

# Simple Stats Tools



# Simple Stats Tools

[Simple Stats Tools](#) Copyright © by . All Rights Reserved.

---

# Cover





---

**Title page**

**Simple Statistics Tools**  
**(for Sociologists)**





---

# Dedication

To my all my students, former, present, and future.



---

# Acknowledgements



---

## Preface

I have dedicated this book to my statistics students, former and future, all of them. Future, because it's all for them; they'll be the ones making use of it.

Former, because over the years they have been showing me (and, in many cases, telling me in no uncertain terms and with great emotion) how their first experience with statistics went. Because, somehow, along the way they have also taught me how to teach statistics to *them*. Not to a mass of generalized “undergraduate social science students in an introductory stats class,” with my initial preconceived idea of these students’ abilities, prior knowledge and needs, no — but to the actual *them*, the very real people I see in my classes. During the almost ten years of “SOCI 2365 Introduction to Social Research Statistics” instruction at Kwantlen Polytechnic University, I have learned how best to approach teaching stats to *my* students, in accordance to their actual academic needs and their actual academic abilities.

So who are the students in my classes? (Forgive me, now I'll have to generalize after all.) The typical student in my introductory stats class tends to be there because they have to (the course is compulsory for our major, along with a handful of others); is majoring sociology; is likely “not very good with math” and, therefore, has delayed taking the course as much as possible because, understandably, they are terrified. I could have used “she is” instead of the

gender-neutral “they are”— I typically have more female than male students. That is not to say that students not fitting this profile don’t take up my class; they do, and they’re not few. This example simply gives me the opportunity to give you a taste of what the book will be about: statistics and sociology.

See, “tends to”, “likely”, and “on average” are all terms with specific statistical meaning (as much as they can be misused and misinterpreted in conventional, everyday usage) — but you’ll have to go further into the book for that. However, I can easily tell you that I also have students, many of them, who are *not* majoring in the social sciences, are in their second year (as they are supposed to), are great with math, and who find the course easy. Of course, many of my students are also male. Obviously, none of what I just said contradicts the description of my typical student (and if it’s not obvious, you *definitely* need this book). The “typical student” description is simply based on a brief statistical profile of an average class I usually have. The various characteristics I listed may or may not be statistically associated with each other, not to mention anything about *causal* association. (Were you perhaps thinking that, say, women in my classes are the ones “not good with math” while men “find it easy”? I actually never said, not even implied, that. But now you see how easily statistical information can be misinterpreted and how statements based on statistical information can be taken to mean more than they actually do.)

Why sociology though? The description above can lead us to a few questions (i.e., we can formulate hypotheses), like, are students majoring sociology (or other social sciences, except economics) really more likely to say they

are “not good at math” than, say, students in the natural sciences? For that matter, are women on average more likely to major in social sciences and humanities than in the STEM (science, technology, engineering, and mathematics) fields? The answers to these questions can be found through statistical analysis (both are “yes” by the way) but the explanations (or theories) — i.e., *why* we observe the relationships between gender, major, and perceived math ability — are profoundly sociological.

In a similar vein, throughout this book I will bring up questions of sociological relevance, I will refer to sociological theories, research and findings, I will give sociological examples, and ultimately I will use sociological data.

Why does that matter? Stats is stats, right?.. Hmm, yes, and no — and in the case of applied statistics, as the current text is, rather no. Yes; if you go by the table of contents, you’ll see what one typically sees in a generic introductory statistics book (for social scientists); statistics is a set of tools, and it can be presented as generically and as generally as possible. However, like any tool, its value is higher the more specialized it is (you *can* take an ailing tooth out with a hammer yet arguably it’s better to use specialized dental equipment). Like any tool, it also matters what it is used for and how.

In other words, in this book the statistics instruction will be specialized: from a sociologist

(granted, herself specialized in social statistics) for sociologists. (If you are neither a sociology student or sociology instructor, you can take this as sort of a *caveat emptor* clause: buyer beware.) To the extent that sociology itself is a rather broad discipline and its use of statistics is equally as broad, one could use the book as an introduction to social science statistics. However, I do not go out of my way to engage in statistical instruments more frequently used in, say, criminology or psychology (i.e., small-size court case data, or experiment data, etc.).

I'll give you a different example: If you open an introductory psychology textbook, you will likely find a chapter on Sexuality and Gender. Yet "gender" and "sexuality" are also huge topics in sociology, and any introductory sociology textbook also has a chapter on them. There will be some overlap in the treatment of the topic by the two disciplines, but you'd be wrong to expect everything — or even most — to be the same.

