Data Analysis in the Psychological Sciences: A Practical, Applied, Multimedia Approach.

Castro Ruiz, Leyre; Mordkoff, J Toby

Data Analysis in the Psychological Sciences: A Practical, Applied, Multimedia Approach

Data Analysis in the Psychological Sciences: A Practical, Applied, Multimedia Approach

LEYRE CASTRO AND J TOBY MORDKOFF

UNIVERSITY OF IOWA LIBRARIES
IOWA CITY

Contents

Introduction 1

Main Body

Unit 1. Introduction to Statistics for Psychological Science 3
J Toby Mordkoff and Leyre Castro

Unit 2. Managing Data 19
J Toby Mordkoff and Leyre Castro

Unit 3. Descriptive Statistics for Psychological Research 28
J Toby Mordkoff and Leyre Castro

Unit 4. Descriptive Statistics with Excel 56
J Toby Mordkoff

Unit 5. Statistics with R: Introduction and Descriptive Statistics 60
Leyre Castro

Unit 6. Brief Introduction to Statistical Significance 66
Brief Introduction to Statistical Significance
Leyre Castro and J Toby Mordkoff

Unit 7. Correlational Measures 71
Leyre Castro and J Toby Mordkoff

Unit 8. Scatterplots and Correlational Analysis in R 91
Leyre Castro

Unit 9. Simple Linear Regression 102
Leyre Castro and J Toby Mordkoff

Unit 10. Simple Linear Regression in R 121
Leyre Castro

Glossary 127
Leyre Castro

Welcome to **Data Analysis in the Psychological Sciences!**

This open resources textbook contains **10 Units** that describe and explain the main concepts in statistical analysis of psychological data. In addition to conceptual descriptions and explanations of the basic analyses for descriptive statistics, this textbook also explains how to conduct those analyses with common statistical software (Excel) and open-source free software (R).

We hope that you enjoy these materials. If you have any questions, issues, or comments, feel free to contact us:

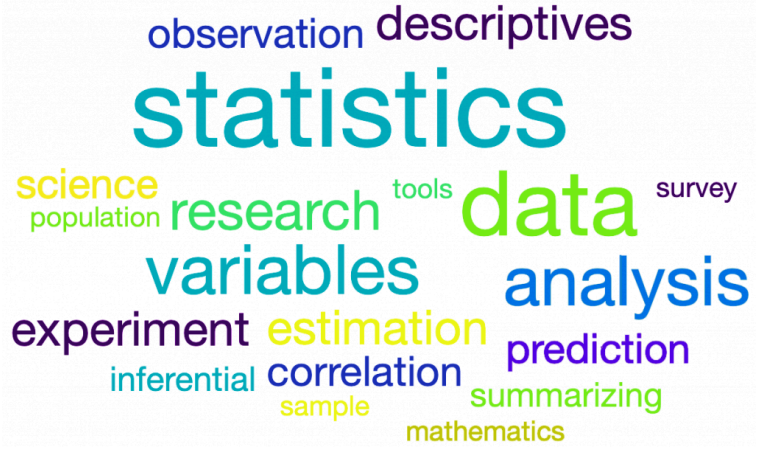
Leyre Castro (leyre-castroruiz@uiowa.edu)

J Toby Mordkoff (jonathan-mordkoff@uiowa.edu)

Unit 1. Introduction to Statistics for Psychological Science

J TOBY MORDKOFF AND LEYRE CASTRO

Summary. This unit introduces some of the basic concepts of statistics as they apply to psychological research. These concepts include data and variables, populations and samples, and the distinction between descriptive and inferential statistics.



Statistics to Better Understand the

World

We live in a data-driven world. We use data in science, sports, business, politics, public health; we collect data from surveys, polls, and ratings; and we measure how much we walk, drive, eat and sleep. We want to know how things are now and how to guide our future actions. Before making a decision, we gather information. Frequently, this means gathering numbers: fuel efficiency, engine capacity, acceleration rate, and market price if we are planning to buy a car; admission rate and likelihood of having a job after graduating if we are deciding what college to attend; or how many miles we run and changes in our heart rate if we are trying to improve our physical fitness. In summary, we are constantly measuring our world, because we want to understand it and make the best decisions.

But collecting numbers and measurements is not enough. We need to try to make sense of all that information, which can sometimes be overwhelming. We don't want to draw false conclusions or make poor decisions. Thus, careful analysis of the data is necessary, to be able to comprehend their meaning, to be able to find patterns, to evaluate actions and behaviors, to uncover trends and estimate the future. Statistics will help us to see through the forest of data more clearly and achieve these goals.

We need to understand and make sense of the data regardless of our biases, wishes, and preferences. That is, we need to be objective and see what the data are telling us, regardless of what we might want to see. That's why we need statistics. Nonetheless, statistics may not give you a simple, single answer. But they will help you condense information so that it's easier to manage and understand. Statistics may not capture the full richness, nuances, color, and texture of the world, but they are the most powerful tool that we have to

objectively understand what we are and what happens around us.

Psychology as an Empirical Science

Psychological science aims to understand human (and animal, in general) behavior from an empirical point of view. That is, we use objective observations and measurements to test our **theories** and **hypotheses** about how humans behave. But psychological science will only be as good as the quality of the observations that we use are.

Anecdotes or intuition or "common sense" may seem to serve us sometimes in our daily lives. But they are subject to all kind of biases. That type of information is inadequate to build a solid and trustable discipline. We need to acquire information using a scientific methodology.

When we do science, it is not about our opinions or wishes. We try to understand events and behaviors by finding patterns in empirical observations and carefully analyzing how those observations are related. Yes, we have opinions, and beliefs, and desires, and intuitions. But none of these can help explain reality.

Science can help us to understand and explain events and behaviors. We develop hypotheses and we conduct detailed empirical observations to test those hypotheses. These detailed observations are our data. Careful statistical analysis of our data will allow us to support or reject our hypotheses. This knowledge will allow us in turn to build theories to understand events and behaviors. These theories will also allow us to make new predictions –to generate new hypotheses that will further advance our comprehension of those events and behaviors.

Statistics may sometimes give us an imperfect

representation of reality. We measure different people at different times in different places, so it may be sometimes difficult to extract conclusions that can apply to absolutely everybody. That's why we need to learn not only to conduct our statistical analysis, but to critique and evaluate others'. Being objective while keeping a critical eye will help us to make the best of our statistical understanding of our world.

Data and Variables

The word ***data*** (which is technically plural) refers to any organized set of observations. An ***observation*** is a value –which can be a number or a label, such as "red"– that tells you something about something. In some cases, an observation specifies the value of a person on some dimension, such as their age or their current level of depression. In other cases, an observation can refer to the behavior of a person, such as what they said when asked a particular question, whether they answered a question correctly, or how quickly they responded on a trial in a laboratory experiment. It does not matter if the observation was collected by watching the person, talking to them, or having a computer measure something like a response time; whenever you get a new piece of information, the value that you got is an observation.

Each observation specifies a value on a dimension, such as age, favorite color, level of depression, correctness, or speed of response. In statistical parlance, these dimensions are referred to as *variables*. A ***variable*** is anything that has more than one possible value.

The values of some variables differ mathematically because they tell you the amount of something. These variables are called ***quantitative***, because they specify a quantity; examples of quantitative variables include age and response time. Most

quantitative variables also have *units*. The units make it clear what the numbers refer to (i.e., how much of *what*). Note that it is quite possible to have two variables that refer to the same general concept, such as time, but use very different units; age vs response time is a good example. For age, we usually use years as the unit, but we could use months or even weeks if we are concerned with infants' ages. For response time, we usually use milliseconds (i.e., thousandths of a second), but we could use minutes if we are asking people to wait in a self-control task. It is crucial to keep track of the units. It would (probably) be very wrong to say that a person took 350 years to make a response or that they were only 19 milliseconds old.

Interval vs Ratio Scales

For most quantitative variables, the value of zero is special, because it means *none* of the thing being measured. Examples of this include response time, number of siblings, or number of study hours per week. In these cases, you are allowed to make ratio statements, such as "Person X took twice as long to respond than Person Y" or "Person X has three times as many siblings as Person Y"; for this reason, any variable with a meaningful zero is referred to as using a *ratio scale*. But not all quantitative variables use ratio scales. If you measure temperature in either Fahrenheit or Celsius, then zero does not mean no heat. (If you want a ratio scale for temperature, you need to use Kelvin, instead.) So you cannot say that 50° F is "twice as warm" as 25° F. But you can say that the difference between 50° F and 45° F is the same as the difference between 55° F and 50° F,

because the steps between values are all the same. For this reason, these variables are said to use an *interval scale* (which is short-hand for *equal-interval scale*). In psychological science, interval scales are often used to measure how much you like something or how much you agree or disagree with some statements.

The distinction between ratio and interval scales does not matter for any of the statistics that we are going to discuss. It only makes a difference to how you talk about the data.

Other variables take on values that differ in ways that are categorical, instead of numerical, in which case the variable is called ***qualitative***, since it specifies a quality. Examples of qualitative variables include favorite food or TV show, ethnic group, and handedness. Some qualitative variables require a label or something like units. It's not enough to know that the value for a person is "strawberry shortcake"; you need to know if this refers to the food or to the cartoon character. The key is to always think of an observation as specifying the value of something on some dimension. If it isn't obvious to what dimension the value refers, make sure to include it somewhere.

There is one other type of variable, which can be thought of as being somewhere between quantitative and qualitative. These are called ***ordinal*** variables. An ordinal variable specifies a position in a sequence or list, and uses a number to do this, such as 1st, 2nd, 3rd, etc., but these numbers do not refer to an amount of something (so they're not quantities). The best example of this in psychology is birth order, which is known to be very important to a variety of things, from attachment (to

the primary caregivers) to performance in school. Assuming a family with three children, none of whom are twins, one child is 1st, another is 2nd, and the last is 3rd. This might seem numerical, but it is not either a ratio or interval scale (see previous box). The 2nd-born is not "twice" the 1st-born in any meaningful way (so it is not a ratio scale), and the gap in time between the 1st and 2nd child does not have to be the same as the gap between 2nd and 3rd (so it isn't an interval scale). In general, it would be inaccurate to take 1st, 2nd, 3rd, etc, as specifying amounts or quantities in the usual sense. But they are more than just categories, because they have an order, which allows us to conduct certain analyses that don't work for qualitative categories, such a favorite color or handedness. Therefore, we keep ordinal variables separate from the other two types.

Discrete vs Continuous Variables

Another way in which variables differ is in terms of what specific values are possible. Some variables are ***discrete***, which means that only certain values are possible. All ordinal variables are discrete by definition, because the only possible values are 1st, 2nd, 3rd, etc. Likewise, every qualitative variable that is used in psychology is also discrete, because there's a limited number of possible values in every case, even for complicated variables, such as ethnic group. Finally, some quantitative variables are also discrete, such as number of siblings, which must be a whole number.

Other quantitative variables are ***continuous***, which means that any value between two end-points is

possible. The best example of this is response time. Although it's true that we usually round the value of response time to the nearest millisecond, an infinite number of values are theoretically possible because time flows continuously, instead of moving in steps, so response time is a continuous variable.

In a very small number of situations, the distinction between discrete and continuous quantitative variables can be important. When one of these situations arises, it will be discussed; for everything else, this distinction can and will be ignored, and we'll refer to both types of quantitative variable as simply being quantitative.

The quantitative vs. qualitative vs. ordinal distinction is inherent to the variable and does not depend on how the variable is being used. Another way that variables differ, however, does depend on the role that the variable is playing. On one side, there are ***independent*** or ***manipulated*** *variables*. These are properties, characteristics, or qualities that are entirely determined or set by the experimenter. For example, in an experiment concerning the effects of sleep deprivation on something like mood, the number of hours of sleep –if it is controlled by the experimenter– is an independent variable (IV). On the other side, there are ***dependent*** or ***measured*** *variables*. These are properties, characteristics, or qualities that are observed as they occur. In the case of the sleep-deprivation experiment, "grumpiness" on a ten-point scale might be the dependent variable (DV).

Note that IV and DV are not automatic labels for a given property, characteristic, or quality. Some variables can be used

as either an IV or DV, depending on the context. For example, hours of sleep can be the IV in a sleep-deprivation experiment, where the amount of sleep is controlled and restricted by the experimenter, or it can be a DV in a study of depression or anxiety, where the participant is asked how much sleep they have been getting in order to take this information into account to better evaluate depression or anxiety. It is crucial that you correctly identify how a given variable is being used in a given line of research. The best way to do this is to remember the alternative names for IV and DV and ask: was this variable ***manipulated*** or was it ***measured*** by the researcher?

Correlational Study vs. Experiment

All research can be classified as taking one of two approaches. If the researcher measures two or more variables, and doesn't use any manipulations, then they are conducting what is called a ***correlational study.*** Note that the label for the research is "correlational study" even if the analysis involves statistics other than correlations; it's the label for the approach, not the analysis. In a correlational study, all of the variables are DVs, even when one is being used as the predictor and the other is being used as the (predicted) outcome.

In contrast, if the researcher manipulates one or more variables and then measures one or more other variables, then they are conducting what is called an ***experimental study***. In an experiment, there is at least one IV and at least one DV. The defining attribute of an experiment is that at least one variable is being manipulated. The values of the

IV in an experiment are what create the different conditions or groups.

There is also what appears to be a third class of variables that is somewhere between an IV and DV. These ***subject variables*** (SVs) are properties, characteristics, or qualities that vary across research subjects, but are relatively stable within subjects (across time) and/or are extremely difficult or impossible to manipulate. Some classic examples are handedness, ethnicity, and sex and/or gender, but "higher-order" examples also exist, such as socio-economic status. The value of an SV is measured, not manipulated, which makes them like a DV, but they are sometimes treated as if they had been manipulated, which makes them a bit like IVs. In other analyses, SVs play a very special role –known as a *covariate*– that is different from both an IV or DV.

The last way in which variables differ has no agreed-upon label; we shall refer to it as *kind*. The first kind of data is ***raw***; these are single observations as they were originally recorded. This can be the response time on one trial in an experiment or the answer to a single item on a questionnaire. This might be surprising, but this kind of data is not used very often in research.

The second kind of data are ***summary scores***; these are created from multiple observations of the same thing under the same set of conditions. The best example of this is how response time is usually measured. It is very rare to run just one trial in each condition of a response-time experiment. Usually, there's 20 or more trials in each condition and the average time across all (correct) trials is kept for analysis. In other words, the 20 or more raw pieces of data are converted to one summary score. (The reason that this is done, as you will see later, is that

summary scores are much less "noisy" than raw scores, which increases the power of the statistical analysis.)

The last kind of data are ***condensed*** or ***composite*** or ***scale scores*** (any of these labels is fine). These are created from multiple observation of different things. Examples of this include the standard measures of depression and anxiety. For both of these, the person is asked many different questions, from which a single (numerical) value can be calculated. This is similar to a summary score in that many raw scores are somehow combined; the difference is that a summary score is based on many measures of the same thing, whereas a condensed score is based on many measures of different things. The reason that condensed scores are used is that it is often impossible to measure all aspects of a psychological construct, such as depression or anxiety, in just one question; to get all of the aspects of the construct of interest, many questions are needed.

The statistical procedures that we will be using make no distinction between the three kinds of data. Put bluntly, the statistical procedures do not care about the kind of data. It is also quite possible for different kinds of data to be mixed in a given analysis. For example, *personality* might be defined in terms of five [condensed] scores (as is done for the Big Five model of personality) and then used in combination with [raw] gender and [summarized] mean number of miles driven per day. Why would one do this? Maybe in an attempt to understand what causes road-rage.

Although the kinds of data being used may have no effect on the method of analysis, they can have profound effects on interpretation. For example, because most information-processing experiments use the mean of response time across many trials (as opposed to the individual, raw values from each of the trials), the proper conclusions from these experiments concern "average" performance and not specific acts. When a difference is observed between two condition means in a

typical response-time experiment, the correct conclusion is that "average" performance in one of the conditions is better than in the other, not that performance is "always" or even "usually" better.

Populations and Samples

Warning: the technical meaning of the word *population* is very different from the everyday meaning.

Recall from earlier that an observation tells you something about something. Taking the above discussion of variables into account, we can now update this to say that an observation tells you the value of a variable. But who or what has this particular value of this particular variable? That is the domain of population and samples.

In the context of research, the word ***population*** refers to the set of all creatures, objects, or events that have values of the variable(s) that you are investigating. If you're interested in depression as measured by a standard questionnaire, then the population will probably be all living people. If you're doing research on the neural connections between two different parts of the brain using drugs or lesions, then your population might be all white laboratory rats, instead of all people. In both of these examples, the word *population* refers to every creature to which your results might apply. This isn't very different from the everyday use of the word *population*.

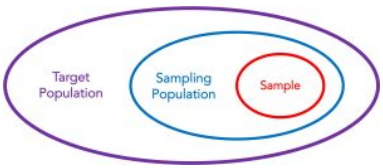
In contrast, if you are doing research on word recognition, then your population might be the set of all words in the language you're using. The words that you use all have values of variables. Each word is from a part-of-speech (i.e., noun, verb, adjective, etc., which is a qualitative variable); each word also has a number of letters and a number of syllables (which

are both discrete quantitative variables). So, the set of all words in a language is also a population.

Similarly, if you are studying road-rage by doing field observations, then your population might be something like intersections. These also have values on variables, such as number of lanes or length of red-light. Finally, in some (rare) cases, the population can be the set of all events of a certain type, such as all hurricanes in the last 100 years. These also have values on variables, such as maximum wind-speed or amount of rainfall.

One thing that all populations have in common is that it is difficult to impossible to include every member or instance in your research. This is where samples come in. A ***sample*** is a subset of a population – it is the subset that was actually included in your experiment or correlational study. In the ideal case, your sample will be "representative" of the whole population. (A sample is said to be representative when all members of the population have an equal chance of being included.) In some situations, this might not be possible. This brings us to another distinction. The term ***target population*** refers to the set of all people, objects, or events in which you are interested and/or want to make statements about. The term ***sampling population*** refers to the set of all people, objects, or events that have a chance of being included in your sample. In the ideal case, the sampling population will be the same as the target population. In actual research, they are often different.

The typical relationship between the target population, sampling population, and sample is shown in the picture to the right. Note that it is possible (in a few, rare situations) for the sampling population to be as large as and, therefore, the same as the target population, but the sample is always a (relatively small) subset of the sampling population. This is what makes it a sample.



What allows us to draw conclusions about the entire target population from just the data in a sample? As suggested by the previous picture, this is a two-step process. The first step goes from the sample to the sampling population (e.g., from the people who actually signed up for your experiment to all students in the Elementary Psychology Subject Pool). This is achieved by a certain type of statistics, which will be defined next and will be described in more detail in subsequent units. The second step goes from the sampling population to the target population (e.g., from all students in the Elementary Psychology Subject Pool to all people who have the variables that you investigate). This requires something called ***external validity*****,** which is not a statistical issue.

