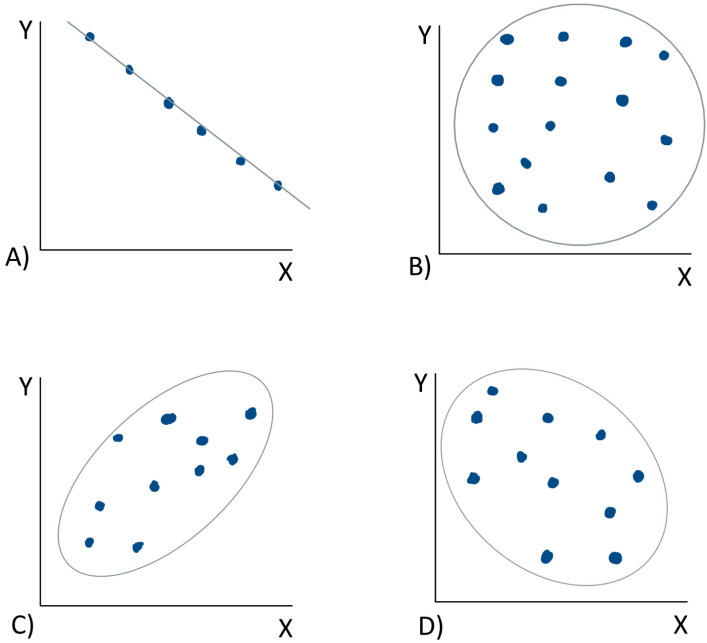
To find out if there appears to be a positive correlation, you can ask yourself "are those that score high on one variable likely to score high on the other?" Here we see an example: what is the relationship between feline friendliness and number of scritches received? As you can see, when cat friendliness is high, the cuddles received is also high. There is a clear positive trend line. This make sense – people may be more likely to offer cuddles to a cat that solicits them.





A downward slope indicates a negative directionality. To find out if there appears to be a negative correlation, you can ask yourself "are those that score high on one variable likely to score low on the other?" Here you can see an example: what is the relationship between feline aloofness and the number of scritches received? There is a clear negative trend line. This makes logical sense, because people may be less likely to offer cuddles to a cat that keeps to itself.

When we look at a scatter plot, we want to ask ourselves two questions: one about the apparent strength of the relationship between the variables, and the other about the direction of the relationship. Let us take a look at a few examples.



In graph A), if we ask "are variables X and Y strongly or weakly related?" We would say strongly related. This is because the points on the scatter plot are in a perfect line. There is no distance between the points and the trend line. It is a perfect **correlation**. If we ask "is the trend line positive or negative in slope?" We would say that it is negative in slope. As scores on variable X increase, scores on variable Y do the opposite – they decrease. We might expect such a relationship if we plotted speed against time. The faster something is, the less time it takes. In the next example, graph B), if we ask "are variables X and Y strongly or weakly related?" We would say weakly related. There is no clear linear trend that can be visually discerned – it just looks like a random scatter of dots. With no trend line, the question about directionality is irrelevant. This **correlation** is close to zero, so it is neither positive nor negative as a directional relationship. If we look at example C), the strength is not quite as perfect as in the first example, but the dots would not be very distant from a trend line through them, so this would be a fairly strong **correlation**. As scores on variable X go up, so

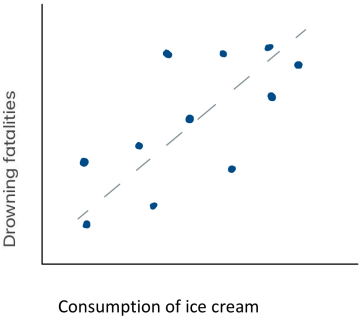
do those on variable Y, making this a positive **correlation**. Now it is your turn. In example D), would this be a strong relationship or weak? Or somewhere in between? And do you see a positive or a negative slope to a trendline that runs through the cloud?

Correlational analysis seeks to answer the question "how closely related are two variables." This is a very useful analytical approach when we have two numeric variables and we wish to analyze the patterns in how they co-vary. However, **correlational** analyses have limitations that it is vital to be aware of.

First, the **correlational** method we will cover in this course is only capable of detecting linear relationships. Patterns that have a curve to them will not be captured by the **correlation** formula we will use.

Secondly, **correlation** does *not* equal causation. **Correlational** research designs do not allow for causal interpretations, because the third variable problem renders **correlational** analyses vulnerable to spurious results. When we measure two variables at the same time and plot them against each other, what we can do is *describe* their relationship. We can even test whether the strength of their relationship is significantly different from zero. However, we cannot determine whether X *causes* Y.

For example, if we measured the consumption of ice cream as well as drowning deaths on a sample of days throughout the year, we might determine that there is a strong positive relationship between the two variables. Consumption of ice cream and drowning deaths are apparently closely related phenomena. But does consumption of ice cream cause drowning deaths? That seems a little far fetched. Could there be another explanation for the pattern? Is there a third variable that could in fact explain the trends in each of the two variables measured here? What might cause people to consume more ice cream as well as put themselves at greater risk for drowning? Warm weather perhaps? If we were to plot temperate against ice cream and drowning deaths, would we see positive correlations with each? Very likely. With this third variable connecting the two, it would be a logical fallacy to interpret the apparent **correlation** shown here as meaningful.



But then again, is it possible that consuming ice cream could be a risk factor for drowning? Did an elder ever tell you that you should not eat right before swimming, because you might cramp up and drown? Maybe there is some truth to that.

So how could we find out whether there is a true causal relationship between two variables? In order to make cause-effect conclusions, we must use an experimental design. Two major features of experimental research designs eliminate the logical fallacies associated with **correlations**. First, an experiment makes use of random assignment of participants to conditions, because that controls for extraneous variables like the third variable of temperature in this example. And secondly, an experiment manipulates the independent variable, to establish a cause, and then measure effects.

Requirements for cause-effect conclusions

A true experiment requires the following elements in order to control for extraneous variables and establish cause-effect directionality:

- random assignment of participants to conditions (or randomization of order of conditions in repeated measures designs)
- manipulation of the independent variable

In our ice cream and drowning study here, how could we make it into an experiment, to allow for causal conclusions? First we would have to assign our participants randomly into the experimental and control groups. There must be no systematic bias in who is given ice cream and who is not. Second, we would have to manipulate independent variable – we would have to have those participants in the experimental group eat ice cream. Then we would put all participants in water, at the same temperature, and see how many of them drown. We would calculate

the average number of drowning events in the ice-cream-eating vs. the control group, and run a t-test or ANOVA to find out if they are significantly different from each other.

Of course, you might be thinking, "would this be ethical?" At least I hope you are thinking that. Of course not! It would not make sense to allow people to drown, just to answer this empirical question. In fact, that is exactly why **correlation** exists.

"Herp Derp :D" by O hai :3 is licensed under CC BY 2.0

Often, practical or ethical limitations make an experiment prohibitively difficult or impossible. If we are limited to **correlational** techniques in a particular research study, then we simply cannot draw cause-effect inferences.

So, a major take-home point of this lesson is... don't be like this guy.

Okay, so how do we go about calculating **correlation**? Well, similar to ANOVA, we can think of the process conceptually as the partitioning of variance. But this time, what counts as good variance is **covariance**. This is the systematic variance that both variables X and Y have in common. Because it is variance that is explained by the co-relation of the two variables, we will put **covariance** in the "good" bucket. The random variance that is unexplained by the relationship between X and Y, the distance between the dots and the trend line, that is the variance that we will put in the "bad" bucket. A conceptual formula for the correlation coefficient **r** would be covariablity of X and Y divided by the variability of X and Y separately.

$$r = \frac{\textit{Covariability of X and Y}}{\textit{Variability of X and Y separately}}$$

Once we find **r**, another statistic that provide helpful information is **r squared**. r^2 is the proportion of variability in one variable that can be

explained by the relationship with the other variable. Make note of this fact, because the proportion of variability explained by a **correlation** is a very helpful metric.

Now we can examine what form a hypothesis test would take in the context of a **correlational** research design. Such a hypothesis test asks the question, "how unlikely is it that the **correlation** coefficient is actually zero?"

In step 1, in order to keep the hypothesis in a form similar to what we did before, we can identify the populations in a particular way. Population 1 will be "people like those in the sample," and population 2 will be "people who show no relationship between the variables." That way, the research hypothesis can be set up as "The correlation for population 1 is [greater than/less than/different from] the correlation for population 2. The null hypothesis can be "The correlation for population 1 is the same as the correlation for population 2."

In step 2, we need to find the characteristics of the comparison distribution, and in this case we need the correlation coefficient **r**, which can range from -1 to 1. An **r** value of 0 indicates there is no **correlation** whatsoever between the two measured variables. An **r** of 1 is a perfect positive **correlation**, and an **r** of -1 is a perfect negative correlation. Most **correlations** in real life fall closer to 0 than to 1 or -1.

$$r = \frac{\sum(Z_X \times Z_Y)}{N}$$

This correlation coefficient formula makes use of Z-scores, which is a great way to review these standardized scores covered in an earlier chapter. Recall that

$$Z = \frac{X - M}{SD}$$

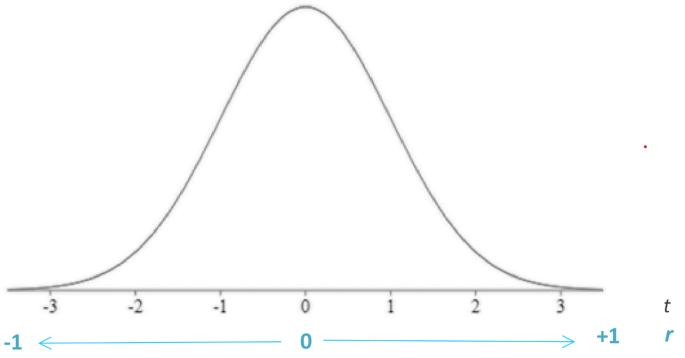
, where

$$SD = \frac{\sum(X - M)^2}{N}$$

For each variable, X and Y, we must calculate the mean and standard deviation of the variable, so each score can be translated to Z-scores. Only then can they be cross-multiplied and then summed in the **r** formula.

Once we calculate the **r** value for a **correlation**, we can test the

statistical significance of this value, based on how extreme it is on the t distribution. An **r** of 0 is placed in the centre of the t distribution, as the comparison distribution mean, and positive one and negative one are placed at either tail of the distribution.



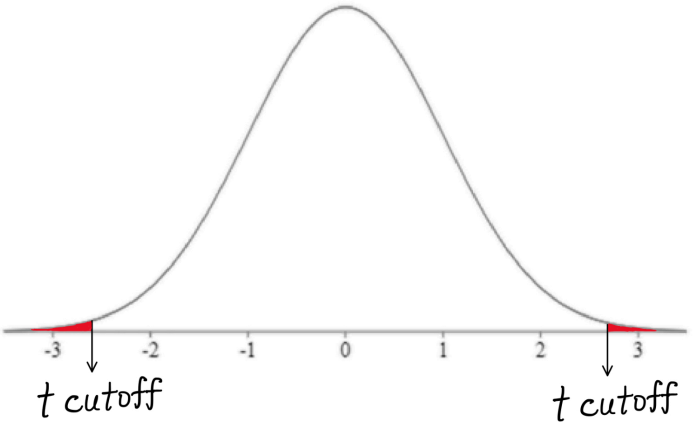
The further out we get into the appropriate tail, the better our chance of rejecting the null hypothesis of a zero correlation. The bad news is, we are back to the t-test, which means we have to think about directionality. The good news is, this is a great opportunity to refresh ourselves on how the t-test works.