Simply put, psychologists and sociologists generally tend to ask different questions, to approach a topic differently, to have different concerns, to have different preferred methods for collecting and analyzing (quantitative) data, and to even reach different conclusions, and to therefore offer different theories (as one would expect from two separate disciplines). Why wouldn't we want specialized statistics for each discipline?

Think of this book as a crash course in statistics. As such, I make these promises:

- 1) I promise to include only what is absolutely necessary.



2) I promise to skip on fluff and padding and any other material that is not strictly relevant to the exposition.

3) I promise to avoid repetitiveness as much as possible and instead explain everything only once but slowly and patiently.

Given my promise, this book provides a necessarily brief introduction to statistics. It is also a conventional introduction in that, as almost all such books, it does not include all there is about some of the more complex concepts, i.e., it is not entirely truthful.

Don't get alarmed by this admission. Rather, think of this introduction as your first date with statistics. No one tells all and bares all their secrets on a first date, do they? (...Or it might be their last.) Some things need to be revealed at a later time, once you've come to know your love interest better. Statistics is like that too. Some advanced concepts and relatively new developments in the discipline would only make sense to you only after the initial period of getting to know it has passed; then you can learn more "truthfully" and understand in what way and why the tools and concepts were simplified when they were first introduced to you.

And if you never get to "a second date" with statistics, never fear. What you will learn from this brief introduction will be quite practical "in real life" and still will serve you well. (You'll just know there is more to what you've learned — but that's the case with everything, no?) You will learn the basics of summarizing data and extracting useful information out of it; how data can be manipulated and how and why not to do that; how and when you can generalize from data and the limits to your generalizations;

what role probability and uncertainty play in statistics; how to interpret basic statistical information; what to look for in existing statistical reports; and how to execute a basic statistics report on your own. You will learn how to talk about statistics, and how to write about statistics. Finally, you will learn where to go from here, should you ever feel like going on a second date with statistics after all.

Given the purposefully streamlined content, some will not like this book. If you are an instructor (or a student) looking for theoretically comprehensive and expansive introductory treatment of statistics, this is not the book for you — but you also know many such books exist, freely available online or otherwise. Statisticians will likely be severely displeased by some of the things missing here, as compared to a truly conventional introductory statistics text.

But this indeed is why this book exists at all: to only include what I've discovered my students need in order to have a basic working knowledge about the most useful and most frequently used simple stats tools.

---

# Introduction

This book is intended to be your “first date” with statistics. It might end up as your *last* date with statistics too, so I’ll try to make the most of it while given the chance.

The book is organized as follows. Applied statistics is about data. Chapters 1 and 2 introduce you to concepts like variables and data sets and the type of information collected wherein, and generally cover all the preliminaries you need to know in order to start ‘doing’ statistics. Chapters 3 and 4 follow with the ways we can summarize and describe data. Altogether, this first part of the book is usually called *descriptive statistics*; it allows us to learn things from and about data that in many cases we cannot readily see just from looking at it.

I have devoted Chapters 5 and 6 to some theoretical concepts which are necessary to continue with the rest of the book, i.e, the part usually referred to as *inferential statistics*. You see, statistics would have a rather limited value if all it allowed us to do were to summarize or *describe* data (as useful as that is). The real power of statistics comes from *prediction* and *estimation* (i.e., *inference*), the subjects of the latter part of the book. In Chapters 7 through 10 you will learn how and why we can know things that go beyond the actual data we have; how likely they are and how confident we can be in this newfound knowledge; what it means for variables to be

statistically associated, and finally, whether we can identify causes and effects in the social world with any amount of certainty.

At this point, when promising all this to my students I usually feel like a charlatan at a county fair: *Come one, come all, I'll look at my crystal ball and the palm of your hand and tell you things I cannot possibly know*. After all, yes, alright, describing data you can see is one thing — but this *inference* thing?.. However, the more you learn about statistics and statistical tools and methods, the less (and less, and less) it will feel like charlatanry (I promise). Like many things in science, it only *looks like* charlatanry at first blush because you lack the knowledge of the principles that make the seemingly impossible, possible. In reality, what you will be learning in this book is not even all that complicated. If you don't believe me yet, check it yourself — just promise to go consecutively and patiently through all the parts until the end — no skipping!