Standard Symbols

The distinction between samples and populations is very important, as is the difference between **descriptive statistics** –which are summaries of samples– and **inferential statistics** –which provide estimations of populations. To help keep these separate, we use different symbols for each.

For samples and descriptive statistics, we use Roman letters. For example, we use the letter *n* for the number of observations. We use the letter *M* (or *m* or *mn*) for the mean of the sample, which is the most popular way to get an average value of a quantitative variable. We use the letter *r* for the correlation between two quantitative variables in a sample.

For populations, we use Greek letters, instead. The mean of the population –which is what we are usually trying to estimate– uses μ (mu, pronounced "mee-you" mashed together in one syllable) which is the Greek version of (lower-case) *m*. Likewise, the value of a correlation in the (entire)

population uses ρ (rho, pronounced “row”), which is a Greek (lower-case) *r*.

Certain upper-case Greek letters are used for mathematical operations. For example, if you want to say “add up the values of all of these numbers,” you can do this by using $\sum$ (“sig-muh”), which is an upper-case *S*, short for the word *sum*. Similarly, if you want to say “multiply all of these values together,” you can use $\prod$ (“pie”), which is an upper-case *P*, short for the word *product*.

Descriptive vs Inferential Statistics

Assume that you are interested in knowing the average hours of sleep that people are getting per night. As discussed above, you are not going to learn this by measuring the hours of sleep for every living person; you aren’t even going to measure this for every student in the Elementary Psychology Subject Pool. Instead, you will probably take a relatively small sample of students (e.g., 100 people), ask each of them how many hours of sleep they usually get, and then use these data to calculate an estimate of the average for everybody.

The process outlined above is best thought of as having three phases or steps: (1) collect the sample, (2) summarize the data in the sample, and then (3) use the summarized data to make an estimate of the entire population. The issues related to collecting the sample, such as how one ensures that the sample is representative, will not be discussed in these modules; they are not part of statistics. The second step is referred to as ***descriptive statistics*** and will be the focus of the units in this textbook. The third step is called ***inferential statistics*** and will covered in future materials.

The purpose and value of descriptive statistics is that they organize and summarize large sets of data. They allow

researchers to communicate quickly. Instead of having to list every value of every variable for every participant, the researcher can report a few numbers or draw a few pictures that capture most of the important information. Because they are summaries, they leave out some detail; but, because they are summaries, they can be very brief and efficient.

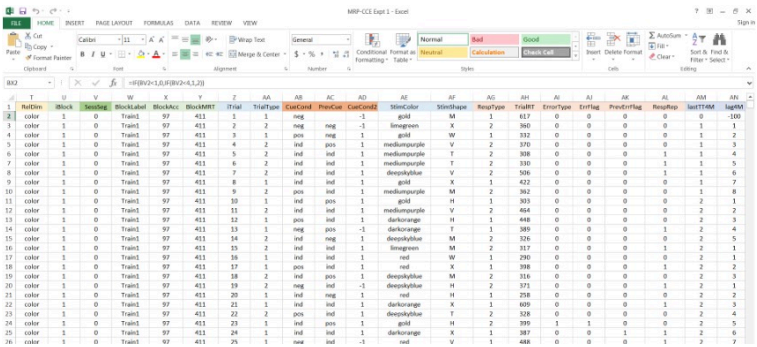
The main limitation of descriptive statistics is that they only describe (or make statements about) the actual sample. In other words, descriptive statistics never go beyond what we know for sure.

In contrast, inferential statistics allow us to go beyond the data in hand and calculate estimates of (or make statements about) the population from which the sample was taken. Although this is the ultimate goal of the research, it's important to note in advance that inferential statistics aren't guaranteed to be 100% accurate; they are educated guesses or estimations, not deductive conclusions. Whenever you use an inferential statistic, you need to keep track of how wrong you might be.

Unit 2. Managing Data

J TOBY MORDKOFF AND LEYRE CASTRO

Summary. This unit discusses the distinction between raw data and the pre-processed values that are used for the subsequent analysis. The various formats and rules with regard to managing and storing data are also reviewed.



Prerequisite Units

Unit 1. Introduction to Statistics for Psychological Science

Psychological Data

Psychology is an empirical science. This means that psychological theories are tested by comparing their

predictions to actual observations, which are often referred to as **data**. If the data match the predictions, the theory survives; if the data fail to confirm the predictions, the theory is falsified (which is a fancy way of saying *disproved*). Because of their crucial role, psychological data must be handled carefully and treated with respect. They must be organized and stored in a logical and standardized manner and format, allowing for their free exchange between researchers. One useful saying that summarizes empirical science is that "while researchers are encouraged to argue about theory, everyone must agree on the data."

In some (very rare) instances, the relevant data are quite simple, such as a single yes-or-no answer to one question that was asked of a sample of people. In such a case, a list of the *yes* or *no* responses would be sufficient. Usually, however, quite a lot of information is gathered from each participant. For example, besides an answer to a yes-or-no question, certain demographic information (e.g., age, sex,, etc.) might also be collected. Alternatively, maybe a series of questions are asked. Or the questions might require more complicated responses, which could be **numerical** (e.g., "on a scale of 1 to 10, how attractive do you find this person to be?") or **qualitative** (e.g., "what is your favorite color?"). Or maybe the participant is asked to perform a task many times, instead of just once, with multiple pieces of information (e.g., response time and response choice) recorded on every, separate trial.

The variety and potential complexity of psychological data raises at least two issues. One issue concerns the storage of data: should all of the data from every participant be stored together in one file or should the data from each participant be stored separately? Another issue concerns the distinction between what can be called "raw data" – which are the individual observations as they were originally collected – and "pre-processed data" (or "data after pre-processing") – which are the values that will be used for the formal analysis.

Pre-processing Data

Starting with the second issue, imagine an **experiment** in which participants must press the left button if the stimulus is blue and the right button if the stimulus is orange. The location of the stimulus is irrelevant – only the color matters – but sometimes the stimulus appears on the left and sometimes it appears on the right. (This task is based on work by J. Richard Simon of the University of Iowa, starting in the 1960s.) The participants are asked to respond as quickly as possible while making very few errors.

In such an experiment, it is typical for each of the four possible combinations of stimulus color and stimulus location to be used on a very large number of trials (e.g., 50 or more of each). Thus, the "raw data" from a single participant would be hundreds of sets of stimulus conditions (color and location) with multiple response values (which button was pressed and how long was required to make the response). But the analysis would not concern these individual sets of trial observations; instead, the raw data would be converted to a small number of summary scores (see Unit 1), such as average response time and percent correct for each of the four combinations of stimulus color and stimulus location.

Note that averaging is not the only change to the data that might occur during pre-processing. In the above example, the raw data were recorded in terms of stimulus color (blue vs orange), stimulus location (left vs right), response time (in milliseconds), and which response was made. During pre-processing, the stimulus locations might be re-coded in terms of whether the stimulus appeared near the correct response for the trial (which is often referred to as "congruent" or "compatible") or on the opposite side ("incongruent" or "incompatible"), because that is what most theories predict to be important. Likewise, the value of which response was made

would be converted to whether the response was correct. In other words, pre-processing in this case is doing two things: it is taking a huge number of pieces of raw data and reducing it down to just a few summary scores, and it is converting the data from the format of the events during the experiment to the format about which the theories make predictions.

Another example of pre-processing occurs when a long questionnaire is used to measure a small number of psychological constructs, such as measures of depression and anxiety. This often requires that participants provide ratings (e.g., on a 7-point scale) of how well each of many different statements (e.g., "I often feel sad" or "I'm often quite worried") applies to them. In this case, the raw data are the individual ratings for each of the questions, but what is needed for the analysis are the condensed scores for each of the small number of constructs. During pre-processing, each participant's answers to many questions are somehow combined to create these values, and then these are what are used for the actual analysis.

Pre-processing Data vs Analyzing Data

The key to the distinction between pre-processing data and subsequent analysis is best thought about in terms of *why* each is done. The purpose of pre-processing is to convert the raw data into the format for which the relevant theories make predictions. Pre-processing also simplifies matters by reducing the total number of pieces of data. Note that the theories are not yet being tested; the data are only being prepared for the subsequent test. For example, very few theories make predictions about

response times on individual trials; they make predictions about average response time, instead. If the theories did make predictions about individual trials, then the raw data would not be pre-processed; they'd be left as individual response times.

Likewise, few theories make predictions about the answers a participant might give to a single, specific question, such as "how often do you skip breakfast?"; they make predictions about the underlying psychological construct, such as depression. As above, pre-processing the numerous, separate answers into one or two measures of psychological constructs is not only simplifying the data by reducing the number of values to be analyzed, it is also converting the data into the form that matches the theory.

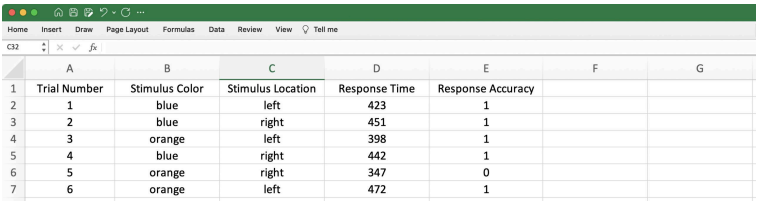
File Formats

In both examples in the previous section, pre-processing produces a small number of values from a very large amount of raw data. This brings us to the second issue raised above: how many files should be used? To answer this question, we need to discuss the various formats for files.

Psychological data are usually stored in large tables, often referred to as "spreadsheets," which can be viewed and edited using various software packages, such as Excel. (All of the pictures below are screen-shots of parts of Excel spreadsheets.) Although the details of these tables vary

considerably, they all obey one simple rule: each column in a spreadsheet always contains a single, specific piece of information, which does not change across rows, and each usually has a header (i.e., a special top row) that indicates what exactly is in every box in the column.

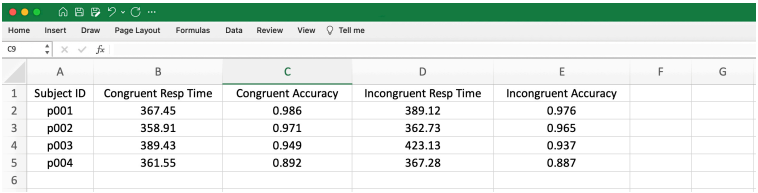
In the case of a response-time experiment, the raw data are usually produced by the software that runs the experiment, with each participant's data in a separate file. A complicated example of this was provided at the beginning of this unit; a much simpler example that matches the experiment from above is provided in Figure 2.1:



	A	B	C	D	E	F	G
1	Trial Number	Stimulus Color	Stimulus Location	Response Time	Response Accuracy		
2	1	blue	left	423	1		
3	2	blue	right	451	1		
4	3	orange	left	398	1		
5	4	blue	right	442	1		
6	5	orange	right	347	0		
7	6	orange	left	472	1		

Figure 2.1. Part of a data file containing raw data.

Note how each column has a label value in the first row, while each subsequent row contains the information related to a single trial. You know this, because the first column tells you how the rows are defines. It is standard good practice to do this: have the first column in the spreadsheet specify what is contained in each row. In general, the raw data from a response-time experiment will use this format. From these data, the summary values would be calculated – repeated for each participant, separately – and then these calculated values would placed in a new file that uses a different format:



	A	B	C	D	E	F	G
1	Subject ID	Congruent Resp Time	Congruent Accuracy	Incongruent Resp Time	Incongruent Accuracy		
2	p001	367.45	0.986	389.12	0.976		
3	p002	358.91	0.971	362.73	0.965		
4	p003	389.43	0.949	423.13	0.937		
5	p004	361.55	0.892	367.28	0.887		
6							

Figure 2.2. Data file with summary values for each participant.

As was true the raw-data files, each column in this sheet contains a specific piece of information, as indicated by the header row at the top. In contrast to the separate files for each participant, in which each subsequent row was a trial, in this file, each row holds the data for a particular participant. This is the standard format for the data that will be used for the actual analysis: each participant gets a row in the spreadsheet.

Wide vs Long Format

The technical label for a file that places all of the values for each participant on one (and only one) row is "wide format," because these spreadsheets can often have a very large number of columns. This is the standard in psychology. An alternative format uses multiple rows for each participant, with some columns being used to indicate the condition(s) under which the subset of the data were collected. This is known as "long format." Here is an example using the same values as previously, but now in long format:

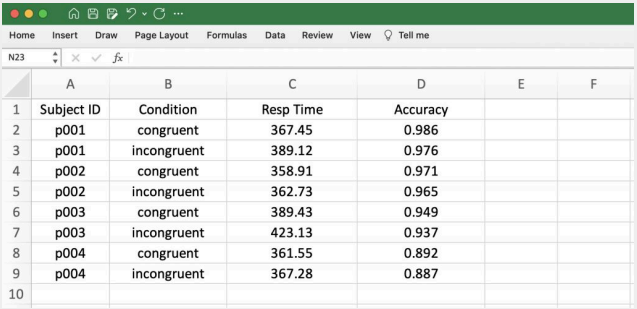


Figure 2.3. Data file with the same values as in Figure 2.2., but in a long format.

Long format is rarely used and the software packages that still employ this format for certain analyses usually have a built-in procedure for converting one format to the other. For this reason, and to maintain consistency, psychological data are almost always stored in wide format.

In contrast to response-time experiments, in which each participant performs hundreds of trials and separate files are used for each participant's raw data, all of the data from a questionnaire study is usually stored in one file. These files use the wide format: each item on the questionnaire gets a separate column, each participant get one row, and the first row in the file provides the label for each of the columns. When condensed scores are calculated (e.g., a measure of depression) based on the raw data in several columns, these can be added to the same file as a new, separate column, or placed in a new file that holds only the values that are needed for the subsequent analysis.

Data Storage

Before anything else, here is the cardinal rule: *never throw any information away*; keep all of the raw data, even if you don't currently have a need or use for them. There are two main reasons for this: first, you might discover a use for these data later; second, someone else might already have a use for them. (It might also appear a bit suspicious if you collect certain data and then delete them, making them unavailable to others who might ask about them.)

This rule applies to all of the data as originally collected. If, for example, you omit a few trials of a response-time experiment during pre-processing, maybe because the response was abnormally fast or slow (i.e., an outlier), you do not delete the trial from the raw-data file; you merely skip over it during pre-processing. Likewise, if you decide to omit all of the data from a certain participant, maybe because they did not appear to follow instructions, you still keep the file that holds their raw data; you just don't include their values in the file for subsequent analysis.

In situations where you have large amounts of raw data, such as response-time experiments or long questionnaires, the files that contain the raw data can stored separately from the single, small file that holds the pre-processed values that are ready for analysis. As discussed above, the raw-data files might use a format that is different from the final file – that is fine, as long as the header row in every file makes it clear what's contained in each column. If condensed scores were added to a raw-data file, as is often done with questionnaire data, you can save two versions, if you wish: one with everything and another with only the final values that are needed for the analysis – that, also, is fine, as long as keep all of the raw data somewhere.

Unit 3. Descriptive Statistics for Psychological Research

J TOBY MORDKOFF AND LEYRE CASTRO

Summary. This unit briefly reviews the distinction between descriptive and inferential statistics and then discusses the ways in which both numerical and categorical data are usually summarized for psychological research. Different measures of center and spread, and when to use them, are explained. The shape of the data is also discussed.

Prerequisite Units
Unit 1. Introduction to Statistics for Psychological Science
Unit 2. Managing Data

Introduction

Assume that you are interested in some attribute or characteristic of a very large number of people, such as the average hours of sleep per night for all undergraduates at all universities. Clearly, you are not going to do this by measuring the hours of sleep for every student, as that would be difficult

to impossible. So, instead, you will probably take a relatively small **sample** of students (e.g., 100 people), ask each of them how many hours of sleep they usually get, and then use these data to estimate the average for all undergraduates.

The process outlined above can be thought of as having three phases or steps: (1) collect a sample, (2) summarize the data in the sample, and (3) use the summarized data to make the estimate of the entire population. The issues related to collecting the sample, such as how one ensures that the sample is representative of the entire population will not be discussed here. Likewise, the way that one uses the summary of a sample to calculate an estimate of the population will not be explained here. This unit will focus on the second step: the way in which psychologists summarize data.

The general label for procedures that summarize data is *descriptive statistics*. This can be contrasted with procedures that make estimates of population values, which are known as *inferential statistics*. Thus, descriptive and inferential statistics each give different insights into the nature of the data gathered. Descriptive statistics describe the data so that the big picture can be seen. How? By organizing and summarizing properties of a data set. Calculating descriptive statistics takes unordered observations and logically organizes them in some way. This allow us to describe the data obtained, but it does not make conclusions beyond the sample. This is important, because part of conducting (good) research is being able to communicate your findings to other people, and descriptive statistics will allow you to do this quickly, clearly, and precisely.

To prepare you for what follows, please note two things in advance. First, there are several different ways that we can summarize a large set of data. Most of all: we can use numbers or we can use graphical representations. Furthermore, when the data are numerical, we will have options for several of the summary values that we need to calculate. This may seem

confusing at first; hopefully, it soon will make sense. Second, but related to the first, the available options for summarizing data often depend on the type of data that we have collected. For example, numerical data, such as hours of sleep per night, are summarized differently from categorical data, such as favorite flavors of ice-cream.

The key to preventing this from becoming confusing is to keep the function of descriptive statistics in mind: we are trying to summarize a large amount of data in a way that can be communicated quickly, clearly, and precisely. In some cases, a few numbers will do the trick; in other cases, you will need to create a plot of the data.

This unit will only discuss the ways in which a single set of values are summarized. When you collect more than one piece of information from every participant in the sample –e.g., you not only ask them how many hours of sleep they usually get, but also ask them for their favorite flavor of ice-cream– then you can do three things using descriptive statistics: summarized the first set of values (on their own), summarize the second set of values (on their own), and summarize the relationship between the two sets of values. This unit only covers the first two of these three. Different ways to summarize the relationship between two sets of values will be covered in Units 7 and 8.

Summarizing Numerical Data

The most-popular way to summarize a set of numerical data –e.g., hours of sleep per night– is in terms of two or three aspects. One always includes values for the ***center*** of the data and the ***spread*** of the data; in some cases, the ***shape*** of the data is also described. A measure of center is a single value that attempts to describe an entire set of data by identifying

the central position within that set of data. The full, formal label for this descriptive statistic is *measure of central tendency*, but most people simply say "center." Another label for this is the "average."

A measure of spread is also a single number, but this one indicates how widely the data are distributed around their center. Another way of saying this is to talk about the "variability" of the data. If all of the individual pieces of data are located close to the center, then the value of spread will be low; if the data are widely distributed, then the value of spread will be high.