In step 3, we find the cutoff score using the t tables. For correlation degrees of freedom will be *N*-2, where N is the number of people in the sample. This is so, because we have two measured (numeric) variables, each of which has *N*-1 scores that are free to vary.

In step 4, the t-test is calculated as **r** divided by S_r, where S_r quantifies the unexplained variability.

$$t = \frac{r}{S_r}$$

Step 5 is the decision: we reject the null hypothesis if the t-test result falls in the shaded tail beyond the cutoff.

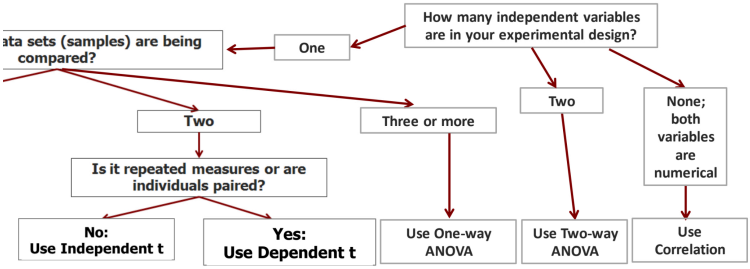


We could express our hypothesis test results on the relationship between income and grades in the following manner:

"We found that there was a significant positive correlation between family income and student grade average ($r = 0.65$, $t_{11} = 2.97$, $p < 0.05$)."

Notice that our interpretation is *not* that we found higher family income results in a higher grade average. Why not? Well, as we said before, causal conclusions require experimental design. To draw such a conclusion regarding the relationship between family income and student grade average, we would need to randomly assign students into family income conditions, wealthy or poor, then measure the effects of that manipulation on their grades. Just like our drowning example, this seems not only logistically challenging, but also rather unethical. So, we are limited to **correlation** here for a reason, and thus we simply need to characterize our findings as a relationship or pattern, rather than a statement of cause and effect.

As we put the final branch to our decision tree, we now have a decision flow for the situation of no independent variables. If both variables are numerical, you must use correlation to test their relationship.



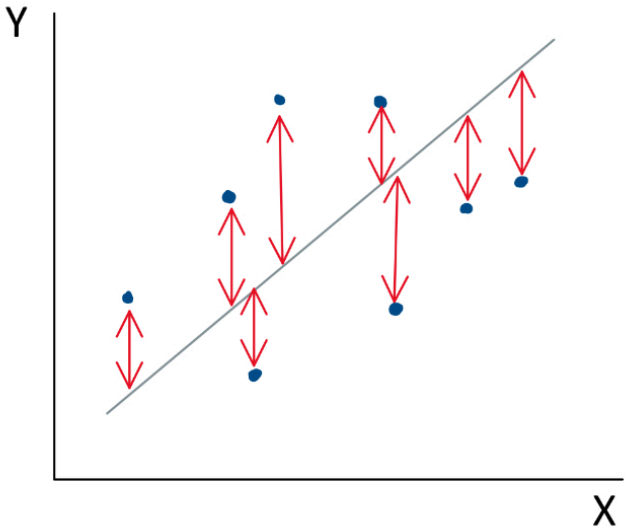
10b. Regression

In the next part of the chapter, we will examine the statistical technique of **regression**. **Regression** allows us to extend the findings of a **correlation** to predict the future from the past.

Once we have calculated a **correlation**, a **regression** allows us to predict how an individual would perform on one variable based on their performance on another variable. In an example of a correlation between income and grades, the **regression** would allow us to see what grade level would be achieved by an individual with a family income level that was not actually collected in our dataset. We could also identify the income level based on a given grade level.

The **regression** line is a line through our scatter plot that can be described with an equation. The equation has two components: slope and intercept. The slope says how many units up (or down) the line goes for each unit over. The intercept says where the line hits the y axis.

The **regression** line is a line that "best fits" the data points that we have collected. Mathematically, it is the line that minimizes the squared deviations (i.e. error) of the individual points from the line.



To find the equation for the **regression** line, you can calculate slope *b* and then intercept *a* using the formulas shown.

$$b = \frac{(X - M_X)(Y - M_Y)}{SS_X}$$

Steps to find b, slope of regression line

1. For each individual, find the deviation of the X score from the mean.*
2. For each individual, find the deviation of the Y

score from the mean.*

3. For each individual, multiply the deviation of X by the corresponding deviation of Y
4. Add together the products from step 3 for all individuals.
5. Divide this sum by SS_x.*

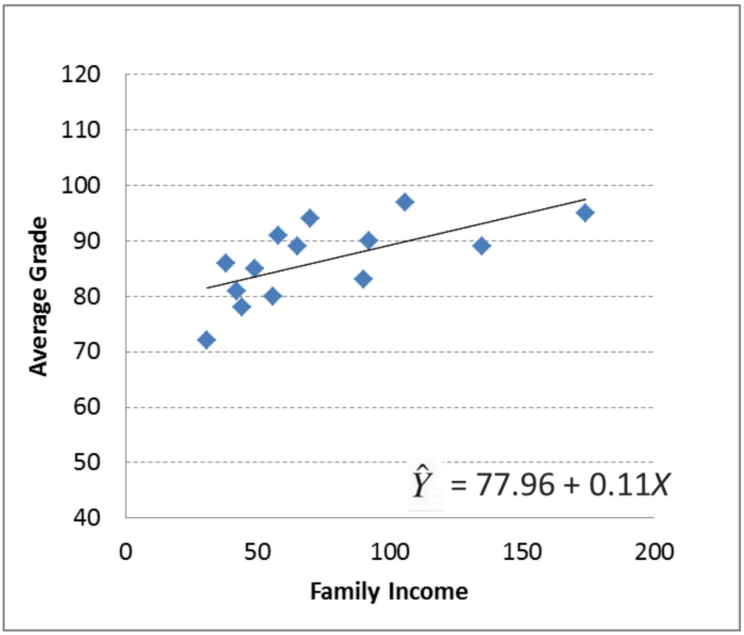
*These calculations should already be completed for correlation.

$$a = M_Y - (b)M_X$$

Once *a* and *b* are calculated, we can plug these numbers into the **regression** line equation.

$$\widehat{Y} = a + b(X)$$

Here I will show you the regression line equation for our family income vs. grade example.



b is 0.11, which means that for every one unit of Family Income, the line goes up 0.11 unit of Average Grade. *a* is 77.96, which means that the line meets the y axis at a height of 77.96.

The line equation allows us to plot the precise **regression** line on the scatter plot. To plot a regression line, pick two X values that are on the low and the high end of the scale. Plug those into the line equation to find the corresponding Y values that are on the line.

Using the regression line, you can predict X value from Y values and Y values from X values. This means that even if you did not have someone in your dataset with a family income of 105, you can figure out what a student's average grade would have been if they had that family income. Likewise, if you had no one in your dataset with an average grade of 75, you can figure out what their family income would have been if they had that grade. Note that these are just predictions. They are imperfect, and do not take into account other factors or individual variability.

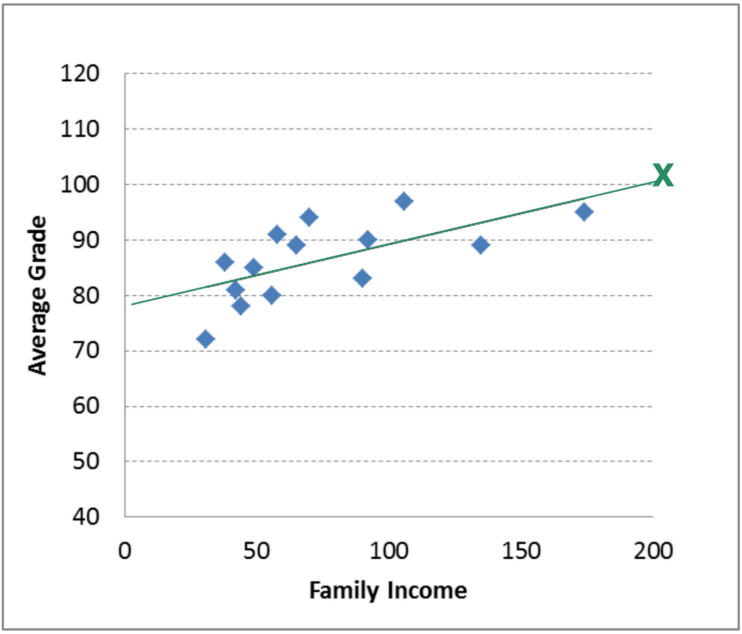
Here we will try try predicting the average grade (Y) for a student

who has a family income of 200. To do this, we will plug 200 in for X in the regression line equation (as shown here).

$$\widehat{Y} = 77.96 + 0.11(X)$$

$$\widehat{Y} = 77.96 + 0.11(200)$$

$$\widehat{Y} = 100.55$$



The result is a grade of 100.55. Of course getting a grade average above 100% is impossible (at least at many institutions). In this case, our prediction shows a "ceiling effect". This means that there is a maximum average grade that we hit before we hit a maximum family income. Therefore, the **regression** line equation becomes useless above a family income of around 190.

Now, we can try predicting family income (Y) for a student with an average grade of 60 (X). To do this, you must plug in 60 for Y in the equation, then solve for X.

$$\widehat{Y} = 77.96 + 0.11(X)$$

$$60 = 77.96 + 0.11(X)$$

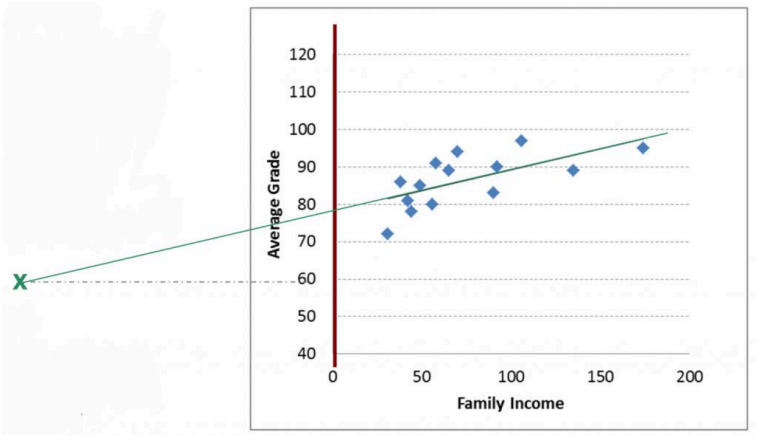
Notice that to rearrange the equation to solve for X, you first have to move intercept (a) over:

$$(60 - 77.96) = 0.11(X)$$

Then you have to divide by the slope:

$$\frac{(60 - 77.96)}{0.11} = X$$

Now you are ready to solve for X: -159. The result of finding X for the Y of 60 is a negative income! This is, of course, impossible (or very unlikely). Here we can see the floor effect.



This means that there is a minimum family income that we reach before reaching the minimum grade. So the **regression** line becomes useless below an average grade of 77.96 (the Y intercept). Floor and ceiling effects are common problems for **regression**, and you should watch out for these problems when you use this technique. We can see that the **regression** line for this particular dataset is useful to make predictions for the grade of 80-100 average grade and the range of 0-190 income level.

Now of course, predictions are not perfect. **Regression** allows for a prediction of one variable from another variable. As we can see in

our scatterplots, not every real data point is exactly on the regression line. The actual data point might be different. Why is that? Because, unless it's a perfect **correlation**, some variability in the real data is not accounted for by the regression equation. We can estimate just how accurate our predictions are by looking at **r squared**. $\mathbf{r^2}$ is the proportion of variance in one variable explained by its relationship with the other variable. The rest is the amount that is not accounted for.