So, ready to go?

---

# Chapter 1 Variables and Their Measurement

Naturally, we start with preliminaries. Before you learn the tools of any trade, you need to learn about your subject matter, i.e., on what you will be applying those tools. In this chapter I introduce you to the “building blocks” of statistics: the concept of variables and some related vocabulary. You will learn what variables are and about their levels of measurement (what nominal, ordinal, interval, and ratio scales are); how to determine the level of measurement of an actual existing variable and whether you should treat variables as discrete or as continuous for the purposes of statistical analysis.

Think of this chapter as the one establishing the main characters of a fictional story — the characters might seem too many at first, appearing too fast one after the other, so initially it might be hard to keep track of them and who is who and who does what. In time, however, the more you read about them (and sometimes going back to re-read key passages) they become familiar to you; then and only then you can comfortably follow their story.



---

You can think of a **variable as a characteristic that varies across individual elements**. For example, hair colour varies across individuals: black, blonde, brown, red, grey (or practically any colour if we include the wonders of hair dying). If we go by other physical characteristics, we can easily see that height, weight, body type, skin colour, age, etc. are all *variables*.

Then what about social/economic characteristics like level of education, annual income, occupation, employment, citizenship, marital status, political party affiliation, union membership, participation in sports (to name a few)...? All variables. Or, what about personal opinions and preferences? You might love chocolate a lot but your friend might not care for it; another friend might like it but just a little... Your friend might try to convince you that classical music is great but you might find it terribly boring, preferring rock instead. You might be a dog person and might frequently extol the virtues of dogs in comparison to cats, to the dismay of your cat-loving significant other. You might think that legalizing marijuana in Canada was the right decision but your parents might feel it was a profound mistake on part of the government. Clearly, opinions and preferences vary, so we can add ‘opinion on marijuana legalization’, ‘liking of chocolate’, ‘preferred music genre to listen to’, and ‘favourite pet animal’ to our ever growing list of variables.

So far, you might decide that variables only apply to

*people*: after all, all the examples mentioned above discuss characteristics that vary across human beings. However, this is absolutely not the case, as we can just as easily see that other things can have varying characteristics. For example, *universities* can differ in their student enrollment numbers, instructor-to-student ratios, type of degrees awarded, geographical location, source of funding, presence of medical school, percentage of international students, etc. *Countries* vary on population size, climate, geographical/geopolitical location, language, GDP (gross domestic product), level of human development, presence of minority groups, immigration (and emigration) rates, fertility and mortality rates, access to universal healthcare, average education level, age of majority, freedom of press, type of government... you get the picture. Clearly, variables apply to *elements* of anything that may be compared on characteristics which vary across these elements (hence the somewhat clumsy definition I started with).

Researchers refer to ***units of analysis*** when they want to specify the elements they study: When we have information about characteristics of *people*, we say that the unit of analysis is “individual”. When instead of people, we study *countries*, the unit of analysis is “country”, and so on.



---

## 1.2 Concepts, Measurement, and Operationalization

You might be wondering why we even need to introduce a concept such as variables. Can't we simply call them *characteristics*, if that's what they are? The short answer is that we use the language of variables when we engage in formal research, but the reason is not solely scientific jargon. *Variables*, as opposed to *characteristics*, imply measurement.

You see, sociologists and other social scientists study *concepts* (i.e., ideas, notions) that are more often than not abstract. If I say "I want to know if the average height of Canadians has changed over time", it's easy for you to suggest that I first collect information about people's heights (perhaps actually measure them, if I don't trust self-reports). By doing that, you might not realize it but what you have done is actually offer *a way to measure a concept*, which is what we call with the mouthful of a word **operationalization**. In other words, you have *operationalized* the abstract concept (height of Canadians) through the actual, physical measurement of individuals' heights (in centimeters or in inches) in real life.

So operationalization is that easy, right? Unfortunately, no, not really.

What if, instead of average height of Canadians, I had

wanted to study how poverty has changed in Canada over time? Or homelessness? How about income? Or people's attitudes to immigration? Or their religiosity? What about if I wanted to study self-esteem of adolescents? Or social status among Canadian university students? Or bullying in high school?