What makes this a little bit complicated is that there are multiple ways to mathematically define the center and spread of a set of data. For example, both the mean and the median (discussed in detail below) are valid measures of central tendency. Similarly, both the variance (or standard deviation) and the inter-quartile range (also discussed below) are valid measures of spread. This might suggest that there are at least four combinations of center and spread (i.e., two versions of center crossed with two version of spread), but that isn't true. The standard measures of center and spread actually come in pairs, such that your choice with regard to one forces you to use a particular option for the other. If you define the center as the mean, for example, then you have to use variance (or standard deviation) for spread; if you define the center as the median, then you have to use the inter-quartile range for spread. Because of this dependency, in what follows we shall discuss the standard measures of center and spread in pairs. When this is finished, we shall mention some of the less popular alternatives and then, finally, turn to the issue of shape.

Measures of Centerand Spread Based on Moments

The mean and variance of a set of numerical values are (technically) the first and second moments of the set of data. Although it is not used very often in psychology, the term "moment" is quite popular in physics, where the first moment is the center of mass and the second moment is rotational inertia (these are very useful concepts when describing how hard it is to throw or spin something). The fact that the mean and variance of a set of numbers are the first and second moments isn't all that important; the key is that they are based on the same approach to the data, which is why they are one of the standard pairs of measures for describing a set of numerical data.

Mean

The mean is the most popular and well known measure of central tendency. It is what most people intend when they use the word "average." The mean can be calculated for any set of numerical data, discrete or continuous, regardless of units or details. The mean is equal to the sum of all values divided by the number of values. So, if we have ***n*** values in a data set and they have values x_1, x_2, ..., x_n, the mean is calculated using the following formula:

$$\bar{X} = \frac{\sum X_i}{n}$$

where $\sum$ is the technical way of writing "add up all of the X values" (i.e., the upper-case, Greek letter $\sum$ [sigma] tells you to calculate the sum of what follows), and *n* is the number of pieces of data (which is usually referred to as "sample size").

The short-hand for writing the *mean of X* is $\bar{X}$ (i.e., you put a bar over the symbol, X in this case; it is pronounced "ex bar"). As a simple example, if the values of X are 2, 4, and 7, then = 13, n = 3, and therefore $\bar{X}$ = 4.33 (after rounding to two decimal places).

Before moving forward, note two things about using the mean as the measure of center. First, the mean is rarely one of the actual values from the original set of data. As an extreme example: when the data are **discrete** (e.g., whole numbers, like the number of siblings), the mean will almost never match any of the specific values, because the mean will almost never be a whole number as well.

Second, an important property of the mean is that it includes and depends on every value in your set of data. If any value in the data set is changed, then the mean will change. In other words, the mean is "sensitive" to all of the data.

Variance and Standard Deviation

When the center is defined as the mean, the measure of spread to use is the *variance* (or the square-root of this value, which is the *standard deviation*). Variance is defined as the average of the squared deviations from the mean. The formula for variance is:

$$Variance\ of\ X = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

where $\bar{X}$ is the mean (see above). In words, you take each piece of data, subtract the mean, and square the result; do this for each of the ***n*** pieces of data and add up the results, then divide by one less than the number of pieces of data. More technically, to determine the variance of a set of scores, you have to 1) find the mean of the scores, 2) compute the deviation scores (the difference between each individual score and the

mean), 3) square each of the deviation scores, 4) add up all of the squared deviation scores, and 5) divide by one less than the number of scores. Thus, for example, the variance of 2, 4, and 7 (which have a mean of 4.33, see above) is:

$$(2 - 4.33)^2 + (4 - 4.33)^2 + (7 - 4.33)^2 = 12.6667$$

$$\text{and then divide by } (3-1) \rightarrow 6.33$$

Note that, because each sub-step of the summation involves a value that has been squared, the value of variance cannot be a negative number. Note, also, that when all of the individual pieces of data are the same, they will all be equal to the mean, so you will be adding up numbers that are all zero, so variance will also be zero. These both make sense, because here we are calculating a measure of how spread out the data are, which will be zero when all of the data are the same and cannot be less than this.

> Technically, the value being calculated here is the ***sample variance***, which is different from something known as the *population variance*. The former is used when we have taken a **sample**; the latter is used when we have measured every possible value in the entire **population**. Since we never measure every possible value when doing psychological research, we do not need the formula for population variance and can simply refer to the sample variance as *variance*.

As mentioned above, some people prefer to express this measure of spread in terms of the square-root of the variance, which is the standard deviation. The main reason for doing this is because the units of variance are the square of the units of the original data, whereas the units of standard deviation are the same as the units of the original data. Thus, for example, if

you have response times of 2, 4, and 7 seconds, which have a mean of 4.33 seconds, then the variance is 6.33 seconds^2 (which is difficult to conceptualize), but also have a standard deviation of 2.52 seconds (which is easy to think about).

Conceptually, you can think of the standard deviation as the typical distance of any score from the mean. In other words, the standard deviation represents the ***standard*** amount by which individual scores deviate from the mean. The standard deviation uses the mean of the data as a baseline or reference point, and measures variability by considering the distance between each score and the mean.

Note that similar to the mean, both the variance and the standard deviation are sensitive to every value in the set of data; if any one piece of data is changed, then not only will the mean change, but the variance and standard deviation will also be changed.

Practice

Let's calculate now the mean and the standard deviation of the two variables in the following dataset containing the number of study hours before an exam (X, Hours), and the grade obtained in that exam (Y, Grade), for 15 participants. To calculate the mean of Hours, $\bar{X}$, we sum all of the values for Hours, and divide by the total number of values, 15. To calculate the mean of Grade, $\bar{Y}$, we sum all of the values for Grade,

and divide by the total number of values, 15. Doing so, we obtain the mean for Hours ($\bar{X} = 13.66$) and the mean for Grade ($\bar{Y} = 86.466$). Table 3.1. shows each of the scores, and the deviation scores for each X and Y score. The deviation scores, as explained above, are calculated by subtracting the mean from each of the individual scores.

Participant	Hours	Grade	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$
P1	8	78	(8 – 13.66) = -5.66	(78 – 86.46) = -8.46
P2	11	80	(11 – 13.66) = -2.66	(80 – 86.46) = -6.46
P3	16	89	(16 – 13.66) = 2.34	(89 – 86.46) = 2.54
P4	14	85	(14 – 13.66) = 0.34	(85 – 86.46) = -1.46
P5	12	84	(12 – 13.66) = -1.66	(84 – 86.46) = -2.46
P6	15	86	(15 – 13.66) = 1.34	(86 – 86.46) = -0.46
P7	18	95	(18 – 13.66) = 4.34	(95 – 86.46) = 8.54
P8	20	96	(20 – 13.66) = 6.34	(96 – 86.46) = 9.54
P9	10	83	(10 – 13.66) = -3.66	(83 – 86.46) = -3.46
P10	9	81	(9 – 13.66) = -4.66	(81 – 86.46) = -5.46
P11	16	93	(16 – 13.66) = 2.34	(93 – 86.46) = 6.54
P12	17	92	(17 – 13.66) = 3.34	(92 – 86.46) = 5.54
P13	13	84	(13 – 13.66) = -0.66	(84 – 86.46) = -2.46
P14	12	83	(12 – 13.66) = -1.66	(83 – 86.46) = -3.46

Table 3.1. Number of study hours before an exam (X, Hours), and the grade obtained in that exam (Y, Grade) for 15 participants. The two most right columns show the deviation scores for each X and Y score.

Once we have the deviation scores for each participant, we square each of the deviation scores, and sum them.

For Hours:

$$(-5.66)^2 + (-2.66)^2 + (2.34)^2 + (0.34)^2 + (-1.66)^2 + (1.34)^2 + (4.34)^2 + (6.34)^2 + (-3.66)^2 + (-4.66)^2 + \ldots$$

$$\ldots(2.34)^2 + (3.34)^2 + (-0.66)^2 + (-1.66)^2 + (0.34)^2 = 166.334$$

We then divide that sum by one less than the number of scores, 15 – 1 in this case:

$$166.334/14 = 11.66$$

So, 11.66 is the variance for the number of hours in our sample of participants.

In order to obtain the standard deviation, we calculate the square root of the variance:

$$\sqrt{11.66} = 3.42$$

We follow the same steps to calculate the standard deviation of our participants' grade. First, we square each of the deviation scores (most right column in Table 3.1), and sum them:

$$(-8.46)^2 + (-6.46)^2 + (2.54)^2 + (-1.46)^2 + (-2.46)^2 + (-0.46)^2 + (8.54)^2 + (9.54)^2 + (-3.46)^2 + \ldots$$

$$\ldots (-5.46)^2 + (6.54)^2 + (5.54)^2 + (-2.46)^2 + (-3.46)^2 + (1.54)^2 = 427.734$$

Next, we divide that sum by one less than the number of scores, 14:

$$427.734/14 = 30.55$$

So, 30.55 is the variance for the grade in our sample of participants.

In order to obtain the standard deviation, we calculate the square root of the variance:

$$\sqrt{30.55} = 5.53$$

Thus, you can **summarize the data** in our sample saying that the **mean hours of study time are 13.66**, with a **standard deviation of 3.42**, whereas the **mean grade is 86.46**, with a **standard deviation of 5.53.**

Measures of Centerand Spread Based on Percentiles

The second pair of measures for center and spread are based on percentile ranks and percentile values, instead of moments. In general, the percentile rank for a given value is the percent of the data that is smaller (i.e., lower in value). As a simple example, if the data are 2, 4, and 7, then the percentile rank for 5 is 67%, because two of the three values are smaller than 5. Percentile ranks are usually easy to calculate. In contrast, a percentile value (which is kind of the "opposite" of a percentile rank) is much more complicated. For example, the percentile value for 67% when the data are 2, 4, and 7 is something between 4 and 7, because any value between 4 and 7 would be larger than two-thirds of the data. (FYI: the percentile value is this case is 5.02.) Fortunately, we won't need to worry about the details when calculating that standard measures of center and spread when using the percentile-based method.

Median

The median –which is how the percentile-based method defines *center*– is best thought of the middle score when the data have been arranged in order of magnitude. To see how this can be done by hand, assume that we start with the data below:

65 54 79 57 35 14 56 55 77 45 92

We first re-arrange these data from smallest to largest:

14 35 45 54 55 **56** 57 65 77 79 92

The median is the middle of this new set of scores; in this case, the value (in blue) is 56. This is the middle value because there are 5 scores lower than it and 5 scores higher than it. Finding the median is very easy when you have an odd number of scores.

What happens when you have an even number of scores? What if you had only 10 scores, instead of 11? In this case, you take the middle two scores, and calculate the mean of them. So, if we start with the following data (which are the same as above, with the last one omitted):

65 54 79 57 35 14 56 55 77 45

We again re-arrange that data from smallest to largest:

14 35 45 54 **55** **56** 57 65 77 79

And then calculate the mean of the 5th and 6th values (tied for the middle, in blue) to get a median of 55.50.

In general, the median is the value that splits the entire set of data into two equal halves. Because of this, the other name for

the median is *50th percentile* –50% of the data are below this value and 50% of the data are above this value. This makes the median a reasonable alternative definition of center.

Inter-Quartile Range

The inter-quartile range (typically named using its initials, IQR) is the measure of spread that is paired with the median as the measure of center. As the name suggests, the IQR divides the data into four sub-sets, instead of just two: the bottom quarter, the next higher quarter, the next higher quarter, and the top quarter (the same as for the median, you must start by re-arranging the data from smallest to largest). As described above, the median is the dividing line between the middle two quarters. The IQR is the distance between the dividing line between the bottom two quarters and the dividing line between the top two quarters.

Technically, the IQR is the distance between the 25th percentile and the 75th percentile. You calculate the value for which 25% of the data is below this point, then you calculate the value for which 25% of the data is above this point, and then you subtract the first from the second. Because the 75th percentile cannot be lower than the 25th percentile (and is almost always much higher), the value for IQR cannot be negative number.

Returning to our example set of 11 values, for which the median was 56, the way that you can calculate the IQR by hand is as follows. First, focus only on those values that are to the left of (i.e., lower than) the middle value:

14 35 45 54 55 56 57 65 77 79 92

Then calculate the "median" of these values. In this case, the

answer is 45, because the third box is the middle of these five boxes. Therefore, the 25th percentile is 45.

Next, focus on the values that are to the right of (i.e., higher than) the original median:

14 35 45 54 55 56 **57 65 77 79 92**

The middle of these values, which is 77, is the 75th percentile. Therefore, the IQR for these data is 32, because 77 – 45 = 32. Note how, when the original set of data has an odd number of values (which made it easy to find the median), the middle value in the data set was ignored when finding the 25th and 75th percentiles. In the above example, the number of values to be examined in each subsequent step was also odd (i.e., 5 each), so we selected the middle value of each subset to get the 25th and 75th percentiles.

If the number of values to be examined in each subsequent step had been even (e.g., if we had started with 9 values, so that 4 values would be used to get the 25th percentile), then the same averaging rule as we use for median would be used: use the average of the two values that tie for being in the middle. For example, if these are the data (which are the first nine values from the original example after being sorted):

14 **35 45** 54 **55** 56 **57 65** 77

The median (in blue) is 55, the 25th percentile (the average of the two values in green) is 40, and the 75th percentile (the average of the two values in red) is 61. Therefore, the IQR for these data is 61 – 40 = 21.

A similar procedure is used when you start with an even number of values, but with a few extra complications (these complications are caused by the particular method of calculating percentiles that is typically used in the psychology). The first change to the procedure for calculating IQR is that

now every value is included in one of the two sub-steps for getting the 25th and 75th percentile; none are omitted. For example, if we use the same set of 10 values from above (i.e., the original 11 values with the highest omitted), for which the median was 55.50, then here is what we would use in the first sub-step:

14 35 45 54 55 56 57 65 77 79

In this case, the 25th percentile will be calculated from an odd number of values (5). We start in the same way before, with the middle of these values (in green), which is 45. Then we adjust it by moving the score 25% of the distance towards next lower value, which is 35. The distance between these two values is 2.50 –i.e., (45 – 35) x .25 = 2.50– so the final value for the 25th percentile is 42.50.

The same thing is done for 75th percentile. This time we would start with:

14 35 45 54 55 **56 57 65 77 79**

The starting value (in red) of 65 would then be moved 25% of the distance towards the next higher, which is 77, producing a 75th percentile of 68 –i.e., 65 + ((77 – 65) x .25) = 68. Note how we moved the value away from the median in both cases. If we don't do this –if we used the same simple method as we used when the original set of data had an odd number of values– then we would slightly under-estimate the value of IQR.

Finally, if we start with an even number of pieces of data and also have an even number for each of the sub-steps (e.g., we started with 8 values), then we again have to apply the correction. Whether you have to shift the 25th and 75th percentiles depends on original number of pieces of data, not the number that are used for the subsequent sub-steps. To

demonstrate this, here are the first eight values from the original set of data:

14 **35** **45** **54** 55 56 57 65

The first step to calculating the 25th percentile is to average the two values (in green) that tied for being in the middle of the lower half of the data; the answer is 40. Then, as above, move this value 25% of the distance away from the median –i.e., move it down by 2.50, because (45 – 35) x .25 = 2.50. The final value is 37.50.

Then do the same for the upper half of the data:

14 35 45 54 **55** **56** **57** **65**

Start with the average of the two values (in red) that tied for being in the middle and then shift this value 25% of their difference away from the center. The mean of the two values is 56.50 and after shifting the 75th percentile is 56.75. Thus, the IQR for these eight pieces of data is 56.75 – 37.50 = 19.25.

Note the following about the median and IQR: because these are both based on percentiles, they are not always sensitive to every value in the set of data. Look again at the original set of 11 values used in the examples. Now imagine that the first (lowest) value was 4, instead of 14. Would either the median or the IQR change? The answer is *No, neither would change*. Now imagine that the last (highest) value was 420, instead of 92. Would either the median or IQR change? Again, the answer is *No*.

Some of the other values can also change without altering the median and/or IQR, but not all of them. If you changed the 56 in the original set to being 50, instead, for example, then the median would drop from 56 to 55, but the IQR would remain 32. In contrast, if you only changed the 45 to being a 50,

then the IQR would drop from 32 to 27, but the median would remain 56.

The one thing that is highly consistent is how you can decrease the lowest value and/or increase the highest value without changing either the median or IQR (as long as you start with at least 5 pieces of data). This is an important property of percentiles-based methods: they are relatively insensitive to the most extreme values. This is quite different from moments-based methods; the mean and variance of a set of data are both sensitive to every value.

Other Measures of Center and Spread

Although a vast majority of psychologists use either the mean and variance (as a pair) or the median and IQR (as a pair) as their measures of center and spread, occasionally you might come across a few other options.

Mode

The mode is a (rarely-used) way of defining the center of a set of data. The mode is simply the value that appears the most often in a set of data. For example, if your data are 2, 3, 3, 4, 5, and 9, then the mode is 3 because there are two 3s in the data and no other value appears more than once. When you think about other sets of example data, you will probably see why the mode is not very popular. First, many sets of data do not have a meaningful mode. For the set of 2, 4, and 7, all three different values appear once each, so no value is more frequent than any other value. When the data are continuous and measured precisely (e.g., response time in milliseconds), then this problem will happen quite often. Now consider the

set of 2, 3, 3, 4, 5, 5, 7, and 9; these data have two modes: 3 and 5. This also happens quite often, especially when the data are discrete, such as when they must all be whole numbers.

But the greatest problem with using the mode as the measure of center is that it is often at one of the extremes, instead of being anywhere near the middle. Here is a favorite example (even if it is not from psychology): the amount of federal income tax paid. The most-frequent value for this –i.e., the mode of federal income tax paid– is zero. This also happens to be the same as the lowest value. In contrast, in 2021, for example, the mean amount of federal income tax paid was a little bit over $10,000.

Range

Another descriptive statistic that you might come across is the range of the data. Sometimes this is given as the lowest and highest values –e.g., "the participant ages ranged from 18 to 24 years"– which provides some information about center and spread simultaneously. Other times the range is more specifically intended as only a measure of spread, so the difference between the highest and lowest values is given –e.g., "the average age was 21 years with a range of 6 years." There is nothing inherently wrong with providing the range, but it is probably best used as a supplement to one of the pairs of measures for center and spread. This is true because range (in either format) often fails to provide sufficient detail. For example, the set of 18, 18, 18, 18, and 24 and the set of 18, 24, 24, 24, and 24 both range from 18 to 24 (or have a range of 6), even though the data sets are clearly quite different.

Choosing the Measures of Center and

Spread

When it comes to deciding which measures to use for center and spread when describing a set of numerical data –which is almost always a choice between mean and variance (or standard deviation) or median and IQR– the first thing to keep in mind is that this is not a question of "which is better?"; it is a question of **which is more appropriate** for the situation. That is, the mean and the median are not just alternative ways of calculating a value for the center of a set of data; they use different definitions of the meaning of center.