Just as we can include multiple factors in ANOVA, we can also include multiple predictive variables in a **regression**. We will not attempt that in this course, but if you take more advanced statistics course you will see that the more variables you include, each explaining a piece of the variability in the criterion variable, the more precise your **regression** model will become. Here, we are using just one predictive variable, and our $\mathbf{r^2}$ is likely to be well shy of 100% explained variance. So in that case, we can expect our **regression** to be only modestly accurate.

Chapter Summary

This chapter introduced you to the statistical techniques of **correlation** and **regression**. We saw how we can detect and describe the strength and direction of the relationship between two numeric variables, and to run a hypothesis test to find out if the **correlation** is significantly different from zero. Finally, we saw that **regression** can generate a linear model allow for the prediction of one variable from the other. A key reminder: **correlation** does *not* equal causation. These techniques suit research designs that do not meet the requirements of experimental design, and as such, our conclusions regarding the statistical findings must avoid cause-effect language.

Key terms:

correlation	**regression**	**r squared**
covariance	**r**	

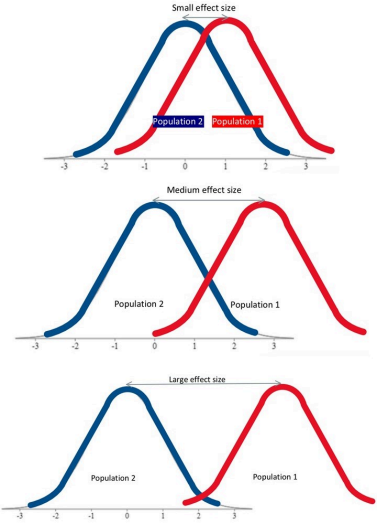
11. Beyond Hypothesis Testing

11. Beyond Hypothesis Testing

In this chapter we will introduce some big concepts that go beyond hypothesis testing. The three big concepts we need to cover to finish off our coverage of inferential statistics are **effect size**, **power** and **confidence intervals**. Effect size measures just how big a difference between means is, or just how much variability a **regression** model explains. **Effect size** is a vital piece of any inferential statistics to complement tests of statistical significance. **Power** is a critical concept whenever we test a statistical hypothesis – the power to find statistical significance if the research hypothesis is in fact true. And **confidence intervals** are another approach to inferential statistics that offers an alternative to hypothesis testing. All three of these concepts are extremely important, and they only got left until the end so that we could get all the way through our tour of the full range of statistical tests and techniques first, before zooming out again to these big, important ideas that complement hypothesis testing procedures.

When a hypothesis test reveals that there is a significant difference between means, the question remains... how big was the difference?

Measures of **effect size** seek to quantify just how big the effect was of an experiment. For each test of statistical significance, there is a corresponding test of **effect size**. We already looked at r-squared as the measure of **effect size** for a correlation, but we have not yet discussed **effect size** measures for differences between means. **Effect size** metrics increase with greater differences between the means we are comparing. But how much of a distance between two means is a good, or large **effect size** and how much is not so good, or small **effect size**?



Cohen's d is the most popular measure of effect size, and it is very easy to use. To calculate **Cohen's d**, we just take the difference between two means and divide by the standard deviation for the distribution of individuals. The formula shown here is to find the effects size of the difference between means in a situation in which we know the comparison population standard deviation, so the same scenario in which we would conduct a single-sample Z-test to determine whether the difference between means is statistically significant.

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

Of course, we do not know μ_1, the research population mean, so we must use the sample mean *M* as the estimate in its place.

Once we calculate **Cohen's d**, we can place it on this relative scale to judge the **effect size**. A **Cohen's d** of 0.2 is small, 0.5 is medium, and 0.8

Cohen's d	
0.20	Small
0.50	Medium
0.80	Large

is considered large. So if our two means are a whole standard deviation apart, then we have a very good **effect size**. Together with a finding of statistical significance from a Z-test, we would have excellent evidence to suggest there is a true difference between the means we were comparing. If, on the other hand, we have a significant Z-test results, but a small **effect size**, then the difference between the means might be statistically significant, but perhaps it is not such an important difference. This is a common pattern when a research study has a very large sample size, such that it is "over-powered." The concept of **power** will come up a little later in the chapter.

Such a situation came up when researchers conducted **effect size** analyses on the many clinical trials published to establish the efficacy of antidepressant drugs like SSRIs on the symptoms of depression. Once they looked at **effect size**, they realized that although antidepressants had a statistically significant effect on symptoms, it was a fairly small effect. This called into question the true clinical efficacy of these medications, the most widely prescribed class of psychotropic medication on the planet. Now scientific journals are far more cautious and require researchers to report not just tests of statistical significance, but also tests of **effect size**, to help the reader interpret the results.

The next concept on our bucket list is **power**. Statistical **power** is defined as the probability that the study will produce a statistically significant result if the research hypothesis is true. In other words, when there is a true difference between means, will we be able to detect that? Or will we miss that fact and make a Type II error, failing to reject the null hypothesis when it is false? **Power** is important, and it directly depends on two factors we have already focused on. **Effect size**, or the difference between means, is one factor that impacts **power**. The other factor is sample size, which translates into the width of the distributions being compared. As you will see, **power** has to do with minimizing the amount of overlap between distributions, which can be minimized with a) a robust difference between means and b) narrow distributions.

Factors affecting power

Two factors play into power, the probability of detecting statistical significance when we should.

- **Effect size**: Difference between means
- **Sample size**: Width of distributions

Recall that in hypothesis testing comparing two means, we use the distribution of sample means as the comparison distribution. The standard deviation (i.e. the width of the distribution) is found by dividing by the square root of sample size.

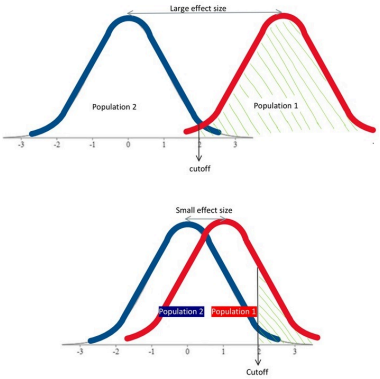
$$\sigma_M = \frac{\sigma}{\sqrt{N}} or S_M = \frac{S}{\sqrt{N}}$$

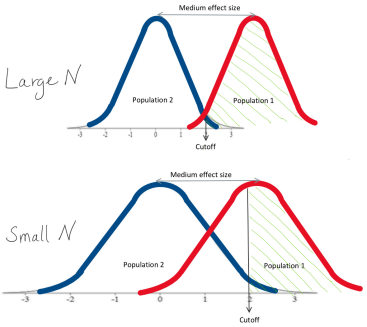
Thus, as sample size increases, the distribution becomes narrower.

Some illustrations will help you picture how these two elements play into power. First, let us examine the relationship between effect size and power, in a simple situation of comparing the means of two distributions.

In these images, the shading under the population 1 distribution represents **power**. Where does the shading start? Where the cutoff sample score falls on the population 2 distribution. The cutoff score for

a Z-test that is two-tailed, with .05 significance level, would be at +/- 1.96. Here the shading starts at about Z=2 on the population 2 curve, and continues to the end of the population 1 curve. As you can see here, if the distance between distributions is robust, so the **effect size** is large, then we have good **power**. We will be have a pretty good chance of detecting a true significant difference. We might think of this as having enough "good variance". On the other hand, if we have a very modest distance between distributions, so the **effect size** is small, then the distributions overlap too much, leaving little shaded area. In this scenario, then, we have little chance of rejecting the null hypothesis even when we should.





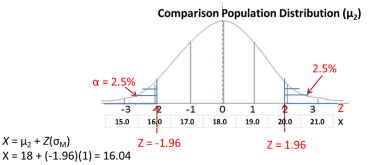
The other factor that determines statistical **power** is sample size, which plays into the width of sampling distributions. When we have a large sample size, the distributions are narrower, and thus there will be less overlap between them. With **effect size** held constant, just widening the distributions is enough to bring us from very good power to rather poor power. We can think of this as having too much "bad variance".

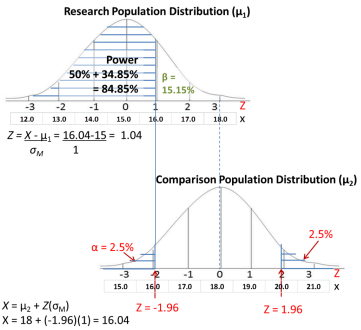
So of course, to maximize **power** we would like to have a large **effect size** and a large sample size. In reality, of course, we have little control over **effect size**. As an example, either our insomnia medication helps people sleep a lot longer, or it only helps them sleep a little longer. There is little we can do about just how much of an impact our medication has over sleep. What we researchers can do, though, is make sure that our statistical analysis is based on data from many participants. Even with a moderate **effect size**, if we have a large sample, we will have nice narrow distributions and thus have a very

good shot at detecting a statistically significant result if our medication does help people sleep... even if it is only a modest boost.

Let us take a look at how we can calculate power precisely for the simplest research design we have encountered – a Z-test.

First, we draw out the comparison population distribution using the mean and standard deviation we identified from Step 2 of hypothesis testing. On a sketch of the normal distribution, map Z-scores to raw scores (X). In this example, the mean of the comparison distribution is 18, with a standard deviation of 1. As you can see, as Z-scores go up by one, so does the raw score. And as the Z-scores go down by one, so does the raw score.





Next, we draw out the research population distribution using the research sample mean *M* that we calculated for step 4 of hypothesis testing as centre and σ_M as standard deviation. Align the distributions according to the raw scores. Here the sample mean is 15, so we line it up the centre of this new distribution with a raw score of 15 on the comparison distribution.

Next, draw a line straight up from your cutoff score. Shade in everything from there toward the side of the distribution that is away from the middle of the comparison distribution.

Either visually estimate the Z-score where the shading starts on the research population distribution, or find the precise Z-score mathematically by converting the Z-score cutoff to a raw score on comparison distribution, then converting that raw score to a Z-score on the research population distribution. I have shown you what those conversions would look like in the illustrations.

Finally, find the shaded area on the research population distribution by looking up the appropriate area in the Z tables. We have shaded more than half the curve, so we just need to add to 50% the area from the mean to the Z-score for the calculated Z-score of 1.04. As you can

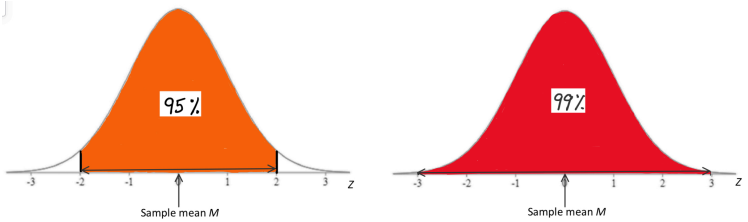
see in the table the % mean to Z is 34.85%, so the total shaded area is 84.85%

The area you shaded corresponds to your **power**, the probability of detecting a significance difference if it existed. So in this particular example, we have pretty good **power**. Our chance of detecting statistical significance if it is the correct conclusion is about 85%, which leaves 15% chance of making a Type II error (symbolized as $\boldsymbol{\beta}$). We are generally not too worried about Type II error, as long as it is less than 50% risk, so this would be considered adequate statistical power.