I'm sure you have no trouble understanding the concepts as *abstract ideas* — but how do you *measure* them? <sup>1</sup>There are various ways one can measure concepts. At the most fundamental level, this depends on what the chosen method of inquiry (or, research) is, *qualitative* or *quantitative*. We shouldn't reify the boundary between quantitative and qualitative methods, however. Many scientists mix their methods, employing both methods in a single study with considerable success. Social scientists use statistics predominantly when they have chosen a quantitative method of collecting and analyzing data, so here we'll focus on the quantitative operationalization of concepts.

#### *Do it! 1.1 Measuring Homelessness*

Imagine you really do want to study the prevalence of homelessness in your city (or any of the abstract concepts mentioned above). Before you decide how to collect

information about it, you have to choose about what *exactly* you will be gathering information. How are you going to define *being homeless* in order to measure homelessness? In a word, how are you going to *operationalize* homelessness? Make a list of possible definitions. What are the various aspects of homelessness, which you may choose to consider in your definition or not, that make defining homelessness difficult?

All in all, operationalizing a concept boils down to choosing a working (i.e., operational), *measurable* definition of a concept within a given study. Most concepts can be (and regularly are) defined differently by different researchers. What matters is that the definition of any concept is provided and is used consistently within each individual study.

If the *Do It!* exercise above seems too abstract still, perhaps one easier way to understand the operationalization of concepts into measurable variables with concrete definitions is to imagine a survey question about what you want to study. Sometimes one such question can provide the operationalization/definition of the concept under study. Other times a single question is not enough and a set of questions can help a researcher measure what they want to study.

Let's say you want to study *income* (perhaps as a part of a larger study on poverty). You want to ask people about their income but how exactly? Will you be asking about personal or household income? Are there types of income you have in mind — from salary, from rent, from

interest, etc.? Is it weekly, bi-monthly, or annual income? Is it income *before* or *after* taxes? For that matter, do you mean only taxable income? Furthermore, what kind of answers would you accept? Will the respondents provide an exact number? Or will you provide a set of multiple-choice answers from which the respondents will choose?

For example, you can measure income in a hypothetical study (through a survey question) like this: “What is your household’s annual after-tax income (from any source)?” This means that you have chosen to operationalize the abstract concept *income* through the specific, measurable variable *annual household after-tax income*.

The types of possible answers you choose to accept for the question are also part of the measurement. Example 1.1. below offers three options to operationalize income.

#### *Example 1.1. Operationalizing Income*

Q1. What is your household’s annual after-tax income (from any source)?

- a) \$0 – \$50,000;
- b) \$50,001 – \$100,000;
- c) \$100,001 – \$150,000;
- d) \$150,001 – \$200,000;
- e) \$200,001 or more;

Q2. Is your annual household after-tax income (from any source) less than \$50,000?

- a) Yes;
- b) No.

Q3. What is your annual household after-tax income (from any source)?

.... [Any number provided by the respondent will be recorded.]

The multiple choices provided in *Q1* in the example above can contain any number of categories to choose from. I have chosen to go by 50 thousand dollars to create the categories, but I could have done so by as little as, say, five thousand dollars to as much as 500 thousand dollars (and I would have ended with a different number of possible answers). If we need the actual dollar amount of the income reported by each respondent, we'll chose to ask *Q3*.

The way we choose to create categories or not depends on the type of answers that will be suitable for our study and what type of information we want. As well, *Q2* offers only two possible answers, yes or no. If the relevant information for our study is whether household annual income is below or above \$50,000 (say, because the average such income has already been established as \$50,000), *Q2* would be the way to go.

Keep in mind that how a variable is operationalized depends not only on the researcher's goals and needs (and

practical considerations like time and money) — but also on their personal beliefs and preferences, the time period in which they live/d and work/ed, etc. Operationalizing concepts considered controversial at a specific time and place can be quite political and itself become a controversy. Consider the following example.

*Example 1.2. Operationalizing Gender*

It should come as not surprise to anyone studying sociology that how people operationalize gender has changed over time. Until recently, the conventional operationalization went something like this:

Q1. Are you...?

- a) Male
- b) Female

With advances in the study of gender and sexuality, over time our understanding of gender changed. Nowadays you are far more likely to see an operationalization similar to the following style of the American Sociological Association when collecting information on their members:

Q2. What is your gender? Select up to two.

- a) Female
- b) Male
- c) Transgender female/Transgender woman
- d) Transgender male/Transgender man
- e) Gender queer/Gender non-conforming

- f) Different identity (please specify) .....
- g) Prefer not to state

In countries like Canada, using *Q1* nowadays would might be considered too restrictive for many purposes, and also offensive by some. On the other hand, in some countries (like in Eastern Europe) choosing to go with *Q2* might be seen as quite controversial and as political activism. Even in Canada the switch to more inclusive gender oprationalization is gradual and quite recent. As you will see later in the book, datasets collected in the past typically use a binary operationalization of gender.