So how should you make this decision? One factor that you should consider focuses on a key difference between moments and percentiles that was mentioned above: how the mean and variance of a set of data both depend on every value, whereas the median and IQR are often unaffected by the specific values at the upper and lower extremes. Therefore, if you believe that every value in the set of data is equally important and equally representative of whatever is being studied, then you should probably use the mean and variance for your descriptive statistics. In contrast, if you believe that some extreme values might be **outliers** (e.g., the participant wasn't taking the study very seriously or was making random fast guesses), then you might want to use the median and IQR instead.

Another related factor to consider is the **shape** of the distribution of values in the set of data. If the values are spread around the center in a roughly symmetrical manner, then the mean and the median will be very similar, but if there are more extreme values in one **tail** of the distribution (e.g., there are more extreme values above the middle than below), this will pull the mean away from the median, and the latter might better match what you think of as the center.

Finally, if you are calculating descriptive statistics as part of a process that will later involve making inferences about the population from which the sample was taken, you might want

to consider the type of statistics that you will be using later. Many **inferential statistics** (including *t*-tests, ANOVA, and the standard form of the correlation coefficient) are based on moments so, if you plan to use these later, it would be probably more appropriate to summarize the data in terms of mean and variance (or standard deviation). Other statistics (including sign tests and alternative forms of the correlation coefficient) are based on percentiles, so if you plan to use these instead, then the median and IQR might be more appropriate for the descriptive statistics.

Hybrid Methods

Although relatively rare, there is one alternative to making a firm decision between moments (i.e., mean and variance) and percentiles (i.e., median and IQR) –namely, hybrid methods. One example of this is as follows. First, sort the data from smallest to largest (in the same manner as when using percentiles). Then remove a certain number of values from the beginning and end of the list. The most popular version of this is to remove the lowest 2.5% and the highest 2.5% of the data; for example, if you started with 200 pieces of data, remove the first 5 and the last 5, keeping the middle 190. Then switch methods and calculate the mean and variance of the retained data. This method is trying to have the best of both worlds: it is avoiding outliers by removing the extreme values, but it is remaining sensitive to all the data that are being retained. When this method is used, the correct label for the final two values are the "trimmed mean" and "trimmed variance."

Measures of Shape for Numerical Data

As the name suggests, the shape of a set of data is best thought about in terms of how the data would look if you made some sort of figure or plot of the values. The most popular way to make a plot of a single set of numerical values starts by putting all of the data into something that is called a **frequency table**. In brief, a frequency table is a list of all possible values, along with how many times each value occurs in the set of data. This is easy to create when there are not very many different values (e.g., number of siblings); it becomes more complicated when almost every value in the set of data is unique (e.g., response time in milliseconds).

The key to resolving the problem of having too many unique values is to "bin" the data. To bin a set of data, you choose a set of equally-spaced cut-offs, which will determine the borders of adjacent bins. For example, if you are working with response times which happen to range from about 300 to 600 milliseconds (with every specific value being unique), you might decide to use bins that are 50 milliseconds wide, such that all values from 301 to 350 go in the first bin, all values from 351 to 400 go in the second bin, etc. Most spreadsheet-based software packages (e.g., Excel) have built-in procedures to do this for you.

As an illustration of this process, let's go back to the set of 11 values we have used in previous examples:

65 55 79 56 35 14 56 55 77 45 92

Based on the total number of values and their range, we decide to use bins that are 20 units wide. Here are the same data in a frequency table:

Bin	1-20	21-40	41-60	61-80
Frequency	1	1	5	3

Once you have a list of values or bins and the number of pieces of data in each, you can make a frequency histogram of the data, as shown in Figure 3.1:

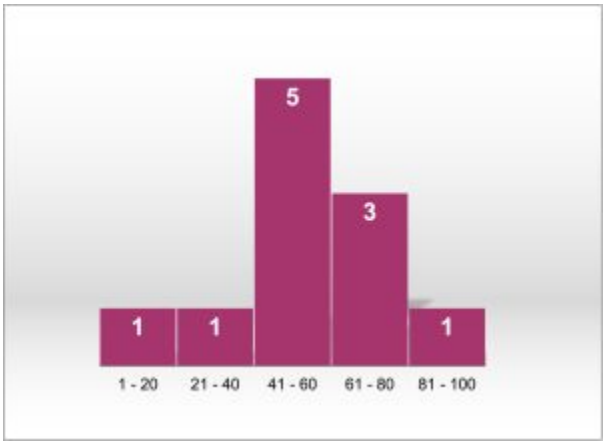


Figure 3.1. Example of a histogram in which the data are grouped into five bins. The numbers inside the bars represent the frequency count, that is, how many data points we have, within each bin.

Based on this histogram, we can start to make descriptive statements about the shape of the data. In general, these will concern two aspects, known as *skewness* and *kurtosis*, as we shall see next.

Skewness

Skewness refers to the lack of symmetry. It the left and right

sides of the plot are mirror images of each other, then the distribution has no skew, because it is symmetrical; this is the case of the normal distribution (see Figure 3.2). This clearly is not true for the example in Figure 3.1. If the distribution has a longer **tail** on the left side, as is true here, then the data are said to have **negative skew**. If the distribution has a longer "tail" on the right, then the distribution is said to have **positive skew**. Note that you need to focus on the skinny part of each end of the plot. The example in Figure 3.1 might appear to be heavier on the right, but skew is determined by the length of the skinny tails, which is clearly much longer on the left. As a reference, Figure 3.2. shows you a normal distribution, perfectly symmetrical, so its skewness is zero; to the left and to the right, you can see two skewed distributions, positive and negative. Most of the data points in the distribution with a positive skew have low values, and has a long tail on its right side. The opposite is true for the distribution with negative skew: most of its data points have high values, and has a long tail on its left side.

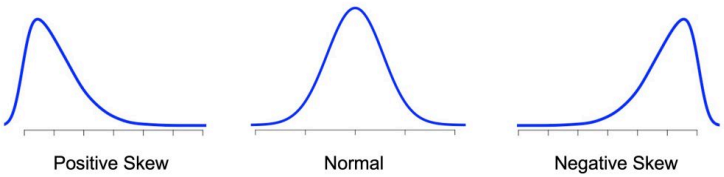


Figure 3.2. An illustration of skewness. A normal distribution (in the middle), is symmetrical, so it has no skew. The distributions with positive and negative skew show a clear lack of symmetry.

Kurtosis

The other aspect of shape, kurtosis, is a bit more complicated. In general, kurtosis refers to how sharply the data are peaked,

and is established in reference to a baseline or standard shape, the normal distribution, that has kurtosis zero. When we have a nearly flat distribution, for example when every value occurs equally often, the kurtosis is negative. When the distribution is very pointy, the kurtosis is positive.

If the shape of your data looks like a bell curve, then it's said to be **mesokurtic** ("meso" means *middle or intermediate* in Greek). If the shape of your data is flatter than this, then it's said to be **platykurtic** ("platy" means *flat* in Greek). If your shape is more pointed from this, then your data are **leptokurtic** ("lepto" means *thin, narrow,* or *pointed* in Greek). Examples of these shapes can be seen in Figure 3.3.

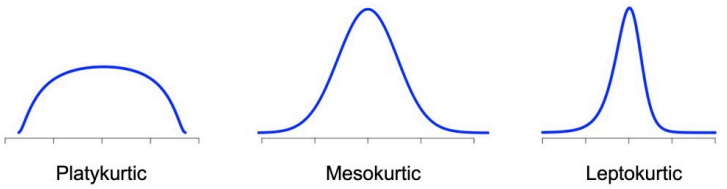


Figure 3.3. An illustration of kurtosis. A normal distribution (in the middle) is mesokurtic, and its kurtosis value is zero. The platykurtic distribution, on the left, is flatter than the normal distribution (negative kurtosis), whereas the leptokurtic distribution, on the right, is more pointed than the normal distribution (positive kurtosis).

Both skew and kurtosis can vary a lot; these two attributes of shape are not completely independent. That is, it is impossible for a perfectly flat distribution to have any skew; it is also impossible for a highly-skewed distribution to have zero kurtosis. A large proportion of the data that is collected by psychologists is approximately normal, but with a long right tail. In this situation, a good verbal label for the overall shape could be *positively-skewed normal,* even if that seems a bit contradictory, because the true normal distribution is actually

symmetrical (see Figures 3.2 and 3.3). The goal is to summarize the shape in a way that is easy to understand while being as accurate as possible. You can always show a picture of your distribution to your audience. A simple summary of the shape of the histogram in Figure 3.1 could be: *roughly normal, but with a lot of negative skew*; this tells your audience that the data have a decent-sized peak in the middle, but the lower tail is a lot longer than the upper tail.

Numerical Values for Skew and Kurtosis

In some rare situations, you might want to be even more precise about the shape of a set of data. Assuming that you used the mean and variance as your measures of center and spread, in these cases, you can use some (complicated) formulae to calculate specific numerical values for skew and kurtosis. These are the third and fourth moments of the distribution (which is why they can only be used with the mean and variance, because those are the first and second moments of the data). The details of these measures are beyond this course, but to give you an idea, as indicated above, values that depart from zero tells you that the shape is different from the normal distribution. A value of skew that is less than –1 or greater than +1 implies that the shape is notably skewed, whereas a value of kurtosis that is more than 1 unit away from zero imply that the data are not mesokurtic.

Summarizing Categorical Data

By definition, you cannnot summarize a set of categorical data (e.g., favorite colors) in terms of a numerical mean and/or a numerical spread. It also does not make much sense to talk about shape, because this would depend on the order in which you placed the options on the X-axis of the plot. Therefore, in this situation, we usually make a frequency table (with the options in any order that we wish). You can also make a frequency histogram, but be careful not to read anything important into the apparent shape, because changing the order of the options would completely alter the shape.

An issue worth mentioning here is something that is similar to the process of binning. Assume, for example, that you have taken a sample of 100 undergraduates, asking each for their favorite genre of music. Assume that a majority of the respondents chose either pop (24), hip-hop (27), rock (25), or classical (16), but a few chose techno (3), trance (2), or country (3). In this situation, you might want to combine all of the rare responses into one category with the label *Other*. The reason for doing this is that it is difficult to come to any clear conclusions when something is rare. As a general rule, if a category contains fewer than 5% of the observations, then it should probably be combined with one or more other options. An example frequency table for such data is this:

Choice	**Pop**	**Hip-Hop**	**Rock**	**Classical**	**Other**
Frequency	24	27	25	16	8

Finally, to be technically accurate, it should be mentioned that there are some ways to quantify whether each of the options is being selected the same percent of the time, including the *Chi-square* (pronounced “kai-squared”) test and relative entropy

(which comes from physics), but these are not very usual. In general, most researchers just make a table and/or maybe a histogram to show the distribution of the categorical values.

Unit 4. Descriptive Statistics with Excel

J TOBY MORDKOFF

Summary. This unit shows you, in a series of short videos, how to calculate descriptive statistics and make simple plots of numerical data using Microsoft Excel.

Prerequisite Units

Unit 1. Introduction to Statistics for Psychological Science
Unit 2. Managing Data
Unit 3. Descriptive Statistics for Psychological Research

Mean and Sample Standard Deviation

The most popular way to define the center and spread of a single set of numerical data is the mean and standard deviation. You can also use the variance for spread, instead of the standard deviation, but then the units of measure are squared, which can be confusing (e.g., what is a squared millisecond?). In any event, we will be calculating the variance on the way to the standard deviation, so both of these will be available.

Here are the formulae that we will use (see Unit 3):

(1)
$$Mean\ of\ X = \bar{X} = \frac{\sum X_i}{n}$$

(2)
$$Sample\ Variance\ of\ X = s_x^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

(3)

$$Sample\ Standard\ Deviation\ of\ X = s_x = \sqrt{s_x^2}$$

Note that we will be calculating the sample values of variance and standard deviation, so we're using the formula with *n* – 1 in the denominator. We will do the calculations on the following data (from Unit 3):

65 54 79 57 35 14 56 55 77 45 92

Watch the video: https://tinyurl.com/Mean-and-Standard-Deviation

Median and Inter-Quartile Range

The other way to define the center and spread of a single set of numerical data is the median and inter-quartile range (IQR). These are both based on percentiles. The median is 50th percentile; the IQR is the difference between the 75th and 25th percentiles.

Note that there are two different ways to calculate a percentile. One method *includes* the median (when calculating percentiles other than the 50th); the other method *excludes* the median. We use the exclusive version of percentiles, because the inclusive version has a tendency to under-estimate the IQR in the population from which the sample was taken. Warning: for many spreadsheets and stats packages, the exclusive version is not the default, so you have to be careful.

For the demonstration, we will use an extremely simple set of data: 1, 2, 3, 4, & 5.

Watch the video: https://tinyurl.com/Median-and-IQR

Numerical Values of Skew and Kurtosis

When center and spread are defined as the mean and standard deviation (or variance), there are parallel definitions of skewness and kurtosis that may be used for the shape of the data. These are also based on the method of moments. The formulae for these are quite complicated, so we will be jumping directly to Excel's built-in functions. Note that a perfect bell curve, known as the normal distribution, has skew = 0 and kurtosis = 0. A new set of (random) data will be used for this demonstration.

Watch the video: https://tinyurl.com/Skew-and-Kurtosis

Simple Frequency Plots (Histograms) of Numerical Data

The more general approach to skew and kurtosis is to make a plot of the data and simply look for asymmetry (skew) and peakedness (kurtosis). As will be shown, in some cases, the data must be "binned" before being plotted.

Watch the video: https://tinyurl.com/Simple-Histograms

Unit 5. Statistics with R: Introduction and Descriptive Statistics

LEYRE CASTRO

Summary. In this unit you will learn the basics of how R works and how to get comfortable interacting with this software. In addition to the information here, next units will include examples of how to conduct in R the different analyses explained in those units.

Prerequisite Units
Unit 1. Introduction to Statistics for Psychological Science
Unit 2. Managing Data
Unit 3. Descriptive Statistics for Psychological Research

Statistical Software for Data Analysis

Using any kind of statistical software will allow you to avoid mistakes and be faster in the computation of your statistical analyses. To start taking advantage of computer software for your data analysis, spreadsheets (like Excel) are good because they allow you to organize the data the way you want, you can sort and filter data, you can count and summarize values,

calculate basic descriptive statistics, and make graphs. But if you want to move beyond summaries and basic graphs, you will need more specialized statistical software. Some common and traditional statistical applications are SPSS and SAS, but they require a (very expensive) commercial license. A non-commercial option is *jamovi*, an open-source application, with point-and-click interfaces for common tasks and data exploration and modeling. But you may have data that do not fit into the rows and columns that standard statistical applications expect or you may have questions that go beyond what the drop-down menus allow you to do. In that case, you will be better off by using a programming language like R because that gives you the ultimate control, flexibility, and much power in analyzing your data in ways that specifically address the questions that matter to you.

What is R?

R is an open-source free programming software. So, R is free, and you can keep updating it without any cost and have it always available in your computer, regardless of the university or company in which you are. In addition, there are many great websites, videos, and tutorials to get you started (and to acquire advanced knowledge) with R, and to explain you how to do statistics with R. We include links to that information down below.

R is very versatile. Because R is, basically, a programming language, it can be used for a variety of things, not just statistics. As you get better at using R for data analysis, you are also learning to program. If you are interested in science, it is highly likely that you will need to learn the basics of computer modeling, or you may want to develop useful apps, or automatize tasks in your business, or conduct surveys online, or communicate information through data visualization. All this can be done with R.

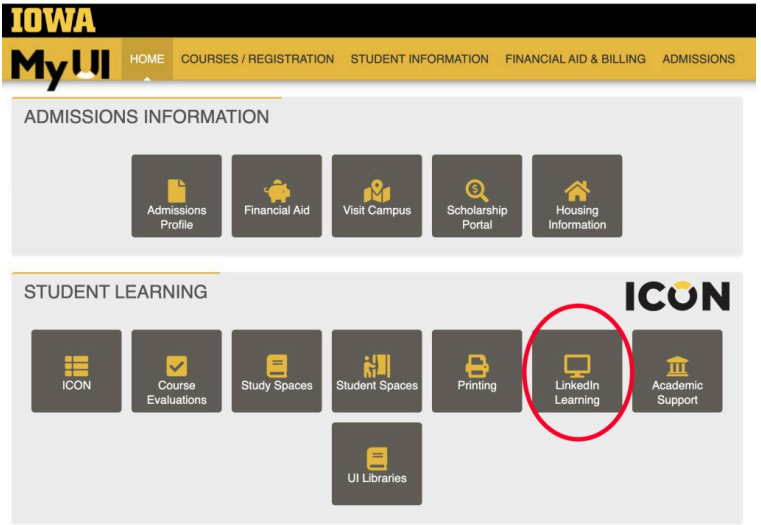
Related to the previous point, R is highly extensible. When you download and install R, you get all the basic "packages," and those are very powerful on their own. Packages are specific units of code that add more functionality to R. Because R is open and so widely used, it has become a standard tool in statistics, so many people write their own packages that extend the system. And these packages are freely available too. There is a large R community for code development and support; indeed, for any kind of special analysis that you need to conduct, be reassured that there will be a package in R, and great explanations about how to perform it. Also, many recent advanced statistical textbooks use R. So, if you learn how to do your basic statistics in R, then you will be a lot closer to being able to use the state-of-the-art methods in psychological statistics and data analysis. In summary, learning R is a very good use of your time.

R is a real programming language. To some people this might seem like a bad or scary thing, but in truth, programming is a core research skill across many of the social and behavioral sciences.

Think about how many surveys and experiments are done online or on computers in a laboratory. Think about all those online social environments which you might be interested in studying. Also, think about how much time you will save and how much accuracy you will gain if you collect data in an automated fashion. If you do not know how to program, then learning how to do statistics using R is a good way to start. Indeed, if you have to or want to learn another programming language in the future, the experience with R will facilitate your learning tremendously.

Learning R

Before moving forward, we highly recommend that you first watch the following tutorial, that you can access through the *LinkedIn Learning* button in your MyUI page:



The tutorial is called **Learning R, by Barton Poulson**, and is 2h, 51m long. It is a very clear and well-paced introduction to R. You will learn to: install R and RStudio, navigate the RStudio environment, import data from a spreadsheet, data visualization, and how to perform a number of data analysis.

In addition, these are some of the best materials available:

Learning Statistics with R, by Danielle Navarro
https://learningstatisticswithr.com/

Excellent, entertaining, and very clear book, freely available online. It includes statistical explanations and how to conduct all the analyses in R. You can download it as a pdf.

R coder. All about R programming
https://r-coder.com/r-tutorials/

Very clear, well-organized, and helpful tutorials for all basic statistics.

Statistics, R programming, and Data Science with Professor Marin

https://www.youtube.com/c/marinstatlectures/playlists?view=50&shelf_id=15

YouTube videos with excellent explanations and examples about how to use R and how to conduct a variety of analyses.