Our final concept is **confidence intervals**. **Confidence intervals** are a complete alternative procedure to hypothesis testing. They offer the same information, but in a different format and with a different flow of logic. Instead of reporting a test value and its level of significance, we can report a range of values, into which the real value would fall with a certain probability. That range of values is the confidence interval. A 95% **confidence interval** states that the research population mean would fall in the range of values 95% of the time. This is a similar idea as the hypothesis test with significance level 0.05. A 99% **confidence interval** states that the research population mean would fall in the range of values 99% of the time. This is a similar idea as the hypothesis test with significance level 0.01.

So how do **confidence intervals** work? Well, the goal of inferential statistics (of any sort related to experimental design) is to estimate a population mean, from which a sample came, and decide if that population mean is different from the comparison population mean. To accomplish this judgement with hypothesis testing, we figure out comparison mean and variability around that mean. We then use the sample mean the find out if the research population mean differs from the comparison population mean. To accomplish the same judgement with **confidence intervals**, we start instead from the sample mean and the variability around that mean. We determine the range in which the research population mean must fall by using the variability measure. We then decide whether the comparison population mean falls within that range or beyond it. So the entire distinction is whether you start with attentional focus on the comparison population mean or the research population mean. And how you represent the results looks different.

Here is a visual representation of **confidence intervals** and how they would look for a situation when we know the comparison population standard deviation, as we would when conducting a Z-test.



With this approach, we would use the normal distribution's areas under the curve to figure out the range we need to capture the research population mean with a particular probability. As you can see, the range from a Z-score of -2 to +2 is pretty close to 95% probability, and the range of -3 to +3 is close to 99%. We place the sample mean, calculated from our research study sample, in the middle of this normal distribution. The task, then, is to add either 2 or 3 standard deviations to the mean to find the upper bound, and subtract either 2 or 3 standard deviations to find the lower bound of the confidence interval. How many standard deviations you must go to be sure you include the population mean depends on how confident you need to be. Are you okay with 95% confidence, or should you have 99% confidence? That decision is based on tolerance for Type I error, just like with significance levels in hypothesis testing.

To find precisely the Z-scores matching 95 or 99% **confidence intervals**, we can use the normal curve areas table. For a 95% **confidence interval**, just like with a two-tailed hypothesis test with significance level of .05, the Z score that precisely corresponds with 2.5% area in the tails on either side of the curve, or 47.5% area from the mean to the Z score, is Z = 1.96. To find the lower end of the **confidence interval**, use this formula:

$$M - (1.96)(\sigma_M)$$

In other words, subtract 1.96 standard deviations from the sample mean. To find the upper end of the interval, use this formula:

$$M + (1.96)(\sigma_M)$$

In other words, add 1.96 standard deviations to the sample mean. Once you have those two limits, you can make the claim that there is a 95% chance that the research population mean corresponding to

this sample mean is between the lower end and the upper end of the **confidence interval**.

Once you have that range, you can also determine whether there is a statistically significant difference between the research and comparison population mean, by answering this question: does the comparison population mean fall within the reported **confidence interval**? If so, there is no significant difference. The two means are too close together. If it falls outside the interval, then you can report a significant difference between means.

So you end up at the same place, whether you use hypothesis testing or **confidence intervals**. The difference is simply in where you start from. Also, hypothesis testing can be one-tailed, but it is uncommon to see directional **confidence intervals**. Finally, if p-values are reported precisely, hypothesis testing offers slightly more information, but if they are not reported, **confidence intervals** offer more information. It is a matter of professional preference which technique you choose as a method for making inferences about data following a research study.

Chapter Summary

In this chapter, we covered three big concepts that are vital to understand in the context of inferential statistics. **Effect size** offers information critical to interpreting a finding of statistical significance. **Power** tells us whether tests of significance are even worth doing, or if we are at too much risk of Type II error. **Confidence intervals** offer a system for statistical inference that represents a full alternative to the hypothesis testing procedure we employed for the past several chapters.

Key terms:

effect size **confidence intervals** **Cohen's *d***
power

12. Afterword

Here we are, at the end of the course. Now, a whole universe of statistics memes are there for you to explore. Seriously, though, I hope you can look back to the beginning of the book and see how far you have come with your understanding of statistical concepts and practical approaches to applying them in a research setting.

Are there more techniques to explore? For sure! Could you go deeper into the mathematical underpinnings of these statistical methods? Totally. But you should now have a good foundation that is ready to apply to typical experimental and correlational research designs.

And hopefully, you have deeply internalized the notion that science is about deep humility and about measuring likely truths in probabilities. Now you see what I meant in the beginning of the course, that this is a course about decision making... making sound judgements from data. And now, hopefully, you have a good basic set of tools to use.

Thanks for your hard work and practice, and your willingness to let me guide you – the key to learning any new skill. I hope it has been an enjoyable journey, and that you continue your exploration of data analysis.

PART II

HOMEWORK ASSIGNMENTS

Homework Chapter 1

Assignment:

Objective: Apply concepts Variable, Value, and Score

1. A geologist employed by a ski resort measures the stability of a rock face near the lodge on a scale of 1 to 100, with 1 meaning it will collapse today, and 100 meaning it will last until the end of time. This particular rock face receives a 64. This is an example of a _____.
2. A geologist employed by a ski resort measures the stability of a rock face near the lodge on a scale of 1 to 100, with 1 meaning it will collapse today, and 100 meaning it will last until the end of time. In this example, stability is the _____.
3. A geologist employed by a ski resort measures the stability of a rock face near the lodge on a scale of 1 to 100, with 1 meaning it will collapse today, and 100 meaning it will last until the end of time. The scale of 1 to 100 represents the _____.

Objective: Apply concepts Nominal, Numeric

4. A nurse in an intensive care unit measures the risk of superbug infection in the ward using a metric based on the percentage chance of a patient becoming ill with such an infection in a two week period of admission. Values range from 0-100%. The variable of superbug infection risk would best be described as _____.
5. A nurse in an intensive care unit measures the risk of superbug infection in the ward by recording which types of fungi/bacteria are present in each patient's screening upon discharge. The variable of superbug infection risk would best be described as _____.

Objective: Apply techniques Grouped Frequency Table, Histogram, Describe Distribution Shape

6. A forensic psychologist measures 20 inmates in a prison on the Hare Psychopathy Checklist – Revised (PCL-R). The results are as

follows: 32, 40, 18, 22, 38, 25, 15, 19, 36, 33, 10, 24, 22, 18, 37, 33, 27, 25, 40, 26. Create a grouped frequency table using value ranges that begin with 1s or 6s and results in 4 rows. Create a histogram using the same ranges. Describe the shape of the smoothed distribution in terms of both symmetry and peaks. The percentage of inmates that scored in the 31-40 range (and thus qualified as "Psychopaths") is _____.

7. A forensic psychologist measures 20 inmates in a prison on the Hare Psychopathy Checklist – Revised (PCL-R). The results are as follows: 32, 40, 18, 22, 38, 25, 15, 19, 36, 33, 10, 24, 22, 18, 37, 33, 27, 25, 40, 26. Create a grouped frequency table using value ranges that begin with 1s or 6s and results in 4 rows. Create a histogram using the same ranges. Describe the shape of the smoothed distribution in terms of both symmetry and peaks. The histogram's X-axis is labeled with _____.
8. A forensic psychologist measures 20 inmates in a prison on the Hare Psychopathy Checklist – Revised (PCL-R). The results are as follows: 32, 40, 18, 22, 38, 25, 15, 19, 36, 33, 10, 24, 22, 18, 37, 33, 27, 25, 40, 26. Create a grouped frequency table using value ranges that begin with 1s or 6s and results in 4 rows. Create a histogram using the same ranges. Describe the shape of the smoothed distribution in terms of both symmetry and peaks. The most appropriate description of the distribution shape is _____.

Homework Chapter 2

Assignment:

Objective: Apply techniques Find/Calculate Mean, Median, Mode

1. Find the mean, median and mode of the following data: 3, 8, 4, 10, 5, 6, 3, 2, 4, 3. The mean is _____.
2. Find the mean, median and mode of the following data: 3, 8, 4, 10, 5, 6, 3, 2, 4, 3. The median is_____.
3. Find the mean, median and mode of the following data: 3, 8, 4, 10, 5, 6, 3, 2, 4, 3. The mode is _____.

Objective: Apply concepts Compare measures of Central Tendency.

4. A researcher finds that their data on the number of errors children make in a corn maze, for which they have prepared using virtual reality headsets before attempting it, include a couple of outliers. Two children make a lot of errors, whereas the vast majority of children only make a few. The distribution of the data can thus be described as unimodal but right-skewed. In this case, what might we expect are the relative locations of the mean and the median? The mean is _____ than the median.
5. A researcher replicates a study on the number of errors children make in a corn maze, for which they have prepared using virtual reality headsets before attempting it. This time some children make very few errors, and a similar number of children make many errors. The distribution of the data can thus be described as bimodal. The best measure of central tendency is likely to be the _____.
6. A researcher replicates a study on the number of errors children make in a corn maze, for which they have prepared using virtual reality headsets before attempting it. Most children make a moderate amount of errors, but some make very few and some make many errors. The distribution of data appears unimodal and quite symmetrical. The mean, median and mode for this dataset will be _____.

Objective: Apply techniques Find/Calculate Variance, Standard Deviation

7. Calculate the variance and standard deviation of the following data: 3, 8, 4, 10, 5, 6, 3, 2, 4, 3. The variance is _____.
8. Calculate the variance and standard deviation of the following data: 3, 8, 4, 10, 5, 6, 3, 2, 4, 3. The standard deviation is _____.

Objective: Apply concepts Compare measures of Variability.

9. A journal editor's instructions to submitting authors includes the guidelines that variability should be reported in units of measurement that are the same as the data (e.g., if number of days is the unit of measurement of the data reported, then the variability reported should also be in terms of number of days). The best way to report variability would be _____.

Objective: Apply concepts Interpret measures of Central Tendency and Variability.

10. A criminologist reports that the number of crimes in downtown Vancouver on summer weekends was M=5.9 (SD=2.2). This indicates that _____.

Homework Chapter 3

Assignment:

Objective: Apply techniques Convert between raw scores and Z-scores

1. On the Multidimensional Scale of Perceived Social Support, the mean of a university student sample was 5.80, with a standard deviation of 0.86. For a student who scored 6.2 on the scale, what is the Z-score? The Z-score is _____.
2. On the Multidimensional Scale of Perceived Social Support, the mean of a university student sample was 5.80, with a standard deviation of 0.86. For a student who scored 4.1 on the scale, what is the Z-score? The Z-score is _____.
3. On the Multidimensional Scale of Perceived Social Support, the mean of a university student sample was 5.80, with a standard deviation of 0.86. For a student whose Z-score is 0, what is the raw score? The raw score is _____.
4. On the Multidimensional Scale of Perceived Social Support, the mean of a university student sample was 5.80, with a standard deviation of 0.86. For a student whose Z-score is -0.80, what is the raw score? The raw score is _____.

Objective: Apply concepts Compare scores on different scales using standard scores.

5. Two siblings wish to compete for bragging rights on who performed best on their standardized test – but one took the LSAT for law school and the other took the MCAT for medical school. Sarah scored 163 (*M*=151.88, *SD*=9.95), but Rachel scored 510 on the MCAT (*M*=500.9, *SD*=10.6). The sibling who scored highest was _____.

Objective: Apply techniques Working with Areas under the Normal Curve

6. Assuming that scores on a creativity measure are normally distributed, with a mean of 90 and a standard deviation of 10, what

is the probability of an individual scoring below 60? The probability is _____.

7. Assuming that scores on a creativity measure are normally distributed, with a mean of 90 and a standard deviation of 10, between which values do the middle 50% of people score? (Hint: the middle 50% would be if you shade outward from the middle of the distribution in either direction until 50% of the distribution is shaded.) The middle 50% score between _____.
8. Assuming that scores on a creativity measure are normally distributed, with a mean of 90 and a standard deviation of 10, what is the probability of an individual scoring above 72? The probability is _____.
9. At an art school, a student is expressing a great deal of pride for being in the 90th percentile on a creativity measure that is normally distributed, with a mean of 90 and a standard deviation of 10. What raw score did the student have? The score is _____.

Homework Chapter 4

Objective: Review concepts Steps of Hypothesis Testing Procedure

1. List and explain the logic of the five hypothesis testing steps. Step 1 of hypothesis testing is _____.
2. List and explain the logic of the five hypothesis testing steps. Step 3 of hypothesis testing is _____.
3. List and explain the logic of the five hypothesis testing steps. Step 5 of hypothesis testing is _____.

Objective: Apply techniques Conducting a hypothesis test

4. In a study, researchers want to compare task performance of participants who see images related to ambiguous partial word stimuli presented later, to other participants who see images that are unrelated. The number of words completed in the task is recorded to see if there is a difference. Data: Normative neuropsychological data have established that the mean words completed by the population of people who have seen unrelated images is 5, with standard deviation 2.4. The individual in this study who saw related images completed 9 words. In this hypothesis test, the two populations are _____.
5. In a study, researchers want to compare task performance of participants who see images related to ambiguous partial word stimuli presented later, to other participants who see images that are unrelated. The number of words completed in the task is recorded to see if there is a difference. Data: Normative neuropsychological data have established that the mean words completed by the population of people who have seen unrelated images is 5, with standard deviation 2.4. The individual in this study who saw related images completed 9 words. In this hypothesis test, the hypotheses are _____.
6. In a study, researchers want to compare task performance of participants who see images related to ambiguous partial word stimuli presented later, to other participants who see images that are unrelated. The number of words completed in the task is recorded to see if there is a difference. Data: Normative

neuropsychological data have established that the mean words completed by the population of people who have seen unrelated images is 5, with standard deviation 2.4. The individual in this study who saw related images completed 9 words. In this hypothesis test, Step 2 characteristics of the comparison distribution would be _____.

7. In a study, researchers want to compare task performance of participants who see images related to ambiguous partial word stimuli presented later, to other participants who see images that are unrelated. The number of words completed in the task is recorded to see if there is a difference. Data: Normative neuropsychological data have established that the mean words completed by the population of people who have seen unrelated images is 5, with standard deviation 2.4. The individual in this study who saw related images completed 9 words. In this hypothesis test, assuming a significance level of .01, the cutoff score(s) in Step 3 is/are _____.
8. In a study, researchers want to compare task performance of participants who see images related to ambiguous partial word stimuli presented later, to other participants who see images that are unrelated. The number of words completed in the task is recorded to see if there is a difference. Data: Normative neuropsychological data have established that the mean words completed by the population of people who have seen unrelated images is 5, with standard deviation 2.4. The individual in this study who saw related images completed 9 words. In this hypothesis test, Step 4 would give the Z-score result of _____.
9. In a study, researchers want to compare task performance of participants who see images related to ambiguous partial word stimuli presented later, to other participants who see images that are unrelated. The number of words completed in the task is recorded to see if there is a difference. Data: Normative neuropsychological data have established that the mean words completed by the population of people who have seen unrelated images is 5, with standard deviation 2.4. The individual in this study who saw related images completed 9 words. In this hypothesis test, the conclusion in Step 5 should be _____.

Homework Chapter 5

Assignment:

Objective: Review concepts Central Limit Theorem

1. According to the Central Limit Theorem, which conditions must be met in order to make the normal curve assumption? The conditions are that _____.

Objective: Apply concepts Distribution of Means

2. If the sample size is 100, then the distribution of means will have a standard deviation that is _____ the standard deviation of the distribution of individuals.

Objective: Apply concepts Decision Matrix and Error Types

3. A researcher predicts that positive emotions will increase the likelihood of forming false memories. In this case, a Type I error would be _____.

Objective: Apply techniques Conducting a Single Sample Z-test

4. A researcher predicts that positive emotions will increase the likelihood of forming false memories. Across metanalysis of many studies has revealed that the norm for false memory creation is 0.72 on the DRM paradigm recognition test, with a standard deviation of 0.29. The scores for the research sample who experienced the induction of positive emotions before completing the DRM recognition test are presented in the table below *(N*=34). Conduct a hypothesis test of the researcher's prediction with significance level .05. The populations should be defined as _____.

0.93
0.96
0.63
0.92
0.77
0.75
0.72
0.98
0.84
0.52
0.92
0.89
0.92
0.82
0.76
0.87
0.89
0.93
0.54
0.52
0.80
0.91
0.94
0.69
0.97
0.88
0.71
0.99
0.93
0.75
0.77
0.97

0.51
0.98

5. For the research scenario described in problem 4, the hypotheses should be defined as _____.
6. For the research scenario described in problem 4, the characteristics of the comparison distribution should be defined as _____.
7. For the research scenario described in problem 4, the cutoff Z score should be _____.
8. For the research scenario described in problem 4, the calculated Z-test result should be _____.
9. For the research scenario described in problem 4, the hypothesis test decision should be _____.
10. For the research scenario described in problem 4, the p-value associated with the Z-test result is _____.
11. For the research scenario described in problem 4, the conclusion as written in a published results section could be _____.

Objective: Apply techniques Conducting a Single Sample t-test

12. A researcher predicts that positive emotions will increase the likelihood of forming false memories. Across metanalysis of many studies has revealed that the norm for false memory creation is 0.72 on the DRM paradigm recognition test. The population standard deviation is unclear. The scores for the research sample who experienced the induction of positive emotions before completing the DRM recognition test are presented in the table below *(N*=34). Conduct a hypothesis test of the researcher's prediction with significance level .05. For this research scenario, the characteristics of the comparison distribution should be defined as _____.

0.93
0.96
0.63
0.92
0.77
0.75
0.72
0.98
0.84
0.52
0.92
0.89
0.92
0.82
0.76
0.87
0.89
0.93
0.54
0.52
0.80
0.91
0.94
0.69
0.97
0.88
0.71
0.99
0.93
0.75
0.77
0.97

0.51
0.98

13. For the research scenario described in problem 12, the cutoff t score should be _____.
14. For the research scenario described in problem 12, the calculated t-test result should be _____.
15. For the research scenario described in problem 12, the hypothesis test decision should be _____.
16. For the research scenario described in problem 12, the p-value associated with the t-test result is _____.
17. For the research scenario described in problem 12, the conclusion as written in a published results section could be _____.

Objective: Apply concepts Decision Tree

18. The Z-test is appropriate to use when _____.

Homework Chapter 6

Assignment:

Objective: Apply concepts Repeated Measures Design

1. A researcher is contemplating various ways to design their study examining the effects of irrigation on crop yield. The first possible design would be to take a sample of yields from plots with a standard irrigation system and compare them to the known yield from plots without irrigation in the same region. Another would be to sample irrigated and non-irrigation crop yields from various regions and compare them, ensuring the plots are matched on the variable of region. A final option would be to sample crop yields from the same plots first before and then after installation of a standard irrigation system and compare the yields prior to irrigation to those following irrigation. The research design that would best represent repeated measures design is _____.

Objective: Apply techniques Conducting a Dependent Means t-test

2. A researcher predicts crop yields sampled from the same plots first before and then after installation of a standard irrigation system will differ. The scores for the research samples before and after irrigation are presented in the table below, in units (bushels per hectare; *N*=28). Conduct a hypothesis test of the researcher's prediction with significance level .10. The populations should be defined as _____.

Plot	Before irrigation	After irrigation
1	163	208
2	131	198
3	160	208
4	183	198
5	168	186
6	135	177
7	179	173
8	143	203
9	161	203
10	154	192
11	162	186
12	163	185
13	166	176
14	181	184
15	132	199
16	149	197
17	169	199
18	148	202
19	169	188
20	158	200
21	143	201
22	150	191
23	178	194
24	147	185
25	170	202
26	171	178
27	169	212
28	180	180

3. For the research scenario described in problem 2, the hypotheses

should be defined as _____.

4. For the research scenario described in problem 2, the characteristics of the comparison distribution should be defined as _____.
5. For the research scenario described in problem 2, the cutoff t score should be _____.
6. For the research scenario described in problem 2, the calculated t-test result should be _____.
7. For the research scenario described in problem 2, the hypothesis test decision should be _____.
8. For the research scenario described in problem 2, the p-value associated with the t-test result is _____.
9. For the research scenario described in problem 2, the conclusion as written in a published results section could be _____.

Objective: Apply concepts Decision Tree

10. The dependent means t-test is appropriate to use when _____.

Homework Chapter 7

Assignment:

Objective: Apply concepts Independent Means Design

1. A researcher is contemplating various ways to design their study examining the effects of irrigation on crop yield. The first possible design would be to take a sample of yields from plots with a standard irrigation system and compare them to a sample of yields from plots without irrigation in the same region. Another would be to sample irrigated and non-irrigation crop yields from various regions and compare them, ensuring the plots are matched on the variable of region. A final option would be to sample crop yields from the same plots first before and then after installation of a standard irrigation system and compare the yields prior to irrigation to those following irrigation. The research design that would require an independent means t-test is _____.

Objective: Apply techniques Conducting a Dependent Means t-test

2. A researcher predicts crop yields from plots with a standard irrigation system will be greater than yields from plots without irrigation in the same region. The scores for the research samples are presented in the table below, in units (bushels per hectare). Conduct a hypothesis test of the researcher's prediction with significance level .05. The populations should be defined as _____.

Plot	Non-irrigated	Plot	Irrigated
1	163	*29*	208
2	131	*30*	198
3	160	*31*	208
4	183	*32*	198
5	168	*33*	186
6	135	*34*	177
7	179	*35*	173
8	143	*36*	203
9	161	*37*	203
10	154	*38*	192
11	162	*39*	186
12	163	*40*	185
13	166	*41*	176
14	181	*42*	184
15	132	*43*	199
16	149	*44*	197
17	169	*45*	199
18	148	*46*	202
19	169	*47*	188
20	158	*48*	200
21	143	*49*	201
22	150	*50*	191
23	178	*51*	194
24	147	*52*	185
25	170	*53*	202
26	171	*54*	178
27	169	*55*	212
28	180	*56*	180

3. For the research scenario described in problem 2, the hypotheses

should be defined as _____.

4. For the research scenario described in problem 2, the characteristics of the comparison distribution should be defined as _____.
5. For the research scenario described in problem 2, the cutoff t score should be _____.
6. For the research scenario described in problem 2, the calculated t-test result should be _____.
7. For the research scenario described in problem 2, the hypothesis test decision should be _____.
8. For the research scenario described in problem 2, the p-value associated with the t-test result is _____.
9. For the research scenario described in problem 2, the conclusion as written in a published results section could be _____.

Homework Chapter 8

Assignment:

Objective: Apply concepts Factors and Levels

1. A psychology honours students is conducting a study to determine whether various categories of social support can impact cognitive performance. They measure response times on the Stroop task in people who report a low, medium, or high degree of social support from friends. In this research design, the factor is _____.
2. A psychology honours students is conducting a study to determine whether various categories of social support can impact cognitive performance. They measure response times on the Stroop task in people who report a low, medium, or high degree of social support from friends. In this research design, the levels are _____.