Descriptive Statistics with R

To start doing descriptive statistics with R, you will find excellent instructions in these specific pages from the websites listed above:

Bar Charts and Pie Charts in R

https://www.youtube.com/watch?v=Eph_Y0BmHU0&list=PLqzoL9-eJTNCzF2A6223SQm0rLYtw6hJE&index=2

Histograms in R

https://www.youtube.com/watch?v=Hj1pgap4UOY&list=PLqzoL9-eJTNCzF2A6223SQm0rLYtw6hJE&index=4

Mean, Standard Deviation, and Frequencies in R

https://www.youtube.com/watch?v=ACWuV16tdhY&list=PLqzoL9-eJTNCzF2A6223SQm0rLYtw6hJE&index=11

Descriptive Statistics in R

https://r-coder.com/r-statistics/

Unit 6. Brief Introduction to Statistical Significance

Brief Introduction to Statistical Significance

LEYRE CASTRO AND J TOBY MORDKOFF

Summary. In this unit we will give you a brief introduction to null-hypothesis testing and the concept of statistical significance.

Prerequisite Units
Unit 1. Introduction to Statistics for Psychological Science
Unit 2. Managing Data
Unit 3. Descriptive Statistics for Psychological Research

Null-Hypothesis Testing and Probability Value

Null-hypothesis testing belongs to the area of **inferential statistics** (not part of this version of Data Analysis in the Psychological Sciences), but you need to know some basic notions to be able to read scientific articles and to advance in your study of statistical analysis of psychological research in the following units.

When you do research, you study **samples** that are selected

from a **population**. The data collected from samples are used to make inferences about that population. Thus, you need to have some way of deciding how meaningful the sample data are.

One common tool to help make that decision is testing for statistical significance, technically named null-hypothesis testing. You are testing your hypothesis against the null hypothesis, that states that there are no differences between groups, or no association between your variables of interest, in the population. The output of this statistical analysis includes what is called the *p* (probability) value. So, in research papers, you may find statements like the these:

- Participants who exercised remembered significantly more words than those who did not, $t(48) = 5.63$, $p < .001$.
- Multi-tasking activity was significantly correlated with sensation seeking as measured by the Sensation Seeking Inventory, $r(275) = .45$, $p = .01$.
- Levels of social support were negatively associated with levels of depression at 12-month follow-up, $r(128) = -.32$, $p = .003$.

The p value tells you the probability of finding a result (a difference between groups or a correlation) in your sample, when that difference or that correlation DOES NOT exist in the population.

To be correct (that is, to correctly infer that a result in your sample can be assumed in the population), that probability has to be low. How low? In social sciences, the typical cut-off value is .05; that is, 5 out of 100. Or, as it is typically written, $p < .05$. So, less than .05 is the probability of finding a difference or a correlation in your sample when that difference or correlation does not exist in the population (so, you are finding it by chance, because peculiar circumstances related to your study or to your sample). In other words, if your *p* value is less than

.05, you would expect that less than 5 out of 100 times that you were to replicate your study (with different samples) you would find a difference or a correlation just by chance and not because it actually exists in the population. When $p < .05$, you will say that your result is statistically significant.

Other Tools to Evaluate your Research Results

Statistics in psychological research, the same as in any other scientific area, are subject to criticism and reevaluation. Some practices are well established, but they may have some flaws, and it may be desirable to move forward and find better a way. Still, those better ways need to be explored, widely adopted, and become part of the well-established set of tools to do psychological research.

In the last years, a debate has developed as to what is the best way to analyze and present the results of psychological research. \Some people have criticized null-hypothesis testing because it encourages dichotomous thinking. That is, an effect is statistically significant or not. But this may not be the best way to approach our results. Some results may be very close to the cut-off value of .05. The *p* value may be = .04 and then we say that the result was statistically significant, but it may be = .06 and then we say that the result was not statistically significant. In the first case, we conclude that an effect exists, whereas in the second case we conclude that it does not. But is this fair? A minimal difference can result in concluding two opposite things. This, among many other issues, is one of the reasons why some researchers favor to include other techniques (e.g., confidence intervals and effect sizes) to evaluate research results.

One of those techniques, confidence intervals, will be

introduced in Unit 7. The confidence interval provides the range of likely values for a population parameter such as the population mean or the correlation between two variables in the population, calculated from our sample data. A confidence interval with a 95 percent confidence level (95% CI) has a 95 percent chance of capturing the population mean. We may find that the mean grade in our sample of students is 86.46, and that the 95% CI ranges from 83.40 to 89.53. We will report it like this:

- The grade in our sample of students was approximately a B letter grade, *M* = 86.46, 95% CI [83.40, 89.53].

In this case, we will expect that the true mean in the population will be no lower than 83.40, and no higher than 89.53. So, we can conclude that the mean grade of our sample is a quite accurate estimation of the mean grade of our population.

How narrow or wide the confidence interval is will allow us to assess how accurate we are in our estimation. If the 95% CI for our mean of 86.46 were from 68.72 to 96.45, then the true mean grade in the population could be as low as a D, or as high as an A. Thus, we cannot very well estimate our population's mean grade. In general, a narrower confidence interval will allow us for a more accurate estimation of the correlation in the population.

Conclusions

Statistical analyses include ways to evaluate how reliable or meaningful they are. Statistical software (like R, SPSS, jamovi, etc.) will give you a *p* value when you compute correlations (Unit 7 and 8) and linear regression (Unit 9 and 10) analyses, so you need to be able to interpret those *p* values. Typically, you

will also obtain confidence intervals for the diverse estimations that are calculated. Keep in mind that, whereas the *p* value leads you to conclude that, for example, a correlation is statistically significant or not, a confidence interval for the same correlation value will give you a possible range of values that are more or less likely to be true in the population.

Unit 7. Correlational Measures

LEYRE CASTRO AND J TOBY MORDKOFF

Summary. In this unit, we will start analyzing how two different variables may relate to one another. We will concentrate on Pearson's correlation coefficient: how it is calculated, how to interpret it, and different issues to consider when using it to measure the relationship between two variables.

Prerequisite Units
Unit 1. Introduction to Statistics for Psychological Science
Unit 2. Managing Data
Unit 3. Descriptive Statistics for Psychological Research
Unit 6. Brief Introduction to Statistical Significance

Measures of Association

On prior units, we have focused on statistics related to one specific variable: the mean of a variable, the standard deviation of a variable, how the values of a variable are distributed, etc. But often we want to know, not only about one variable, but also how one variable may be related to another variable. So, in this unit we will focus on analyzing the possible association between two variables.

Choice of the appropriate measure of association depends

on the types of variables that we are analyzing; that is, it depends on whether the variables of interest are quantitative, ordinal, or categorical, and on the possible range of values of these variables. For example, we may be interested in the differences in healthy lifestyle scores (a continuous **quantitative variable**) between people who graduated from college and people who did not graduate from college (a **categorical variable**). Note that this categorical variable has only two values; because of that, it is called a dichotomous variable. In this situation, when one variable consists of numerical, continuous scores, and the other variable has only two values, we use the ***point-biserial correlation*** to measure the relationship between the variables.

In other cases, we are interested in how two categorical variables are related. For example, we may want to examine whether ethnic origin is associated to marital status (single, partnership, married, divorced). In this situation, when the two variables are categorical, you use a **chi-square** as the measure of association.

We could also be interested in the relationship between **ordinal variables**. If, for example, we want to see if there is a relationship between student rankings in a language test and in a music test, the two variables are ordinal or ranked variables. In this case, when two variables are measured on an ordinal scale, we should use **Spearman's correlation** to measure the strength and direction of the association between the variables.

You need to know that these possibilities exist. However, in this unit we are going to focus on the analysis of the relationship between two variables when both variables are quantitative. In this case, we typically use **Pearson's correlation coefficient**. For example, when we want to see if there is a relationship between time spent in social media and scores in an anxiety scale. But before looking at this statistical analysis in detail, let's clarify a few issues.

Correlation as a Statistical Technique

First of all, we need to distinguish between correlational research and correlational statistical analyses. In a correlational research study, we measure two or more variables as they occur naturally to determine if there is a relationship between them. We may read that "as people make more money, they spend less time socializing with others," or "children with social and emotional difficulties in low-income homes are more likely to be given mobile technology to calm them down," or "babies' spatial reasoning predicts later math skills." The studies supporting these conclusions are correlational because the researcher is measuring the variables "making money," "time socializing," "social and emotional difficulties," and "babies' spatial reasoning skills," rather than manipulating them; that is, the researcher is measuring these variables in the real word, rather than determining possible values of these variables and assigning groups to specific values. When we say that a study is correlational in nature we are referring to the study's design, not the statistics used for analysis. In some situations, the results from a correlational study are analyzed using something other than a correlational statistic. Conversely, the results from some experiments, in which one of the variables was determined by the experimenter, such as the number of dots on a computer screen, can be analyzed using a correlational statistic. Thus, it is important to understand when is appropriate to use a correlational statistical technique.

When we analyze a correlation, we normally are looking at the relationship between two numerical variables, so that we have 2 scores for each participant. If we want to see whether spatial skills at the age of 4 correlate with mathematical skills at the age of 12, then we will have one score for spatial skills at 4, and one score for mathematical skills at 12, for each individual participating in our study.

These scores can be represented graphically in a **scatterplot**. In a scatterplot, each participant in the study is represented by a point in two-dimensional space. The coordinates of this point are the participant's scores on variables X (in the horizontal axis) and Y (in the vertical axis).

How we decide which variable goes on the abscissa (x-axis) or on the ordinate (y-axis) depends on how we interpret the underlying relationship between our variables. We may have a predictor variable that is related to an outcome (also called response or criterion variable), as in the case of spatial skills at the age of 4 (predictor) and mathematical skills at the age of 12 (outcome). In this case, the predictor variable is represented on the x-axis, whereas the outcome variable is represented on the y-axis, as shown in Figure 7.1.

This distinction between predictor and outcome variables may be obvious in some cases; for example, smoking will be the predictor and whether or not lung cancer develops will be the outcome, or healthy diet will be the predictor and cardiovascular disease the outcome. But may not always be; for example, time to complete a task may be related to accuracy in that task, but it is not clear whether accuracy depends on time or time depends on accuracy. In this latter case, it does not matter which variable is represented on the x-axis or the y-axis. However, when creating a scatterplot, be aware that most people will have a tendency to interpret the variable in the x-axis as the one leading to the variable in the y-axis.

A scatterplot provides you with a quick visual, intuitive way to assess the correlation between the two variables. It could be that there is no correlation, or the correlation is positive, or negative, as illustrated in Figure 7.1.

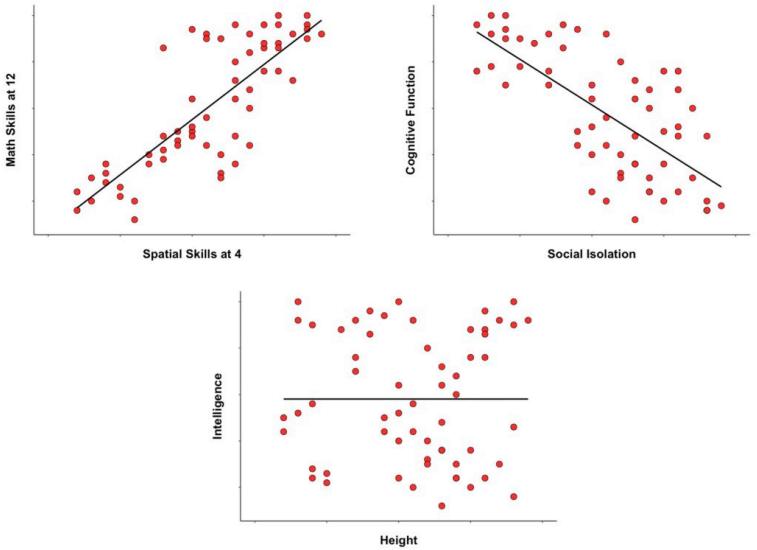


Figure 7.1. Three scatterplots depicting different types of relationships. On the top left, a scatterplot showing a positive relationship between spatial skills at the age of 4 and mathematical skills at the age of 12. On the top right, a scatterplot showing a negative relationship between social isolation and cognitive functioning in the elderly. On the bottom, a scatterplot showing no relationship between height and intelligence. The straight lines show the trend of the linear relationship between the variables.

If spatial skills in 4-year olds are related to mathematical skills at the age of 12, so that children who at the age of 4 show poor spatial skills tend to show poor mathematical skills eight years later, and children who at the age of 4 show good spatial skills tend to show good mathematical skills eight years later, then there is a positive correlation between spatial skills at the age of 4 and mathematical skills at the age of 12. This positive correlation, where the two variables tend to change in the same direction, is represented on the top left panel. On the top right panel we see an example of a negative correlation. In this case, it was observed that, in the elderly, the higher

the social isolation, the lower their cognitive functioning. The correlation is negative because the two variables go in opposite directions: as social isolation increases, cognitive functioning decreases. On the bottom panel, height and intelligence are depicted. In this case, the data points are distributed without showing any trend, so there is no correlation between the variables.

Pearson's Correlation Coefficient

Visual inspection of a scatterplot can give us a very good idea of how two variable are related. But we need to conduct a statistical analysis to confirm a visual impression or, sometimes, to uncover a relationship that may not be very obvious. There are different correlational analysis, depending on the type of relationship between the variables that we want to analyze. When we have two quantitative variables that change together—so that when one of the variables increases, the other variable increases as well, or when one of the variables increases, the other variable decreases—the most commonly used correlational analysis is *Pearson product-moment correlation*, typically named *Pearson's correlation coefficient* or *Pearson's r* or just *r*.

In order to use Pearson's correlation coefficient:

1. Both variables must be **quantitative**. If the variables are numerical, but measured along an ordinal scale, Spearman coefficient should be used, instead.
2. The relationship between the two variables should be **linear**. Pearson's correlation coefficient can only detect and quantify linear (i.e., "straight-line") relationships. If the data in the scatterplot show some kind of curvilinear trend, then the relationship is not linear and a more

complicated procedure should be used, instead.

Pearson's correlation coefficient measures the direction and the degree of the linear relationship between two variables. The value of *r* can range from +1 (perfect positive correlation) to -1 (perfect negative correlation). A value of 0 means that there is no correlation. The ***sign* of *r*** tells you the direction (positive or negative) of the relationship between variables. The ***magnitude of r*** tells you the degree of relationship between variables. Let's say that we obtain a correlation coefficient of 0.83 between physical activity (exercise hours per week) and scores on an academic test. What does it mean? Since 0.83 is positive and close to 1.00, you can say that the two variables have a **strong positive** relationship –so **high** number of exercise hours per week are related to **high** scores on the academic test. In a different situation, let's say that we obtain a correlation coefficient of -0.86 between number of alcoholic drinks per week and scores on an academic test. What does it mean? Since -0.86 is negative and close to -1.00, you can say that the two variables have a **strong negative** relationship –so **high** number of alcoholic drinks per week are related to **low** scores on the academic test.

A slightly more complicated way to quantify the strength of the linear relationship is by using the square of correlation: R^2. The reason why this is sometimes preferred is because R^2 is the proportion of the variability in the outcome variable that can be "explained" by the value of the predictor variable. If we obtain an R^2 of 0.23 when analyzing the relationship between spatial skills at the age of 4

(predictor) and mathematical skills at the age of 12 (outcome), we will say that spatial skills at the age of 4 explain 23% of the variability in mathematical skills at the age of 12. This amount of explained variance is one way of expressing the degree to which some relation or phenomenon is present.

Importantly, when you use R^2 as your measure of strength, you can make statements like "verbal working memory score is twice as good at predicting IQ than spatial working memory score" (assuming that the R^2 for verbal WM and IQ is twice as large as the R^2 for spatial WM and IQ). You cannot make statements of this sort based on (*unsquared*) values of *r*.

Conceptually, Pearson's correlation coefficient computes the degree to which change in one numerical variable is associated with change in another numerical variable. It can be described in terms of the covariance of the variables, a measure of how two variables vary together. When there is a perfect linear relationship, every change in the X variable is accompanied by a corresponding change in the Y variable. The result is a perfect linear relationship, with X and Y always varying together. In this case, the covariability (of X and Y together) is identical to the variability of X and Y separately, and *r* will be positive (if the two variables increase together) or negative (if increases in one variable correspond to decreases in the other variable).

To understand the calculations for *r*, we need to understand the concept of **sum of products of deviations** (SP). It is very

similar to the concept of sum of squared deviations (SS) that we saw in Unit 3 to calculate the variance and standard deviation.

This was the formula for the SS of one single variable, X:

$$SS_x = \sum (X_i - \bar{X})^2$$

In order to see the similarities with the SP, it will be even clearer if we write the formula for the SS this way:

$$SS_x = \sum (X_i - \bar{X})(X_i - \bar{X})$$

The formula for the sum of the products of the deviation scores or SP computes the deviations of each score for X and for Y from its corresponding mean, and then multiplies and add those values:

$$SP_{xy} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

In order to show the degree to which X and Y vary together, that is, their covariance (similar to the variance for one variable, but now referring to two variables), we divide by n – 1:

$$cov_{xy} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Practice (1)

Let's calculate the SP and the covariance with the dataset containing the number of

study hours before an exam (X), and the grade obtained in that exam (Y), for 15 participants, that we used previously in Unit 3. The table shows each of the scores, the deviation scores for each X and Y score, and the product of each pair of deviation scores:

Part.	Hours	Grade	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
P1	8	78	(8 – 13.66)	(78 – 86.46)	-5.66 * -8.46 = 47.88
P2	11	80	(11 – 13.66)	(80 – 86.46)	-2.66 * -6.46 = 17.18
P3	16	89	(16 – 13.66)	(89 – 86.46)	2.34 * 2.54 = 5.94
P4	14	85	(14 – 13.66)	(85 – 86.46)	0.34 * -1.46 = – 0.50
P5	12	84	(12 – 13.66)	(84 – 86.46)	-1.66 * -2.46 = 4.08
P6	15	86	(15 – 13.66)	(86 – 86.46)	1.34 * -0.46 = – 0.62
P7	18	95	(18 – 13.66)	(95 – 86.46)	4.34 * 8.54 = 37.06
P8	20	96	(20 – 13.66)	(96 – 86.46)	6.34 * 9.54 = 60.48
P9	10	83	(10 – 13.66)	(83 – 86.46)	-3.66 * -3.46 = 12.66
P10	9	81	(9 – 13.66)	(81 – 86.46)	-4.66 * -5.46 = 25.44
P11	16	93	(16 – 13.66)	(93 – 86.46)	2.34 * 6.54 = 15.30
P12	17	92	(17 – 13.66)	(92 – 86.46)	3.34 * 5.54 = 18.50
P13	13	84	(13 – 13.66)	(84 – 86.46)	-0.66 * -2.46 = 1.62
P14	12	83	(12 – 13.66)	(83 – 86.46)	-1.66 * -3.46 = 5.74

Part.	Hours	Grade	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
P15	14	88	(14 – 13.66)	(88 – 86.46)	0.34 * 1.54 = 0.52

Now, as the term $\sum (X_i - \bar{X})(Y_i - \bar{Y})$ in the formula indicates, we need to add all the products of each pair of deviation scores, the scores in the rightmost column. This total adds up to 251.33. Then, we divide by n -1, that is, by 14:

$$cov_{xy} = \frac{251.33}{14} = 17.95$$

Thus, the covariance between Hours and Grade is 17.95.