Objective: Apply techniques Conducting a Dependent Means t-test

3. A psychology honours students is conducting a study to determine whether various categories of social support can impact cognitive performance. They measure response times on the Stroop task in people who report a low, medium, or high degree of social support from friends. The scores for the research samples are presented in the table below, in seconds. Conduct a hypothesis test of the researcher's prediction with significance level .05. The populations should be defined as _____.

Low	*Med*	*High*
9.15	6.35	7.30
9.00	8.00	7.34
8.56	8.24	9.20
10.10	6.80	6.55
9.02	6.39	7.49
8.18	7.40	7.24
7.30	7.29	10.66
7.88		

4. For the research scenario described in problem 3, the hypotheses should be defined as _____.
5. For the research scenario described in problem 3, the characteristics of the comparison distribution should be defined as _____.
6. For the research scenario described in problem 3, the cutoff F score should be _____.
7. For the research scenario described in problem 3, the calculated F-test result should be _____.
8. For the research scenario described in problem 3, the hypothesis test decision should be _____.
9. For the research scenario described in problem 3, the p-value associated with the F-test result is _____.
10. For the research scenario described in problem 3, the conclusion as written in a published results section could be _____.

Objective: Apply concepts Experimentwise alpha level

11. A psychology honours students is conducting a study to determine whether various categories of social support can impact cognitive performance. They measure response times on the Stroop task in people who report a low, medium, or high degree of social support from friends. If t-tests were used to conduct tests of statistically significant difference between means, at significance level .05, the experimentwise alpha level

would effectively become _____.

Objective: Apply concepts Planned Contrasts and Posthoc tests

12. Given the results of the one-way ANOVA analysis conducted for the research scenario described in problem 3, the researchers would conduct the following analyses to explore which groups differed significantly from each other: _____.

Objective: Review techniques Bonferroni and Scheffé Corrections

13. The Bonferroni correction is typically used for planned contrasts and protects against inflation of Type I error risk by _____.
14. The Scheffé correction is typically used for post-hoc tests and protects against inflation of Type I error risk by _____.

Homework Chapter 9

Coming soon...

Homework Chapter 10

Assignment:

Objective: Review concept Correlation and Causation

1. The two major requirements for causal conclusions from an experimental design are _____.

Objective: Apply concepts Correlation, Causation

2. If a strong positive correlation between the viewing of violent films and antisocial behaviour were observed, many might assume that viewing violent films causes antisocial behaviour. This would be a faulty assumption because _____.

Objective: Apply technique Correlation

3. A researcher seeks to replicate a study that found a significant positive relationship between women's symptoms of obsessive-compulsive disorder and their core beliefs regarding the importance of holding unrelenting standards. The data collected are shown below. Using a scatter plot, examine the strength and direction of the relationship between the two variables. The relationship appears to be _____.

Client	OCD (X)	Beliefs (Y)
1	4.08	7.20
2	4.43	3.69
3	6.56	7.27
4	1.72	3.62
5	3.79	2.83
6	1.51	4.71
7	4.32	0.64
8	6.16	7.62
9	5.65	0.22
10	2.61	5.08
11	3.97	0.42
12	2.79	3.36
13	1.36	1.61
14	1.71	3.15
15	3.10	3.25
16	3.26	4.45
17	3.40	4.11
18	0.81	4.32
19	1.63	0.20
20	4.03	1.98
21	3.24	5.16
22	2.00	6.67
23	3.92	0.82
24	7.03	3.09
25	0.95	0.52
26	0.29	3.14

4. For the research scenario described in problem 3, for a hypothesis test on correlation, assume the Populations are defined as follows. Population 1: Women like those in this study; Population 2: Women for whom there is no (positive) relationship between the two variables. The hypotheses should be defined as _____.
5. For the research scenario described in problem 3, the

characteristics of the comparison distribution should be defined as _____.

6. For the research scenario described in problem 3, assuming a significance level of .01, the cutoff t score should be _____.
7. For the research scenario described in problem 3, the calculated t-test result should be _____.
8. For the research scenario described in problem 3, the hypothesis test decision should be _____.
9. For the research scenario described in problem 3, the p-value associated with the t-test result is _____.
10. For the research scenario described in problem 3, the conclusion as written in a published results section could be _____.

Objective: Apply technique Regression

11. For the research scenario described in problem 3, the slope of the best fit regression line would be _____.
12. For the research scenario described in problem 3, the intercept of the best fit regression line would be _____.

Objective: Apply concepts Prediction

13. For the research scenario described in problem 3, the predicted core beliefs score for a woman with the OCD symptom score of 3.70 would be _____.
14. The accuracy of predictions made from the regression model from the data in the research scenario described in problem 3 would be _____.

Homework Chapter 11

Coming soon...

Key Terms List

Σ

in summation notation, a symbol that denotes “taking the sum” of a series of numbers

α

the probability of making a Type I error; used as shorthand for significance level

β

the probability of making a Type II error; the antithesis of power

Analysis of Variance

also called ANOVA, a system of data analysis that is very flexible and adaptable to a variety of research designs. It is based on a statistical concept called the general linear model and involves the technique of partitioning variance.

bimodal

a descriptor of a distribution indicating that there are two peaks, or two collections of scores

Bonferroni correction

adjustment to avoid inflation of experimentwise risk of Type I error, by dividing significance level by the number of planned contrasts to be conducted

central limit theorem

mathematical theorem that proposes the following: as long as we take a decent-sized sample, if we took many samples (10,000) of large enough size (30+) and took the mean each time, the distribution of those means will approach a normal distribution, even if the scores from each sample are not normally distributed

central tendency

a statistical measure that defines the centre of a distribution with a single score

Cohen's d

a measure of effect size commonly used to quantify the difference between two population means; 0.2 is small, 0.5 is medium, and 0.8 is considered large

confidence intervals

an approach to inferential statistics that serves as an alternative to hypothesis testing. A statement of where a research population mean should lie with a particular probability.

correlation

statistical analysis of the direction and strength of the relationships between two numerical variables

covariance

the variability that two numeric variables have in common

cutoff sample score

critical value that serves as a decision criterion in hypothesis testing

degrees of freedom

the number of scores in a given calculation that are free to vary.

dependent means t-test

a test for statistical significance when comparing mean difference scores to zero in repeated measures or matched pairs designs

dependent variables

a variable you measure to detect a difference/change as a result of the manipulation -- most often it is numeric

descriptive

ways to summarize or organize data from a research study – essentially allowing us to describe what the data are

directional hypothesis

a research prediction that the research population mean will be “greater than” or "less than" the comparison population mean

distribution of means

also called a sampling distribution, is the distribution of many sample means drawn from the population of individual scores

do not reject the null hypothesis

a decision in hypothesis testing that is inconclusive because the sample score is less extreme than the cutoff score

effect size

a measure of how well a statistical model explains variability, apart from statistical significance, e.g. how big a difference between means is, or just how much variability a regression model explains

experimentwise alpha level

the problem of accumulating risk of Type I error with multiple statistical tests on the same data

factor

in ANOVA, a grouping variable used to account for variance among scores; in an experiment a factor is an independent variable

frequency tables

a way to summarize a dataset in table form, to organize the data and make it easy to get an overview of the dataset quickly

general linear model

an extension of the statistical technique linear regression that is

adaptable to various combinations of independent (nominal) and dependent (numeric) variables

grouped frequency table

a frequency table that defines ranges of values in the first column, and reports the frequency of scores that fall within each range

histogram

a graph for summarizing numeric data that essentially is a frequency table that has been turned on its side, with the added benefit of a visual representation of the frequency as the height of the bars in the graph, rather than just a number

homoscedasticity assumption

independent means t-tests require the assumption that the two populations we are comparing have the same variance

hypothesis testing

a formal decision making procedure often used in inferential statistics

independent means t-test

a statistical test used in hypothesis tests comparing the means of two independent samples, created by random assignment of individuals to experimental and control groups

independent variables

a variable you manipulate -- most often it is categorical, or nominal

inferential

analytical tools that allow us to draw conclusions based on data from a research study -- essentially allowing us to make a statement about what the data mean

interaction

the degree to which the contribution of one factor to explaining

variability in the data depends on the other factor; the synergy among factors in explaining variance

left skewed

a descriptor of a distribution that indicates asymmetry, specifically with a low frequency tail leading off to the left

levels

the individual conditions or values that make up a factor, a nominal variable that forms the groups in analysis of variance

levels of statistical significance

the probability level that we are willing to accept as a risk that the score from our research sample might occur by random chance within the comparison distribution. By convention, it is set to one of three levels: 10%, 5%, or 1%.

M

the symbol for the mean (average) of scores in a sample

main effect

the degree to which one of the factors explains variability in the data when taken on its own, independent of the other factor

matched pairs

a research design for which a dependent means t-test may be used to test for a hypothesis test; in this design two separate samples are used, but each individual in a sample is matched one-to-one with an individual in the other sample, most often matching participants on a a possible confounding variable as a way to control for the effects of that variable

mean

the same thing as an average: you add up all the numbers, then divide by how many numbers there were. Conceptually we can think of the mean as the balancing point for the distribution.

median

the midpoint of the scores after placing them in order. The median is a counting-based measure: the point at which half of the scores fall above and half of the scores fall below.

mode

the score(s) that occur(s) most often in the dataset

N

the symbol for the number of scores in a sample

nominal

variables that label or categorize something, and any numbers used to measure these variables are arbitrary and do not indicate quantity or size

non-directional hypothesis

a research prediction that the research population mean will be “different from" the comparison population mean, but allows for the possibility that the research population mean may be either greater than or less than the comparison population mean

normal curve

a theoretical distribution, sometimes called a Z distribution, has a very distinct set of properties that make it a useful model for data analysis (e.g. 2-14-34% area rule)

normal curve assumption

parametric tests like the t-test and Z-test require the assumption that the distribution of means for any given population is normally distributed

null hypothesis

the prediction that the population from which sample came is not different from the comparison population

numeric

variables for which numbers are actually meaningful -- they indicate the size or amount of something

one-tailed test

a hypothesis test in which there is only one cutoff sample score on either the lower or the upper end of the comparison distribution

p-value

the probability of the observed sample score or more extreme occurring at random under the comparison distribution

participant variables

variables used like independent variables in (quasi-)experimental research designs, but which cannot be manipulated or assigned randomly to participants, and as such must not generate cause-effect conclusions

partitioning variance

the allocation of variability among scores in numeric data into different buckets, like treatment effects vs. error, or between-groups vs. within-groups variance

percentile

the score at which a given percentage of scores in the normal distribution fall beneath

planned contrasts

statistical tests of pairwise comparisons among groups, used to follow up on a significant ANOVA result, when researchers know in advance which groups they expect to differ

population

all possible individuals or scores about which we would ideally draw conclusions

posthoc tests

statistical tests of pairwise comparisons among groups, used to follow up on a significant ANOVA result, when researchers do not know in advance which groups they expect to differ and wish to test all possible combinations

power

the probability of rejecting the null hypothesis (i.e. finding statistical significance) if the research hypothesis is in fact true. Depends on effect size and sample size.

probability

in a situation where several different outcomes are possible, the probability of any specific outcome is a fraction or proportion of all possible outcomes

r

correlation coefficient that describes the strength and direction of the relationship between two numeric variables. Can be between -1 and 0 and between 0 and +1.

r squared

proportion of variability in one variable that can be explained by the relationship with the other variable. Can be between 0 and 1.

regression

a statistical model that allows for prediction based on a trend line that "best fits" the data points that we have collected. Mathematically, a regression line is one that minimizes the squared deviations (i.e. error) of each point from the line.