It could be possible to use the covariance between X and Y as a measure of the relationship between the two variables; however, its value is not quickly understandable or possible to compare across studies because it depends on the measurement scale of the variables, that is, it depends on the specific units of X and the specific units of Y. Pearson's correlation coefficient solves this issue, by dividing the covariance by the specific value of the standard deviations of X and Y, so that the units (and scale effects) cancel:

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

With this maneuver, the limits of *r* range between -1 and +1 and, therefore, *r* is easy to interpret, and its value can be used to compare different studies.

Practice (2)

Let's calculate Pearson's correlation coefficient in our case. We know from Unit 3 that the standard deviation of X, the number of hours, is 3.42, and that the standard deviation of Y, the grade, is 5.53. Therefore:

$$r = \frac{17.95}{3.42 * 5.53} = 0.95$$

So, Pearson's correlation coefficient between number of hours studying and grade obtained in our dataset is 0.95. Very high.

As indicated above, the magnitude of *r* tells us how weak or strong the relationship is so, the closer to 0, the weaker the relationship is, whereas the closer to 1 (or -1), the stronger the relationship is. Figure 7.2 shows different scatterplots illustrating different values of *r*.

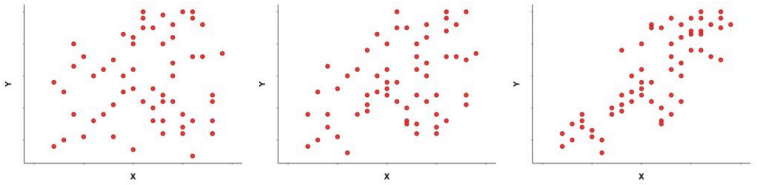


Figure 7.2. Three scatterplots depicting positive relationships of different strength between variables X and Y. On the left, a scatterplot in which r = 0.1. On the center, a scatterplot in which r = 0.4. On the right, a scatterplot in which r = 0.8.

A perfect correlation of +1 or -1 means that all the data points lie exactly on a straight line. These perfect relationships are rare, in general, and very unlikely in psychological research.

Interpretation of Pearson's Correlation Coefficient

The interpretation of the value of the correlation coefficient *r* is somehow arbitrary (see Table 7.1 for typical guidelines). Although most data scientists will agree that an *r* smaller than 0.1 reflects a negligible relationship, and an *r* larger than 0.9 reflects a very strong relationship, how to interpret intermediate coefficients is more uncertain. An *r* value of 0.42 may be weak or strong depending on the typical or possible association found between some given variables. For example height and weight are typically highly correlated, so an *r* value of 0.42 between those variables would be low; however, an *r* value of 0.42 between eating cranberries daily and cognitive capacity would be high, given that so many other variables are related to cognitive capacity. It also may be weak or strong depending on the results of other research studies in the same

area. Thus, a specific *r* value should be interpreted within the context of the specific research.

Value of *r*	Strength
0.0 to 0.09	negligible
0.1 to 0.29	weak
0.3 to 0.49	moderate
0.5 to 0.89	strong
0.9 to 1.00	very strong

Table 7.1. Typical guidelines for the interpretation of *r*

Issues to Consider

Outliers

An outlier is a data point that has a value much larger or smaller than the other values in a data set. It is important to carefully examine the dataset for outliers because they can have an excessive influence on Pearson's correlation coefficient. For example, in Figure 7.3, we can see a data point that has X and Y values much larger than the other data points. Pearson's *r* for this entire data set is *r* = 0.86, indicating a strong relationship between X and Y. However, if we do not include the outlier in the analysis, Pearson's r is greatly reduced, *r* = 0.07, very close to 0, indicating a negligible relationship. You should be able to easily visualize the difference with and without the outlier data point in the scatterplot in Figure 7.3.

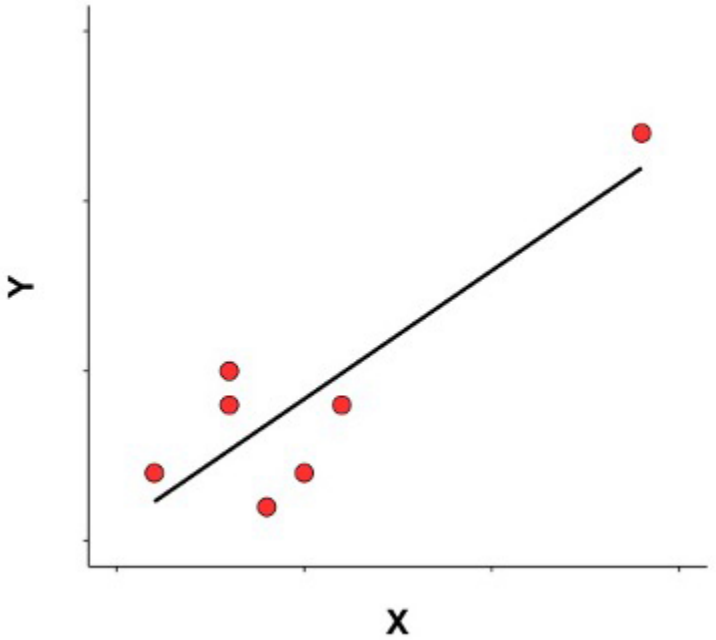


Figure 7.3. Scatterplot of the relationship between two variables, X and Y, with one outlier in the data set.

You always need to examine the distribution of your scores to make sure that there are no relevant outliers. Whether outliers should or should not be eliminated from an analysis depends on the nature of the outlier. If it is a mistake in the data collection, it should probably be eliminated; if it is an unusual, but still possible value, it may need to be retained. One way or the other, the presence of the outlier should be noted in your data report. If the outlier must be retained, an alternative analysis can be Spearman correlation, that is more robust than Pearson's correlation coefficient against outliers.

Restricted Range

You need to be aware that the value of a correlation can be influenced by the range of scores in the dataset used. If a correlation coefficient is calculated from a set of scores that do not represent the full range of possible values, you need to be cautious in interpreting the correlation coefficient. For example, you may be interested in the relationship between family income level and educational achievement. You choose a convenient sample in a nearby school, that happens to be a private school in which most students come from wealthy and very wealthy families. You analyze the data in your sample and find that there is no correlation between family income level and educational achievement. It could be that, indeed, this relationship is not apparent among high-income students, but it would have been revealed if you have included in your sample students from very-low, low, and middle-income families. Figure 7.4 shows a scatterplot that depicts this issue.

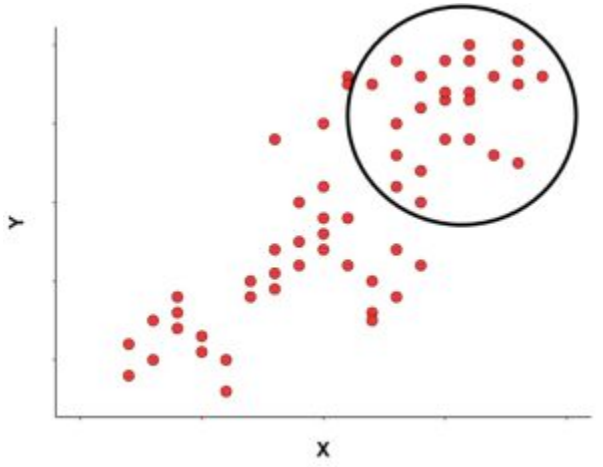


Figure 7.4. Considering all the possible values for X and Y in this scatterplot, r = 0.80. However, if the data set consisted of a restricted range of values, as those high values included in the circle, r would be close to 0.

In general, in order to establish whether or not a correlation exists between two variables, you need a wide range of values for each of the variables. If it is not possible to obtain this wide range of values, then at least you should limit your interpretation to the specific range of values in your dataset.

Correlation Coefficient in the Sample and in the Population

We rarely are interested in the correlation between variables that only exists in a sample. As is true for most situations, we use the correlation coefficient in our **sample** to make an estimate of the correlation in the **population**. How does that work? The sample data allow us to compute *r*, the correlation coefficient for the sample. We normally do not have access to the entire population, so we cannot know or calculate the correlation coefficient for the population (named rho, ρ). Because of this, we use the sample correlation coefficient, *r*, as our estimate of the unknown population correlation coefficient . The accuracy of this estimation depends on two things. First, the sample needs to be representative of the population; for example, the people included in the sample need to be an unbiased, random subset of the population—this is not a statistical issue, but it should always be kept in mind. Second, how precisely a sample correlation coefficient will match the population correlation coefficient depends on the size of the sample: the larger the sample, the more accurate the estimate (provided that the sample is representative).

Whenever some value that is calculated from a sample is used to estimate the (true) value in the population, the question of bias arises. Recall that bias is whether the estimator has a tendency to over-shoot or under-shoot the target value. If possible, we always use the estimator that has the least bias.

In the case of correlation coefficients, the value of ***r* from a sample provides a very good estimate of ρ in the population, as long as the sample size is not very small**. If you are working with samples of fewer than 30 participants, you may wish to adjust the value of *r* when using it as an estimate of ρ. The details of this adjustment are beyond the scope of this unit, but you should be aware that better estimates of ρ are available for small-sample situations.

Before conducting our study, we need to decide the size of our sample. This decision must be informed by the purpose of minimizing inaccurate estimates when we later analyze our sample data. So, we need to plan for the sufficient sample size. It is not easy to give a specific number, because our sample size should be based on prior and expected sizes of the relationship of interest. And, the size of the sample should be large enough to be able to detect small effects, and make sure that our results are not due to chance. Nowadays, there are a variety of software tools that can help you decide the size of the sample. At the moment, just be aware that, if you have a sample that is too small, the correlation coefficient that you obtain from your sample data may be inadequate as the correlation coefficient for the population.

In addition, in order to improve the interpretation of your correlation coefficient *r*, a **confidence interval** will help. That's why is always advisable to include a confidence interval for the obtained coefficient (typically, a 95% confidence interval). The confidence interval provides the range of likely values of the coefficient in the population from which the sample was taken. An *r* value of 0.53 suggests quite a strong relationship between two variables; however, if the 95% confidence interval ranges from 0.02 to 0.82 (as it could be the case with a very small sample), then the strength of the relationship in the population could be negligible ($r = 0.02$) and, therefore, of little importance, or it could be strong ($r = 0.82$) and, therefore, of high relevance. So, if the confidence interval is very wide, it is difficult to make a valuable interpretation of the results. In general, a narrower confidence interval will allow us for a more accurate estimation of the correlation in the population.

Conclusions

A correlation coefficient shows the strength and direction of an association between two variables. Note that a correlation describes a relationship between two variables, but does not explain why. Thus, it should never be interpreted as evidence of a causal relationship between the variables.

If *r* is relatively strong, you can assume that when one variable increases, the other variable will increase as well (for a positive relation) or the other variable will decrease (for a negative relation). But *r* does not allow you to predict, precisely, the value of one variable based on the value of the other variable. To do that, we have another statistical tool: linear regression analysis (Unit 9)

Unit 8. Scatterplots and Correlational Analysis in R

LEYRE CASTRO

Summary. In this unit you will learn how to create scatterplots and how to calculate Pearson's correlation coefficient with R. You will learn how to enter the code and how to interpret the output that R provides.

Prerequisite Units
Unit 5. Statistics with R: Introduction and Descriptive Statistics
Unit 6. Brief Introduction to Statistical Significance
Unit 7. Correlational Measures

Reading Data and Creating a Scatterplot

We have the dataset of 50 participants with different amount of experience (from 1 to 16 weeks) in performing a computer task, and their accuracy (from 0 to 100% correct) in this task. Thus, Experience is the predictor variable, and Accuracy is the outcome variable.

The code line below imports our dataset by using the

read.table function, and assigns the data file to the object *MyData*.

```
  #read data file
MyData <-
read.table("ExperienceAccuracy.txt", header
= TRUE, sep = ",")
```

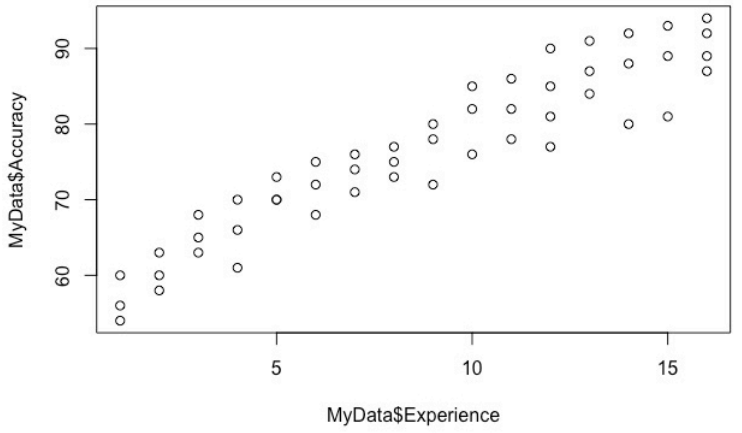
Once R has access to your data file, you can create a scatterplot. There are multiple ways of creating a scatterplot in R (links are included below). Some ways are quick and simple, so you can easily see the graphical representation of how your variables of interest are related; other ways require some additional packages. Let's see here two frequently used options.

1. Using **plot** function

For the simplest scatterplot, you just need to specify the two variables that you want to plot. The first one will be on the x-axis and the second one on the y-axis. Remember that you need to specify, with the $ sign, that your variables, Experience and Accuracy, are within the object MyData.

```
  #basic scatterplot
plot(MyData$Experience, MyData$Accuracy)
```

This is the scatterplot that you will obtain:

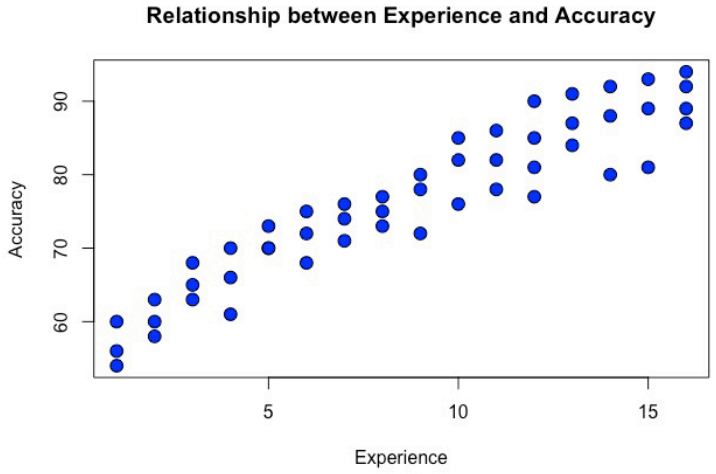


The **plot** function has a number of possibilities to modify and improve the basic scatterplot. For example:

```
  #basic scatterplot
plot(MyData$Experience, MyData$Accuracy,
main = "Relationship between Experience and
Accuracy",
pch = 21,
bg = "blue",
xlab = "Experience",
ylab = "Accuracy")
```

The **main** argument allows you to include a title to the scatterplot. You can also choose the shape of the points, with

pch (you can find the assignment of shapes to numbers in the links included below), and the color, with **bg**. In addition, you can change the labels to the axis, with **xlab** and **ylab**. If you run the script above, you will obtain:



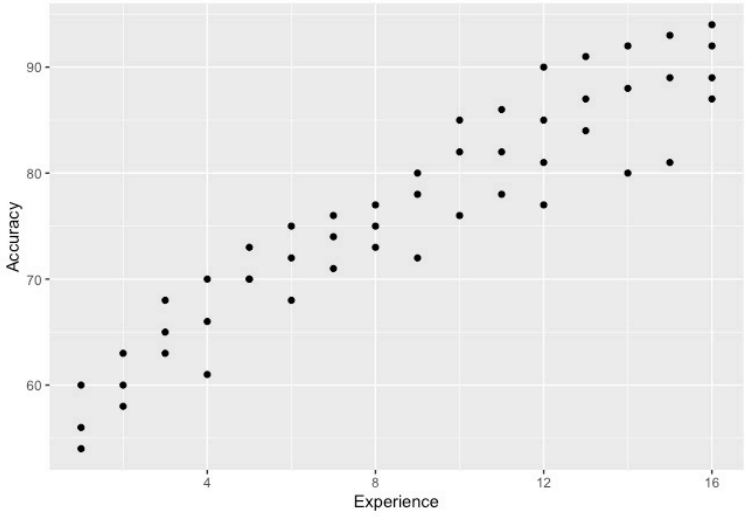
Find more options and information here:
https://r-coder.com/scatter-plot-r/

2. Using the **ggplot** package

First of all, you will need to install the **ggplot** package for R. And, as indicated in the first line of code below, you will to load it when you want to use it, using the **library** function.

```
  #load the ggplot2 package
library(ggplot2)
  #basic scatterplot
scatterplot <- ggplot(MyData) +
aes(x = Experience, y = Accuracy) +
geom_point()
print(scatterplot)
```

In this script, you are just indicating the variables in the x- and y-axis, and that the data are represented by points. To visualize the scatterplot, you have to use **print**. If you run the script above, you will obtain:

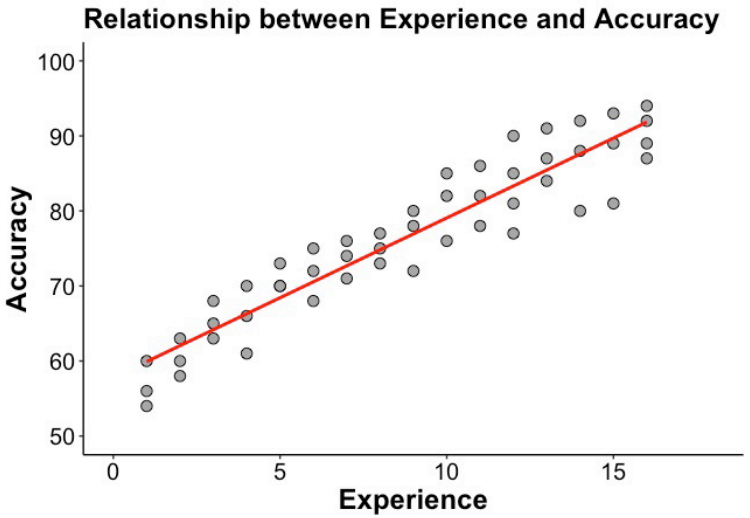


You can elaborate on this scatterplot, and improve as much as you want. The script below shows you some options to modify the data points (with **geom_point**), titles (**xlab** and **ylab**), changes to the scales of the axis (**xlim** and **ylim**), including the regression line (**geom_smooth**), and modifying the text elements with **theme**:

```
  #load the ggplot2 package
library(ggplot2)
  #more elaborated scatterplot
scatterplot <- ggplot(MyData) +
aes(x = Experience, y = Accuracy) +
geom_point(size = 3, fill= “dark grey”,
shape = 21) +
ggtitle (“Relationship between Experience
and Accuracy”) +
xlab (“Experience”) +
ylab (“Accuracy”) +
xlim (0, 18) +
ylim (50,100) +
geom_smooth(method=lm, se = FALSE, color =
“red”, weight = 6) +
theme (axis.text.x =
element_text(colour=”black”,size=15,face=”p
lain”),
axis.text.y =
element_text(colour=”black”,size=15,face=”p
lain”),
axis.title.x =
```

```
element_text(colour="black",size=18,face="b
old"),
axis.title.y =
element_text(colour="black",size=18,face="b
old"),
plot.title =
element_text(colour="black",size=18,face="b
old"),
panel.background = element_rect(fill =
'NA'),
axis.line = element_line(color="black",
size = 0.5)
)
print(scatterplot)
```