reject the null hypothesis

a decision in hypothesis testing that concludes statistical significance because the sample score is more extreme than the cutoff score

repeated measures

also known as within-subjects designs or pre-test post-test design, in which the experiment involves obtaining two separate scores for each individual in a single sample. The same participants are used in all treatment conditions.

research hypothesis

prediction that the population from which the research sample came is different from the comparison population

right skewed

a descriptor of a distribution that indicates asymmetry, specifically with a low frequency tail leading off to the right

sample

the individuals or scores about which we are actually drawing conclusions

Scheffe's correction

in posthoc analyses, an adjustment to correct for inflated experimentwise risk of Type I error, by dividing the F value by the overall degrees of freedom between from the original overall ANOVA analysis

score

a particular individual's value on the variable

standard deviation

a common measure of variability in numeric data. The average distance of a scores from the mean.

standard error of the mean

standard deviation of the distribution of means

statistically significant

the conclusion from a hypothesis test that probability of the

observed result occurring randomly within the comparison distribution is less than the significance level

Sum of Squares

the sum of squared deviations, or differences, between scores and the mean in a numeric dataset

t-distributions

a series of distributions, based on the normal distribution, that adjust their shape according to degrees of freedom (which in turn is based on sample size)

t-test

statistical test to test the differences between two population means. Suitable for single sample design when standard deviation is unknown, or in two-sample designs.

two-tailed test

a hypothesis test in which there are two cutoff sample scores, one on either end of the comparison distribution

Type I error

if we made the decision to reject the null hypothesis when it is true

Type II error

if we made the decision to not reject the null hypothesis but the research hypothesis is true

μM

population mean for the distribution of means

unimodal

a descriptor of a distribution indicating that there is one peak, or a single collection of scores

value

any possible number or category that a variable could take on

variable

a quality or a quantity that is different for different individuals

variance

a common measure of variability in numeric data. The average squared distance of scores from the mean.

Z-scores

standard scores that allow us to transform scores in any numeric dataset, using any scale, into a standard metric

Z-test

statistical hypothesis test suitable for comparing the means of two populations, when the comparison population mean and standard deviation are known

Normal Curve (Z) Area Tables

Z	% mean to Z	% in tail
0.00	0.00%	50.00%
0.01	0.40%	49.60%
0.02	0.80%	49.20%
0.03	1.20%	48.80%
0.04	1.60%	48.40%
0.05	1.99%	48.01%
0.06	2.39%	47.61%
0.07	2.79%	47.21%
0.08	3.19%	46.81%
0.09	3.59%	46.41%
Z	**% mean to Z**	**% in tail**
0.10	3.98%	46.02%
0.11	4.38%	45.62%
0.12	4.78%	45.22%
0.13	5.17%	44.83%
0.14	5.57%	44.43%
0.15	5.96%	44.04%
0.16	6.36%	43.64%
0.17	6.75%	43.25%
0.18	7.14%	42.86%
0.19	7.53%	42.47%
Z	**% mean to Z**	**% in tail**
0.20	7.93%	42.07%
0.21	8.32%	41.68%
0.22	8.71%	41.29%
0.23	9.10%	40.90%
0.24	9.48%	40.52%
0.25	9.87%	40.13%
0.26	10.26%	39.74%
0.27	10.64%	39.36%
0.28	11.03%	38.97%
0.29	11.41%	38.59%
Z	**% mean to Z**	**% in tail**

0.30	11.79%	38.21%
0.31	12.17%	37.83%
0.32	12.55%	37.45%
0.33	12.93%	37.07%
0.34	13.31%	36.69%
0.35	13.68%	36.32%
0.36	14.06%	35.94%
0.37	14.43%	35.57%
0.38	14.80%	35.20%
0.39	15.17%	34.83%
Z	**% mean to Z**	**% in tail**
0.40	15.54%	34.46%
0.41	15.91%	34.09%
0.42	16.28%	33.72%
0.43	16.64%	33.36%
0.44	17.00%	33.00%
0.45	17.36%	32.64%
0.46	17.72%	32.28%
0.47	18.08%	31.92%
0.48	18.44%	31.56%
0.49	18.79%	31.21%
Z	**% mean to Z**	**% in tail**
0.50	19.15%	30.85%
0.51	19.50%	30.50%
0.52	19.85%	30.15%
0.53	20.19%	29.81%
0.54	20.54%	29.46%
0.55	20.88%	29.12%
0.56	21.23%	28.77%
0.57	21.57%	28.43%
0.58	21.90%	28.10%
0.59	22.24%	27.76%
Z	**% mean to Z**	**% in tail**
0.60	22.57%	27.43%

0.61	22.91%	27.09%
0.62	23.24%	26.76%
0.63	23.57%	26.43%
0.64	23.89%	26.11%
0.65	24.22%	25.78%
0.66	24.54%	25.46%
0.67	24.86%	25.14%
0.68	25.17%	24.83%
0.69	25.49%	24.51%
Z	**% mean to Z**	**% in tail**
0.70	25.80%	24.20%
0.71	26.11%	23.89%
0.72	26.42%	23.58%
0.73	26.73%	23.27%
0.74	27.04%	22.96%
0.75	27.34%	22.66%
0.76	27.64%	22.36%
0.77	27.94%	22.06%
0.78	28.23%	21.77%
0.79	28.52%	21.48%
Z	**% mean to Z**	**% in tail**
0.80	28.81%	21.19%
0.81	29.10%	20.90%
0.82	29.39%	20.61%
0.83	29.67%	20.33%
0.84	29.95%	20.05%
0.85	30.23%	19.77%
0.86	30.51%	19.49%
0.87	30.78%	19.22%
0.88	31.06%	18.94%
0.89	31.33%	18.67%
Z	**% mean to Z**	**% in tail**
0.90	31.59%	18.41%
0.91	31.86%	18.14%

0.92	32.12%	17.88%
0.93	32.38%	17.62%
0.94	32.64%	17.36%
0.95	32.89%	17.11%
0.96	33.15%	16.85%
0.97	33.40%	16.60%
0.98	33.65%	16.35%
0.99	33.89%	16.11%
Z	**% mean to Z**	**% in tail**
1.00	34.13%	15.87%
1.01	34.38%	15.62%
1.02	34.61%	15.39%
1.03	34.85%	15.15%
1.04	35.08%	14.92%
1.05	35.31%	14.69%
1.06	35.54%	14.46%
1.07	35.77%	14.23%
1.08	35.99%	14.01%
1.09	36.21%	13.79%
Z	**% mean to Z**	**% in tail**
1.10	36.43%	13.57%
1.11	36.65%	13.35%
1.12	36.86%	13.14%
1.13	37.08%	12.92%
1.14	37.29%	12.71%
1.15	37.49%	12.51%
1.16	37.70%	12.30%
1.17	37.90%	12.10%
1.18	38.10%	11.90%
1.19	38.30%	11.70%
Z	**% mean to Z**	**% in tail**
1.20	38.49%	11.51%
1.21	38.69%	11.31%
1.22	38.88%	11.12%

1.23	39.07%	10.93%
1.24	39.25%	10.75%
1.25	39.44%	10.56%
1.26	39.62%	10.38%
1.27	39.80%	10.20%
1.28	39.97%	10.03%
1.29	40.15%	9.85%
Z	**% mean to Z**	**% in tail**
1.30	40.32%	9.68%
1.31	40.49%	9.51%
1.32	40.66%	9.34%
1.33	40.82%	9.18%
1.34	40.99%	9.01%
1.35	41.15%	8.85%
1.36	41.31%	8.69%
1.37	41.47%	8.53%
1.38	41.62%	8.38%
1.39	41.77%	8.23%
Z	**% mean to Z**	**% in tail**
1.40	41.92%	8.08%
1.41	42.07%	7.93%
1.42	42.22%	7.78%
1.43	42.36%	7.64%
1.44	42.51%	7.49%
1.45	42.65%	7.35%
1.46	42.79%	7.21%
1.47	42.92%	7.08%
1.48	43.06%	6.94%
1.49	43.19%	6.81%
Z	**% mean to Z**	**% in tail**
1.50	43.32%	6.68%
1.51	43.45%	6.55%
1.52	43.57%	6.43%
1.53	43.70%	6.30%

1.54	43.82%	6.18%
1.55	43.94%	6.06%
1.56	44.06%	5.94%
1.57	44.18%	5.82%
1.58	44.29%	5.71%
1.59	44.41%	5.59%
Z	**% mean to Z**	**% in tail**
1.60	44.52%	5.48%
1.61	44.63%	5.37%
1.62	44.74%	5.26%
1.63	44.84%	5.16%
1.64	44.95%	5.05%
1.65	45.05%	4.95%
1.66	45.15%	4.85%
1.67	45.25%	4.75%
1.68	45.35%	4.65%
1.69	45.45%	4.55%
Z	**% mean to Z**	**% in tail**
1.70	45.54%	4.46%
1.71	45.64%	4.36%
1.72	45.73%	4.27%
1.73	45.82%	4.18%
1.74	45.91%	4.09%
1.75	45.99%	4.01%
1.76	46.08%	3.92%
1.77	46.16%	3.84%
1.78	46.25%	3.75%
1.79	46.33%	3.67%
Z	**% mean to Z**	**% in tail**
1.80	46.41%	3.59%
1.81	46.49%	3.51%
1.82	46.56%	3.44%
1.83	46.64%	3.36%
1.84	46.71%	3.29%

1.85	46.78%	3.22%
1.86	46.86%	3.14%
1.87	46.93%	3.07%
1.88	46.99%	3.01%
1.89	47.06%	2.94%
Z	**% mean to Z**	**% in tail**
1.90	47.13%	2.87%
1.91	47.19%	2.81%
1.92	47.26%	2.74%
1.93	47.32%	2.68%
1.94	47.38%	2.62%
1.95	47.44%	2.56%
1.96	47.50%	2.50%
1.97	47.56%	2.44%
1.98	47.61%	2.39%
1.99	47.67%	2.33%
Z	**% mean to Z**	**% in tail**
2.00	47.72%	2.28%
2.01	47.78%	2.22%
2.02	47.83%	2.17%
2.03	47.88%	2.12%
2.04	47.93%	2.07%
2.05	47.98%	2.02%
2.06	48.03%	1.97%
2.07	48.08%	1.92%
2.08	48.12%	1.88%
2.09	48.17%	1.83%
Z	**% mean to Z**	**% in tail**
2.10	48.21%	1.79%
2.11	48.26%	1.74%
2.12	48.30%	1.70%
2.13	48.34%	1.66%
2.14	48.38%	1.62%
2.15	48.42%	1.58%

2.16	48.46%	1.54%
2.17	48.50%	1.50%
2.18	48.54%	1.46%
2.19	48.57%	1.43%
Z	**% mean to Z**	**% in tail**
2.20	48.61%	1.39%
2.21	48.64%	1.36%
2.22	48.68%	1.32%
2.23	48.71%	1.29%
2.24	48.75%	1.25%
2.25	48.78%	1.22%
2.26	48.81%	1.19%
2.27	48.84%	1.16%
2.28	48.87%	1.13%
2.29	48.90%	1.10%
Z	**% mean to Z**	**% in tail**
2.30	48.93%	1.07%
2.31	48.96%	1.04%
2.32	48.98%	1.02%
2.33	49.01%	0.99%
2.34	49.04%	0.96%
2.35	49.06%	0.94%
2.36	49.09%	0.91%
2.37	49.11%	0.89%
2.38	49.13%	0.87%
2.39	49.16%	0.84%
Z	**% mean to Z**	**% in tail**
2.40	49.18%	0.82%
2.41	49.20%	0.80%
2.42	49.22%	0.78%
2.43	49.25%	0.75%
2.44	49.27%	0.73%
2.45	49.29%	0.71%
2.46	49.31%	0.69%

2.47	49.32%	0.68%
2.48	49.34%	0.66%
2.49	49.36%	0.64%
Z	**% mean to Z**	**% in tail**
2.50	49.38%	0.62%
2.51	49.40%	0.60%
2.52	49.41%	0.59%
2.53	49.43%	0.57%
2.54	49.45%	0.55%
2.55	49.46%	0.54%
2.56	49.48%	0.52%
2.57	49.49%	0.51%
2.58	49.51%	0.49%
2.59	49.52%	0.48%
Z	**% mean to Z**	**% in tail**
2.60	49.53%	0.47%
2.61	49.55%	0.45%
2.62	49.56%	0.44%
2.63	49.57%	0.43%
2.64	49.59%	0.41%
2.65	49.60%	0.40%
2.66	49.61%	0.39%
2.67	49.62%	0.38%
2.68	49.63%	0.37%
2.69	49.64%	0.36%
Z	**% mean to Z**	**% in tail**
2.70	49.65%	0.35%
2.71	49.66%	0.34%
2.72	49.67%	0.33%
2.73	49.68%	0.32%
2.74	49.69%	0.31%
2.75	49.70%	0.30%
2.76	49.71%	0.29%
2.77	49.72%	0.28%

2.78	49.73%	0.27%
2.79	49.74%	0.26%
Z	**% mean to Z**	**% in tail**
2.80	49.74%	0.26%
2.81	49.75%	0.25%
2.82	49.76%	0.24%
2.83	49.77%	0.23%
2.84	49.77%	0.23%
2.85	49.78%	0.22%
2.86	49.79%	0.21%
2.87	49.79%	0.21%
2.88	49.80%	0.20%
2.89	49.81%	0.19%
Z	**% mean to Z**	**% in tail**
2.90	49.81%	0.19%
2.91	49.82%	0.18%
2.92	49.82%	0.18%
2.93	49.83%	0.17%
2.94	49.84%	0.16%
2.95	49.84%	0.16%
2.96	49.85%	0.15%
2.97	49.85%	0.15%
2.98	49.86%	0.14%
2.99	49.86%	0.14%
Z	**% mean to Z**	**% in tail**
3.00	49.87%	0.13%
3.01	49.87%	0.13%
3.02	49.87%	0.13%
3.03	49.88%	0.12%
3.04	49.88%	0.12%
3.05	49.89%	0.11%
3.06	49.89%	0.11%
3.07	49.89%	0.11%
3.08	49.90%	0.10%

3.09	49.90%	0.10%
Z	**% mean to Z**	**% in tail**
3.10	49.90%	0.10%
3.11	49.91%	0.09%
3.12	49.91%	0.09%
3.13	49.91%	0.09%
3.14	49.92%	0.08%
3.15	49.92%	0.08%
3.16	49.92%	0.08%
3.17	49.92%	0.08%
3.18	49.93%	0.07%
3.19	49.93%	0.07%
Z	**% mean to Z**	**% in tail**
3.20	49.93%	0.07%
3.21	49.93%	0.07%
3.22	49.94%	0.06%
3.23	49.94%	0.06%
3.24	49.94%	0.06%
3.25	49.94%	0.06%
3.26	49.94%	0.06%
3.27	49.95%	0.05%
3.28	49.95%	0.05%
3.29	49.95%	0.05%
Z	**% mean to Z**	**% in tail**
3.30	49.95%	0.05%
3.31	49.95%	0.05%
3.32	49.95%	0.05%
3.33	49.96%	0.04%
3.34	49.96%	0.04%
3.35	49.96%	0.04%
3.36	49.96%	0.04%
3.37	49.96%	0.04%
3.38	49.96%	0.04%
3.39	49.97%	0.03%

Z	% mean to Z	% in tail
3.40	49.97%	0.03%
3.41	49.97%	0.03%
3.42	49.97%	0.03%
3.43	49.97%	0.03%
3.44	49.97%	0.03%
3.45	49.97%	0.03%
3.46	49.97%	0.03%
3.47	49.97%	0.03%
3.48	49.97%	0.03%
3.49	49.98%	0.02%
3.50	49.98%	0.02%

T distribution tables

Cutoff Scores for the T-distribution

T-Distribution Table (One Tail)

df	.10	.05	.01
1	3.078	6.314	31.821
2	1.886	2.920	6.965
3	1.638	2.353	4.541
4	1.533	2.132	3.747
5	1.476	2.015	3.365
6	1.440	1.943	3.143
7	1.415	1.895	2.998
8	1.397	1.860	2.896
9	1.383	1.833	2.821
10	1.372	1.812	2.764
11	1.363	1.796	2.718
12	1.356	1.782	2.681
13	1.350	1.771	2.650
14	1.345	1.761	2.624
15	1.341	1.753	2.602
16	1.337	1.746	2.583
17	1.333	1.740	2.567
18	1.330	1.734	2.552
19	1.328	1.729	2.539
20	1.325	1.725	2.528
21	1.323	1.721	2.518
22	1.321	1.717	2.508
23	1.319	1.714	2.500
24	1.318	1.711	2.492
25	1.316	1.708	2.485
26	1.315	1.706	2.479
27	1.314	1.703	2.473
28	1.313	1.701	2.467
29	1.311	1.699	2.462
30	1.310	1.697	2.457
60	1.296	1.671	2.390
120	1.289	1.658	2.358

1000	1.282	1.646	2.330
∞	1.282	1.645	2.326

T-Distribution Table (Two Tail)

df	.10	.05	.01
1	6.314	12.706	63.656
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
11	1.796	2.201	3.106
12	1.782	2.179	3.055
13	1.771	2.160	3.012
14	1.761	2.145	2.977
15	1.753	2.131	2.947
16	1.746	2.120	2.921
17	1.740	2.110	2.898
18	1.734	2.101	2.878
19	1.729	2.093	2.861
20	1.725	2.086	2.845
21	1.721	2.080	2.831
22	1.717	2.074	2.819
23	1.714	2.069	2.807
24	1.711	2.064	2.797
25	1.708	2.060	2.787
26	1.706	2.056	2.779
27	1.703	2.052	2.771
28	1.701	2.048	2.763
29	1.699	2.045	2.756
30	1.697	2.042	2.750
60	1.671	2.000	2.660
120	1.658	1.980	2.617

1000	1.645	1.960	2.576
∞	1.645	1.960	2.576

F distribution tables

		$df_{Between}$					
0.10 significance level		1	2	3	4	5	6
df_{Within}	1	39.86	49.50	53.59	55.83	57.24	58
	2	8.53	9.00	9.16	9.24	9.29	9.
	3	5.54	5.46	5.39	5.34	5.31	5.
	4	4.54	4.32	4.19	4.11	4.05	4.
	5	4.06	3.78	3.62	3.52	3.45	3.
	6	3.78	3.46	3.29	3.18	3.11	3.
	7	3.59	3.26	3.07	2.96	2.88	2.
	8	3.46	3.11	2.92	2.81	2.73	2.
	9	3.36	3.01	2.81	2.69	2.61	2.
	10	3.29	2.92	2.73	2.61	2.52	2.
	11	3.23	2.86	2.66	2.54	2.45	2.
	12	3.18	2.81	2.61	2.48	2.39	2.
	13	3.14	2.76	2.56	2.43	2.35	2.
	14	3.10	2.73	2.52	2.39	2.31	2.
	15	3.07	2.70	2.49	2.36	2.27	2.
	16	3.05	2.67	2.46	2.33	2.24	2.
	17	3.03	2.64	2.44	2.31	2.22	2.
	18	3.01	2.62	2.42	2.29	2.20	2.
	19	2.99	2.61	2.40	2.27	2.18	2.
	20	2.97	2.59	2.38	2.25	2.16	2.
	21	2.96	2.57	2.36	2.23	2.14	2.
	22	2.95	2.56	2.35	2.22	2.13	2.
	23	2.94	2.55	2.34	2.21	2.11	2.
	24	2.93	2.54	2.33	2.19	2.10	2.
	25	2.92	2.53	2.32	2.18	2.09	2.
	26	2.91	2.52	2.31	2.17	2.08	2.
	27	2.90	2.51	2.30	2.17	2.07	2.
	28	2.89	2.50	2.29	2.16	2.06	2.
	29	2.89	2.50	2.28	2.15	2.06	1.9
	30	2.88	2.49	2.28	2.14	2.05	1.9
	40	2.84	2.44	2.23	2.09	2.00	1.9
	60	2.79	2.39	2.18	2.04	1.95	1.8

120	2.75	2.35	2.13	1.99	1.90	1.8
infinity	2.71	2.30	2.08	1.94	1.85	1.7

		$df_{Between}$							
0.05 significance level		1	2	3	4	5	6	7	8
df_{Within}	1	161.4	199.5	215.7	224.6	230.2	234	236.8	238.9
	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
	6	6.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10

120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02
infinity	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94

0.01 significance level		$df_{Between}$ 1	2	3	4	5	6	7	8
df_{Within}	1	4052	5000	5403	5625	5764	5859	5928	5982
	2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37
	3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49
	4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80
	5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29
	6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10
	7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84
	8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03
	9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47
	10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06
	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74
	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50
	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30
	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14
	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00

16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99

60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66
infinity	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51

Media Attributions

Chapter 1

Fig. 2.2 from *Pulling Together: A Guide for Front-Line Staff, Student Services, and Advisors* by Ian Cull, Robert L. A. Hancock, Stephanie McKeown, Michelle Pidgeon, and Adrienne Vedan is licensed under CC BY-C 4.0

Chapter 2

Camel Farm in Mongolia 02 by Alexandr frolov is licensed under CC BY-SA 4.0

Chapter 3

"Apple" by Open Grid Scheduler / Grid Engine is licensed under CC0 1.0

"Oranges" by Dious is licensed under CC PDM 1.0

Chapter 4

"Dice" by matsuyuki is licensed under CC BY-SA 2.0

"(not my) toolbox" by erix! is licensed under CC BY 2.0

"Dovetail Dresser" by Didriks is licensed under CC BY 2.0

Chapter 5

"Google Self-Driving Car" by smoothgroover22 is licensed under CC BY-SA 2.0

Chapter 10

"Herp Derp :D" by O hai :3 is licensed under CC BY 2.0

Acknowledgements

Strangely enough, a major acknowledgement for the production of this book should go to the COVID-19 pandemic. Being forced to teach online for the first time meant that I needed to generate scripts for my video lessons. Turns out those scripts, with minor adaptations, made for a basic textbook! Difficulties with a publisher was the other impetus to get me to take a few weeks to adapt the scripts into a text.

BC Campus offers tremendous resources for the creation and adaptation of open educational resources, and they host the textbook.

Kai Zhang is a major contributor to the interactive practice problems and examples in the textbook.

But the biggest thanks by far go to my partner Dimitris, who puts up with my grumps and slumps and cooks for me most days. May you be as fortunate as I, to meet someone who asks no more than I can give, but gives of himself freely.

Next, thanks go to all the students, past, present and future, who share with me what explanations work, inspire new approaches, and even come up with new memes for my use. Nothing spurs me on like the moments of insight that I witness in those learning these concepts, and the realization that they actually enjoy learning something "math-y."

Finally, thanks go to Tiberius, an imperial grey tabby cat, who offers comfort and routine, keeping my life on track with regular demands for food (in exchange for floofy cuddles).