Running this script, you will obtain:



You have plenty of options to improve the visual aspects of your scatterplot. Find more in the following websites:

- How to make a scatterplot with ggplot2

  http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization
  https://www.r-bloggers.com/2020/07/create-a-scatter-plot-with-ggplot/

- How to make any plot using ggplot2

  http://r-statistics.co/ggplot2-Tutorial-With-R.html

Correlational Analysis

Once you have a visual representation of how your variables are related, it is time to conduct the correlational analysis that will allow you to obtain Pearson's correlation coefficient between your variables of interest. In R, you can use the **cor** function, as you can see in this code line:

```
#how to obtain Pearson's r
cor(MyData$Experience, MyData$Accuracy)
```

The output will give you Pearson's r. Simply:

```
0.9390591
```

To obtain the result of the statistical significance test and the confidence intervals for the correlation coefficient, you can use **cor.test**:

```
#how to obtain Pearson's r with
significance test and confidence intervals
cor.test (MyData$Experience,
MyData$Accuracy)
```

You will obtain the following output:

```
Pearson's product-moment correlation

data:  MyData$Experience and
MyData$Accuracy
t = 18.926, df = 48, p-value < 2.2e-16
alternative hypothesis: true correlation is
not equal to 0
95 percent confidence interval:
0.8945274 0.9651350
sample estimates:
```

```
cor
0.9390591
```

The null hypothesis in a correlation test is a correlation of 0, that is, that there is no relationship between the variables of interest. As indicated in the output above, the alternative hypothesis is that that the correlation coefficient is different from zero. The *t* statistic tests whether the correlation is different from zero. We have not seen the t statistic yet, so you only need to pay attention to the *p* value. As we explained here, the cut-off value for a hypothesis test to be statistically significant is 0.05, so that if the *p*-value is less than 0.05, then the result is statistically significant. Here, the *p*-value is very small; R uses the scientific notation for very small quantities, and that's why you see the *e* in the number. For values larger than .001, R will give you the exact *p* value. You just need to know that this number, 2.2e-16, represents a very small value, much smaller than 0.05; so, we can conclude that the correlation between Experience and Accuracy is statistically significant.

In the last line you can see Pearson's correlational coefficient, 0.94, indicating a very strong correlation. And, above, the 95% confidence interval for the correlation coefficient. Following APA style, we typically report the confidence interval this way:

95% CI [0.89, 0.96]

So, we obtained an *r* value of 0.94 in our sample, with a 95% CI between 0.89 and 0.96. That is, you can be 95% confident that the true *r* value in the population is between the values of 0.89 and 0.96. This interval is relatively narrow, and any value within the interval would indicate a very strong correlation, so we have a very accurate estimation of the correlation in the population.

More information about correlation tests in R:

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test

https://www.statmethods.net/stats/correlations.html

https://www.statology.org/correlation-test-in-r/

Unit 9. Simple Linear Regression

LEYRE CASTRO AND J TOBY MORDKOFF

Summary. This unit further explores the relationship between two variables with linear regression analysis. In a simple linear regression, we examine whether a linear relationship can be established between one predictor (or explanatory) variable and one outcome (or response) variable, such as when you want to see if time socializing can predict life satisfaction. Linear regression is a widely-used statistical analysis, and the foundation for more advanced statistical techniques.

Prerequisite Units
Unit 6. Brief Introduction to Statistical Significance
Unit 7. Correlational Measures

Linear Regression Concept

Imagine that we examine the relationship between social media use and anxiety in adolescents and, when analyzing the data, we obtain an *r* of 0.40, indicating a positive correlation between social media use and anxiety. The data points in the

scatterplot on the top of Figure 9.1 illustrate this positive relationship. The relationship is even clearer and faster to grasp when you see the line included on the plot on the bottom. This line summarizes the relationship among all the data points, and it is called regression line. This line is the best-fitting line for predicting anxiety based on the amount of social media use. It also allows us to better understand the relationship between social media use and anxiety. We will see in this unit how to obtain this line, and what exactly means.

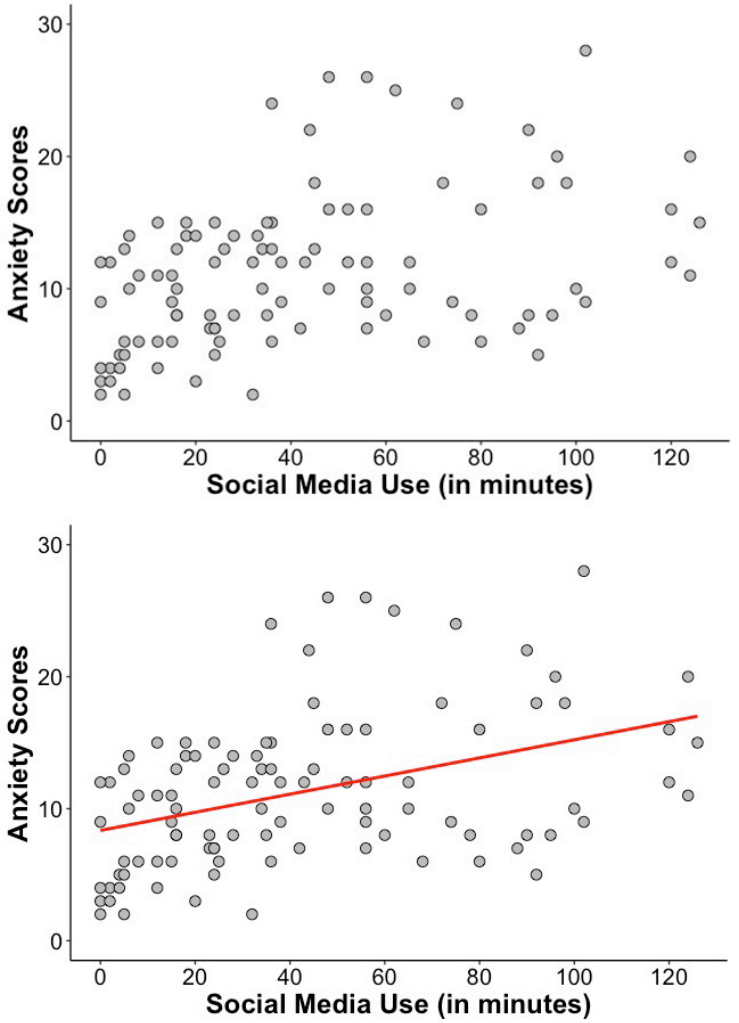


Figure 9.1. Scatterplots showing the relationship between social media use and anxiety scores. The scatterplots are exactly the same, except that the one on the bottom has the regression line added.

A linear regression analysis, as the name indicates, tries to

capture a linear relationship between the variables included in the analysis. When conducting a correlational analysis, you do not have to think about the underlying nature of the relationship between the two variables; it doesn't matter which of the two variables you call "X" and which you call "Y". You will get the same correlation coefficient if you swap the roles of the two variables. However, the decision of which variable you call "X" and which variable you call "Y" does matter in regression, as you will get a different best-fitting line if you swap the two variables. That is, the line that best predicts Y from X is not the same as the line that predicts X from Y.

In the basic linear regression model, we are trying to predict, or trying to explain, a quantitative variable Y on the basis of a quantitative variable X. Thus:

- The X variable in the relationship is typically named **predictor** or **explanatory variable**. This is the variable that may be responsible for the values on the outcome variable.
- The Y variable in the relationship is typically named **outcome or response or criterion variable**. This is the variable that we are trying to predict/understand.

In this unit, we will focus on the case of one single predictor variable, that's why this unit is called "simple linear regression." But we can also have multiple possible predictors of an outcome; if that is the case, the analysis to conduct is called multiple linear regression. For now, let's see how things work when we have one possible predictor of one outcome variable.

Linear Regression Equation

You may be interested in whether the amount of caffeine

intake (predictor) before a run can predict or explain faster running times (outcome), or whether the amount of hours studying (predictor) can predict or explain better school grades (outcome). When you are doing a linear regression analysis, you model the outcome variable as a function of the predictor variable. For example, you can model school grades as a function of study hours. The formula to model this relationship looks like this:

$$\hat{Y} = b_0 + b_1 X$$

And this is the meaning of each of its elements:

$\hat{Y}$ (read as "Y-hat") = the predicted value of the variable Y

b_0 = the **intercept** or value of the variable when the predictor variable X = 0

b_1 = the **slope** of the regression line or the amount of increase/decrease in Y for each increase/decrease in one unit of X

X = the value of the predictor variable

When we conduct a study, the data from our two variables X and Y are represented by the points in the scatterplot, showing the different values of X and Y that we obtained. The line represents the $\hat{Y}$ values, that is, the predicted values of Y when we use the X and Y values that we measured in our sample to calculate the regression equation. Figure 9.2 shows a regression line depicting the relationship between number of study hours and grades, and where you can identify the intercept and the slope for the regression line.

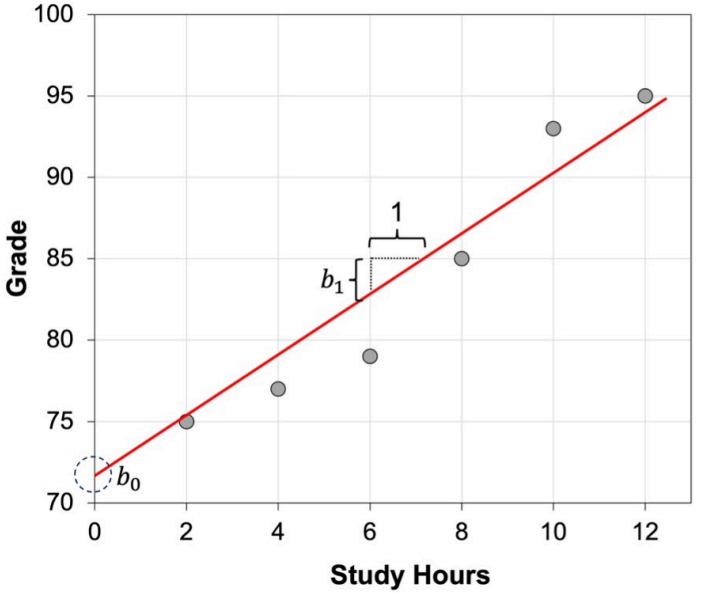


Figure 9.2. Regression line depicting the relationship between number of study hours and grades. For the sake of clarity, only 6 observations are included. Here, the intercept is equal to 72, and the slope is equal to approximately 2. You can check and see that, regardless of where you measure the slope, its value is always the same.

We have defined the intercept as the value of $\hat{Y}$ when X equals zero. Note that this value may be meaningful in some situations but not in others, depending on whether or not having a zero value for X has meaning and whether that zero value is very near or within the range of values of X that we used to estimate the intercept. Analyzing the relationship between the number of hours studying and grades, the intercept would be the grade obtained when the number of study hours is equal to 0. Here, the intercept tells us the expected grade in the absence of any time studying. That's helpful to know. But let's say, for example, that we are examining the relationship

between weight in pounds (variable X) and body image satisfaction (variable Y). The intercept will be the score in our body image satisfaction scale when weight is equal to 0. Whatever the value of the intercept may be in this case, it is meaningless, given that it is impossible that someone weighs 0 pounds. Thus, you need to pay attention to the possible values of X to decide whether the value of the intercept is telling you something meaningful. Also, make sure that you look carefully at the x-axis scale. If the x-axis does not start at zero, the intercept value is not depicted; that is, the value of Y at the point in which the regression line touches the y-axis is not the intercept if the x-axis does not start at zero.

The critical term in the regression equation is b_1 or the slope. We have defined the slope as the amount of increase or decrease in the value of the variable Y for a one-unit increase or decrease in the variable X. Therefore, the slope is a measure of the predicted rate of change in Y. Let's say that the slope in the linear regression with number of study hours and grades is equal to 2.5. That would mean that for each additional hour of study, grade is expected to increase 2.5 points. And, because the equation is modeling a linear relationship, this increase in 2.5 points will happen with one-unit increase at any point within our range of X values; that is, when the number of study hours increase from 3 to 4, or when they increase from 7 to 8, in both cases, the predicted grade will increase in 2.5 points.

The different meanings of β

The letter *b* is used to represent a sample estimate of the β coefficient in the population. Thus b_0 is the sample estimate of β_0, and b_1 is the sample estimate of β_1. As we mentioned in Unit 1, for sample statistics

we typically use Roman letters, whereas for the corresponding population parameters we use Greek letters. However, be aware that β is also used to refer to the standardized regression coefficient, a sample statistic. Standardized data or coefficients are those that have been transformed so as to have a mean of zero and a standard deviation of one. So, when you encounter the symbol β, make sure you know to what it refers.

How to Find the Values for the Intercept and the Slope

The statistical procedure to determine the best-fitting straight line connecting two variables is called regression. The question now is to determine what we mean by the "best-fitting" line. It is the line that minimizes the distance to our data points or, in different words, the one that minimizes the error between the values that we obtained in our study and the values predicted by the regression model. Visually, it is the line that minimizes the vertical distances from each of the individual points to the regression line. The distance or error between each predicted value and the actual value observed in the data is easy to calculate:

$$error = Y - \hat{Y}$$

This **"error" or residual**, as is also typically called, is not an error in the sense of a mistake. It tries to capture variations in the outcome variable that may occur due to unpredicted or unknown factors. If the observed data point lies above the line, the error is positive, and the line underestimates the actual

data value for Y. If the observed data point lies below the line, the error is negative, and the line overestimates that actual data value for Y (see error for each data point in green in Figure 9.3). For example, in Figure 9.3, our regression model overestimates the grade when the number of study hours are 6, whereas it underestimates the grade when the number of study hours are 10.

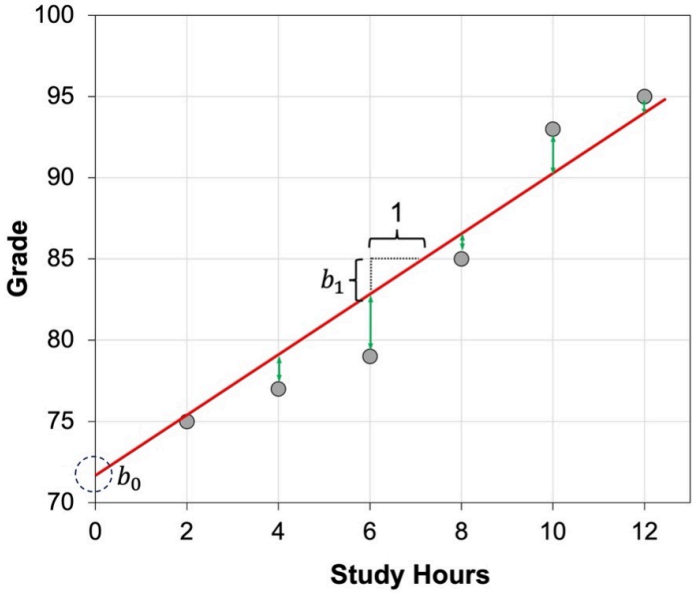


Figure 9.3. The same regression line as in Figure 9.2. depicting the relationship between number of study hours and grades. Here, the errors or residuals, that is, the distance between each predicted value (regression line) and the actual value observed in the data (data points in grey) are indicated with the green lines.

Because some data points are above the line and some are below the line, some error values will be positive while others will be negative, so that the sum of all the errors will be zero. Because it is not possible to conduct meaningful calculations

with zero values, we need to calculate the sum of all the squared errors. The result will be a measure of overall or total squared error between the data and the line:

$$total\ squared\ error = \sum (Y - \hat{Y})^2$$

So, the best-fitting line will be the one that has the smallest total squared error or the smallest sum of squared residuals. In addition, the residuals or error terms are also very useful for checking the linear regression model assumptions, as we shall see below.

Getting back to our calculations, we know that our regression equation will be the one that minimizes the total squared error. So, how do we find the specific values of b_0 and b_1 that generate the best-fitting line?

We start calculating b_1, the slope:

$$b_1 = r(\frac{s_y}{s_x})$$

where $_y$ = the standard deviation of the Y values and $_x$ = the standard deviation of the X values (that you learned to calculate in Unit 3) and *r* is the correlation coefficient between X and Y (that you learned to calculate in Unit 7).

Once we have found the value for the slope, it is easy to calculate the intercept:

$$b_0 = \bar{Y} - b_1\bar{X}$$

Practice

Any statistical software will make these calculations

for you. But, before learning how to do it with R or Excel, let's **find the intercept and the slope of a regression equation** for the small dataset containing the number of study hours before an exam and the grade obtained in that exam, for 15 participants, that we used previously in Unit 3 and Unit 7. These are the data:

Participant	Hours	Grade
P1	8	78
P2	11	80
P3	16	89
P4	14	85
P5	12	84
P6	15	86
P7	18	95
P8	20	96
P9	10	83
P10	9	81
P11	16	93
P12	17	92
P13	13	84
P14	12	83
P15	14	88

As you can tell from the formulas above, we need to know the mean and the standard deviation for each of the variables (see Unit 3). From our calculations in Unit 3, we know that the mean number of hours ($\bar{X}$) is 13.66, and the mean grade obtained in the exam ($\bar{Y}$) is 86.46.

We also know that the standard deviation for hours (s_x) is 3.42, and the standard deviation for grade (s_y) is 5.53. In addition, from our calculations in Unit 7 we know that *r* for the relationship between study hours and grade is 0.95. So, let's calculate first the slope for our regression line:

b_1 = 0.95 * (5.53 / 3.42)

that is:

b_1 = 1.54

Once we have b_1 , we calculate the intercept:

b_0 = 86.46 (the mean grade in the exam) – 1.54 * 13.66 (the mean number of study hours)

that is:

b_0 = 65.47

So, our regression equation is:

$$\hat{Y} = 65.47 + 1.54 * X$$

This means that **the expected grade of someone who studies 0 hours will be 65.47**, and that **for each additional hour of study, a student is expected to increase their grade in 1.54 points**.

Linear Regression as a Tool to Predict Individual Outcomes

The regression equation is useful to make predictions about the expected value of the outcome variable Y given some specific value of the predictor variable X. Following with our

example, if a student has studied for 15 hours, their predicted grade will be:

$\hat{Y}$ = 65.47 + 1.54*15

that is:

$\hat{Y}$ = 88.57

Of course, this prediction will not be perfect. As you have seen in Figure 9.3, our data points do not fit perfectly on the line. Normally, there is some error between the actual data points and the data predicted by the regression model. The closer the data points are to the line, the smaller the error is.

In addition, be aware that we can only calculate the $\hat{Y}$ value for values of X that are within the range of values that we included in the calculation of your regression model (that is, we can interpolate). We cannot know if values that are smaller or larger than our range of X values will display the same relationship. So, we cannot make predictions for X values outside the range of X values that we have (that is, we cannot extrapolate).

Linear Regression as an Explanatory Tool

Note that, despite the possibility of making predictions, most of the times that we use regression in psychological research we are not interested in making actual predictions for specific cases. We typically are more concerned with finding general principles rather than making individual predictions. We want to know if studying for a longer amount of time will lead to have better grades (although you probably already know this) or we want to know if social media use leads to increased anxiety in adolescents. The linear regression analysis provides us with an estimate of the magnitude of the impact of a change in one variable on another. This way, we can better understand the overall relationship.

Linear regression as a statistical tool in both correlational and experimental research

Linear regression is a statistical technique that is independent of the design of your study. That is, whether your study is correlational or experimental, if you have two numerical variables, you could use a linear regression analysis. You need to be aware that, as mentioned at different points in this book, if the study is correlational, you cannot make causal statements.

Assumptions

In order to conduct a linear regression analysis, our data should meet certain assumptions. If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading. These are the four assumptions:

1) The relationship between the variables is linear

The values of the outcome variable can be expressed as a linear function of the predictor variable. The easiest way to find if this assumption is met is to examine the scatterplot with the data from the two variables, and see whether or not the data points fall along a straight line.

2) Observations are independent

The observations should be independent of one another and, therefore, the errors should also be independent. That is, the error (the distance from the obtained Y to the $\hat{Y}$ predicted by the model) for one data point should not be predictable from knowledge of the error for another data point. Knowing whether this assumption is met requires knowledge of the study design, method, and procedure. If the linear regression model is not adequate, this does not mean that the data cannot be analyzed; rather, other analyses are required to take into account the dependence among the data observations.

3) Constant variance at every value of X

Another assumption of linear regression is that the residuals should have constant variance at every value of the variable X. In other words, variation of the observations around the regression line are constant. This is known as **homoscedasticity**. You can see on the scatterplot on the left side of Figure 9.4 that the average distance between the data points above and below the line is quite similar regardless of the value of the X variable. That's what homoscedasticity means. However, on the scatterplot on the right, you can see the data points close to the line for the smaller values of X, so that variance is small at these values; but, as the value of X increases, values of Y vary a lot, so some data points are close to the line but others are more spread out. In this case, the constant variance assumption is not met, and we say that the data show **heteroscedasticity.**

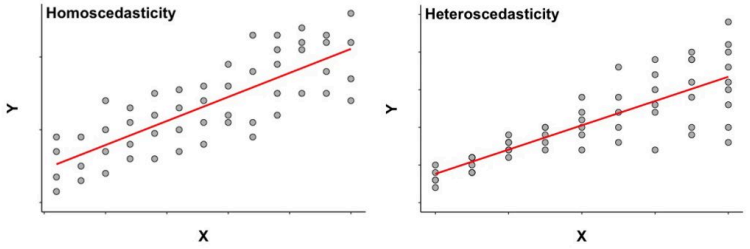


Figure 9.4. On the left, a scatterplot showing homoscedasticity: that is, the variance is similar at all levels of the variable X. On the right, a scatterplot showing heteroscedasticity; in this case, the variance of the observations for the lower values of the variable X is much smaller (the data points are tighter and closer to the line) than the variance for the higher values of the variable X.

When heteroscedasticity (that is, the variance is not constant but varies depending on the value of the predictor variable) occurs, the results of the analysis become less reliable because the underlying statistical procedures assume that homoscedasticity is true. Alternatives methods must be used when the data as heteroscedasticity is present.

When we have a linear regression model with just one predictor, it may be possible to see whether or not the constant variance assumption is met just looking at the scatterplot of X and Y. However, this is not so easy when we have multiple predictors, that is, when we are conducting a multiple linear regression. So, in general, in order to evaluate whether this assumption is met, once you fit the regression line to a set of data, you can generate a scatterplot that shows the fitted values of the model against the residuals of those fitted values. This plot is called a *residual by fit plot* or *residual by predicted plot* or, simply, ***residuals plot***.

In a residual plot, the x-axis depicts the predicted or fitted Y values ($\hat{Y}$s), whereas the y-axis depicts the residuals or errors, as you can see in Figure 9.5. If the assumption of constant

variance is met, the residuals will be randomly scattered around the center line of zero, with no obvious pattern; that is, the residuals will look like an unstructured cloud of points, with a mean around zero. If you see some different pattern, heteroscedasticity is present.

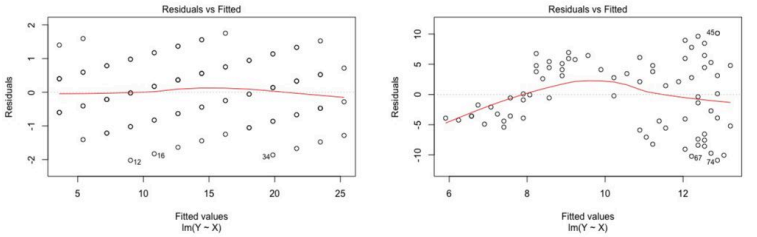


Figure 9.5. On the left, a residuals plot where the errors are evenly distributed without showing any pattern, an indication of homoscedasticity. On the right, a residuals plot where the errors show different patterns depending on the fitted values of Y, a sign heteroscedasticity.

4) Error or residuals are normally distributed

The distribution of the outcome variable for each value of the predictor variable should be normal; in other words, the error or residuals are normally distributed. The same as with the assumption of constant variance, it may be possible to visually identify whether this assumption is met by looking at the scatterplot of X and Y. In Figure 9.4 above, for example, you can see that in both scatterplots the distance between the actual values of Y and the predicted value of Y is quite evenly distributed for each value of the variable X, suggesting therefore a normal distribution of the errors on each value of X. So, note that, even if the constant variance assumption is not met, the residuals can still be normally distributed.

A normal quantile or quantile-quantile or **Q-Q plot** of all of the residuals is a good way to check this assumption. In a Q-Q plot, the y-axis depicts the ordered, observed, standardized, residuals. On the x-axis, the ordered theoretical residuals; that is, the expected residuals if the errors are normally distributed (see Figure 9.6). If the points on the plot form a fairly straight diagonal line, then the normality assumption is met.

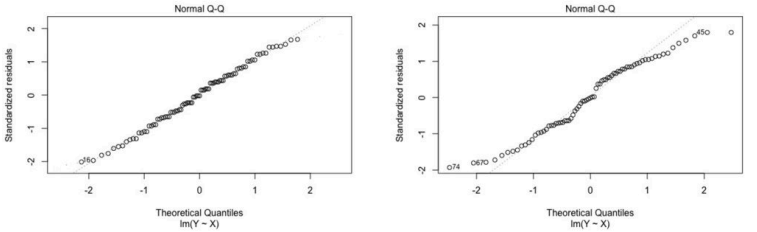


Figure 9.6. On the left, a Q-Q plot showing a normal distribution of the residuals, so that this assumption is met. On the right, a Q-Q plot showing a non-normal distribution of the residuals, so that this assumption is not met.

It is important to check that your data meet these four assumptions. But you should also know that regression is reasonably robust to the equal variance assumption. Moderate degrees of violation will not be problematic. Regression is also quite robust to the normality assumption. So, in reality, you only need to worry about severe violations.

Unit 10. Simple Linear Regression in R

LEYRE CASTRO

Summary. In this unit, we will explain how to conduct a simple linear regression analysis with R, and how to read and interpret the output that R provides.

Prerequisite Units
Unit 5. Statistics with R: Introduction and Descriptive Statistics
Unit 6. Brief Introduction to Statistical Significance
Unit 9. Simple Linear Regression

Simple linear regression analysis in R

Whenever we ask some statistical software to perform a linear regression, it uses the equations that we described above to find the best fit line, and then shows us the parameter estimates obtained. Let's see here how to conduct a simple linear regression analysis with R.

We will use the dataset of 50 participants with different amount of experience (from 1 to 16 weeks) in performing a computer task, and their accuracy (from 0 to 100% correct) in this task (the same dataset that we used in Unit 8). Thus, Experience is the predictor variable, and Accuracy is the outcome variable.

In the script below, the first line imports our dataset by using

the **read.table** function, and assigns the data file to the object *MyData*.

```
  #read data file
MyData <-
read.table("ExperienceAccuracy.txt", header
= TRUE, sep = ",")
#fit linear model
model <- lm(Accuracy ~ Experience, data =
MyData)
summary(model)
#see the residuals plot
plot(model)
```

To conduct the linear regression in which we will model Accuracy as a function of Experience, we use the **lm()** function in R, as you can see in the second line of code in the script. As usual in R, we assign our linear regression model to an object that we will call *model*. Remember that R works with objects, and that this name is arbitrary (you could call the object to which the linear regression model is assigned *seewhatIgot* or *mybestanalysis)*, although it is convenient to use a name relatively simple and relevant for your task at hand. Within the parenthesis, you first include your Y variable, Accuracy in this case, and then your X variable, Experience in this case, connected by the ~ symbol. This reads as "Accuracy as a function of Experience." Then, you indicate where your data are; here, you include the object that you created for the dataset that you are working with, MyData.

In the next line, you ask to see the output for the linear

regression analysis, using the **summary()** on the model. And this is what you will obtain:

```
  Call:
lm(formula = Accuracy ~ Experience, data =
MyData)Residuals:
Min      1Q  Median      3Q     Max
-8.7014 -2.4943  0.6176  2.7505  6.6814

Coefficients:
Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  57.7872     1.1115   51.99
<2e-16 ***
Experience    2.1276     0.1124   18.93
<2e-16 ***
—
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.705 on 48
degrees of freedom
Multiple R-squared:  0.8818, Adjusted R-
squared:  0.8794
F-statistic: 358.2 on 1 and 48 DF,  p-
value: < 2.2e-16
```

* **Call:** The first item shown in the output is the formula R used to fit the data (that is, the formula that you typed to request the linear regression analysis).

* **Residuals:** Here, the error or residuals are summarized. The smaller the residuals, the smaller the difference between the data that you obtained and the predictions of the model. The

Min and **Max** tell you the largest negative (below the regression line) and positive (above the regression line) errors. Given that our accuracy scale is from 0 to 100, the largest positive error being 6.68 and the largest negative error being -8.70 do not seem terribly large errors. Importantly, we can see that the median has a value of 0.61, very close to 0, and the first and third quartiles (**1Q** and **3Q**) are approximately the same distance from the center. Therefore, the distribution of residuals seems to be fairly symmetrical.

* **Coefficients:** This is the critical section in the output. For the intercept and for the predictor variable, you get an estimate that comes along with a standard error, a t-value, and the significance level.

The **Estimate** for the intercept or b_0 is the estimation of the analysis for the value of Y when X is 0; that is the predicted accuracy level (57.78% here) of someone who has 0 weeks of experience with the task. Remember that you should only interpret the intercept if zero is within or very close to the range of values for your predictor variable, and talking about a zero value for your predictor variable makes sense. This may be a bit tricky in our example. The minimum amount of experience with the computer task among our participants is 1 week, so the zero value seems to be close enough. However, realize that someone with zero experience with the task may not know what to do and may not be able to perform the task at all. Thus, the value for the intercept may not make sense. This is something that you have to evaluate when you know your methods and procedures well.

The **Estimate** for our predictor variable, Experience, appears below: 2.12. This is b_1 or the estimated slope coefficient or, simply, the slope for the linear regression. Remember that the slope tells us the amount of expected change in Y for each one-unit change in X. Here, we would say that for each additional week of experience with the task, accuracy is expected to improve 2.12 points.

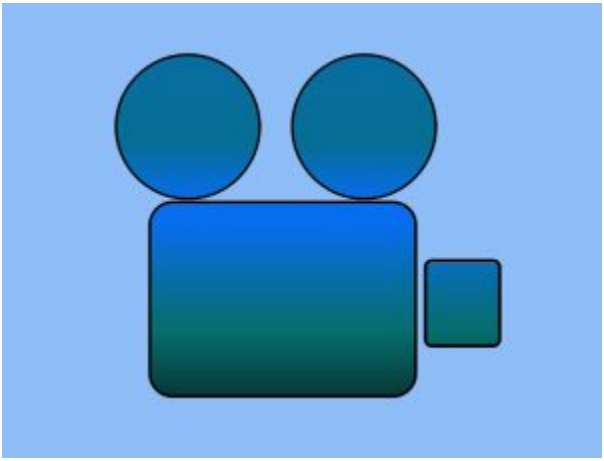
Pay attention to the sign of the estimate for the predictor. If it is positive, then it represents the expected **increase** in the outcome variable by each one-unit increase in the predictor variable. If it is negative, then it represents the expected **decrease** in the outcome variable by each one-unit increase in the predictor variable.

The **Std. Error** (standard error), tells you how precisely each of the estimates was measured. We want, ideally, a lower number relative to its coefficients. The standard error is important for calculating the t-value. The **t-value** is calculated by taking the estimate for the coefficient and dividing it by the standard error (for example, the t-value of 18.93 in the Experience line is 2.12 / 0.11. This t-value is then used to test whether or not the estimate for the coefficient is significantly different from zero. For example, if the coefficient b1 is not significantly different from zero, it means that the slope of the regression line is close to being flat, so changes in the predictor variable are not related to changes in the outcome variable.

Pr(>|t|) refers to the significance level on the *t*-test. Remember that the cut-off value for a hypothesis test to be statistically significant is 0.05 (see Unit 6), so that if the *p*-value is less than 0.05, then the result is statistically significant. Here, the *p*-value is very small, and R uses the scientific notation for very small quantities, and that's why you see the e in the number. You just need to know that this is a tiny value. For values larger than .001, the exact number will appear. The asterisks next to the *p*-values indicate the magnitude of the *p*-value (* for < .05, ** for < .01, and *** for < .001 as described in the **Signif. codes** line).

* At the bottom of the output, you have some measures that also help to evaluate the linear regression model. We have not seen yet what many of those elements mean so, for the time being, just note that R^2 (see Unit 7) is included here, indicating the amount of variance in Accuracy that is explained by Experience. R^2 always lies between 0 and 1. An R^2 of 0 means

that the predictor variable provides no information about the outcome variable, whereas an R^2 of 1 means that the predictor variable allows perfect prediction of the outcome variable, with every point of the scatterplot exactly on the regression line. Anything in between represents different levels of closeness of the scattered points around the regression line. Here, we obtained an R^2 of 0.88, so you could say that Experience explains 88% of the variance in Accuracy. The difference between multiple and adjusted R^2 is negligible in this case, given that we only have on predictor variable. Adjusted R^2 takes into account the number of predictor variables and is more useful in multiple regression analyses.



[See video with a simple linear regression analysis being conducted in R]

Glossary

LEYRE CASTRO

central tendency

Value, normally located around the center of a data distribution, that is most representative of the entire set of scores.

condensed scores

Also named *composite* or *scale* scores. These scores are created from multiple observation of different variables.

confidence interval (CI)

A confidence interval (CI) is as a range of plausible values for the population mean (or another population parameter such as a correlation), calculated from our sample data. A CI with a 95 percent confidence level has a 95 percent chance of capturing the population parameter.

continuous variable

A variable for which there are an infinite number of possible values between two end-points.

correlational analysis

Statistical technique that allows you to determine to what extent two variables are associated so that, when one changes the other one changes as well.

correlational study

A research study in which the researcher measures two or

more variables. The researcher does not set the values of any of the variables.

data

Set of observations representing the values of the variables under study (singular: datum; plural: data).

dependent variable

A variable whose properties, characteristics, or qualities are observed, measured, and recorded, as they occur.

descriptive statistics

Type of statistics used to organize and summarize the properties of a dataset.

discrete variable

A variable that consists of separate, indivisible categories. No values can exist between two neighboring categories.

experimental study

A research study in which the researcher manipulates one or more variables (independent variables) and then measures one or more other variables (dependent variables). In an experiment, there is at least one independent variable and at least one dependent variable.

external validity

Evaluation of how well the results of a study generalize to individuals, contexts, tasks, or situations beyond those in the study itself.

hypothesis

Tentative statement about how variables are related or

how one may cause another [singular: hypothesis; plural: hypotheses].

independent variable

A variable whose properties, characteristics, or qualities are entirely determined or set by the experimenter.

inferential statistics

Statistical analyses and techniques that are used to make inferences beyond what is observed in a given sample, and make decisions about what the data mean.

linear regression

Statistical technique that tries to capture a linear relationship between two (simple linear regression) or more (multiple linear regression) variables.

ordinal variable

A variable whose values specify a position in a sequence or list. It uses a number to do this, such as 1st, 2nd, 3rd, etc., but these numbers do not refer to an amount of something

outcome variable

The variable that we are trying to predict or understand in a regression analysis. Also called response or criterion variable.

outliers

An individual value in a dataset that is substantially different (larger or smaller) than the other values in the dataset.

parameter

A value (normally, a numerical value) that describes a population.

population

The entire set of individuals of interest for a given research question.

predictor variable

The variable that may be responsible for the values of the outcome variable in a regression analysis. Also called explanatory variable.

qualitative variable

A variable whose values do not represent an amount of something; rather, they represent a quality. Also called categorical variable.

quantitative variable

A variable whose values are indicated by numbers.

sample

A set of individuals selected from a population, typically intended to represent the population in a research study.

statistic

A value (normally, a numerical value) that describes a sample.

subject variable

A variable whose properties, characteristics, or qualities vary across research subjects, but are relatively stable

within subjects (across time) and/or are extremely difficult or impossible to manipulate by the experimenter.

summary scores

Scores created from multiple observations of the same variable under the same set of conditions.

tail

The end sections of a data distribution where the scores taper off.

theory

A statement or set of statements that describes general principles about how variables relate to one another.

variability

A quantitative measure of the differences among scores in a distribution. It describes to what extent the scores are clustered together or spread out.

variable

Each of the concepts, notions, dimensions, or elements that can be manipulated, measured, and/or analyzed in a research study.