Before we continue with measurement in the next section, here is a practical tip when working with SPSS.

*SPSS Tip 1.1. Exploring How Variables in a Dataset Have Been Operationalized*

When exploring an existing dataset in SPSS (more on that in Chapter 2), you can see a variable's categories/values in the *Values* column in *Data View*. (You can switch between *Data View* and *Variable View* by clicking on their respective tabs at the bottom of your primary data window.) Clicking on a variable's cell in the *Values* column will open a new window listing all the categories/values through which the variable has been operationalized.





---

## 1.3 Levels of Measurement

Now that you know there are different ways to operationalize concepts, let me introduce another term in respect to variables: *level of measurement*. Each and every variable has a level of measurement. Knowing, or being able to identify, the level of measurement of a variable tells us how it has been operationalized and vice versa: knowing how an existing variable has been operationalized gives us information about its level of measurement.

More importantly, however, knowing and being able to identify **a variables's level of measurement allows us to determine what we can do with that variable in terms of statistical methods and procedures**. This last point is key to doing statistical analysis in a correct and meaningful way. The flip side is also true: misidentifying a variable's level of measurement will inevitably end in erroneous analysis and conclusions (that is, if the analysis can even be performed, as in many cases the statistical software will give an error message).<sup>1</sup>

Why is the level of measurement so important for statistical analysis?

Simply put, variables are not created equal when it

1. The more dangerous -- and quite frequent -- scenario, however, is when the software will execute the analysis and produce results. In that case, without an error message to warn them, the researchers would trust their analysis and results without realizing both are bogus.

comes to levels of measurement. Due to differences in the nature of the information contained within, you can do very little with some variables in terms of analysis while you can do a whole lot more with others.

*Do it! 1.2. Measuring Different Types of Variables*

Imagine you have to analyze the following (individual-level) variables:

- a) religious affiliation,
- b) educational attainment,
- c) exam test scores,
- d) age.

Think of what type of information would be contained within the categories of each of the four variables above. (It might help to imagine the possible answers respondents — say, university students — could give if asked questionnaire questions about each.)

What more (beyond collecting it), if anything, can you do with that information? For example, can you say that one answer is more/bigger than another? Can you identify answers as different or the same as others? Can you do some calculations with the answers?

The exercise above gives you a clue: **there are**

**four levels of measurement.** They are called *nominal*, *ordinal*, *interval*, and *ratio*. Each and every variable has only one level of measurement once it's operationalized.<sup>2</sup> A variable's level of measurement is sometimes also called its measurement *scale*.

The following sub-sections provide details about each measurement scales.

2. Recall, however, that sometimes -- though not always -- one and the same variable can be operationalized in different ways. These different ways can sometimes be at different levels of measurement, depending on the type of information we want to have.



---

### 1.3.1 Nominal Variables

As the name of this level of measurement implies, the information contained in the categories of a nominal-scale variable is solely their... well, *name*. Think about the *religious affiliation* variable from the *Do it! 1.2.* exercise. You have already probably imagined people's possible affiliations in terms of religion (i.e., what religion they subscribe to, if any) as something like *Muslim, Jewish, Christian, Sikh, Hindu, Buddhist, not religious* — though likely (and depending on your own religious affiliation) *not in this particular order*.<sup>1</sup>

Of course, I could have just as easily listed the possible categories (or “questionnaire answers”) as *Christian, Muslim, Jewish, Buddhist, Hindu, Sikh, not religious*. Or, as *Sikh, not religious, Buddhist, Hindu, Jewish, Muslim, Christian*. Or, as... virtually any possible variation in the ordering of the list.

In other words, the information we have about religious affiliation is simply in *identifying* the different categories, and that is *all*. We cannot do much more than count the different answers and specify what they are. We cannot

1. It's also likely that these general categories might have been

**disaggregated** to list variations/denominations, e.g. *Catholic* and *Protestant* instead of simply *Christian*, or *Shia* and *Sunni* instead of simply *Muslim*, etc. For simplicity's sake, I choose to use the most general religious categories in the example.

even use some inherent order to them, as they are only that, *names*.<sup>2</sup>

When researchers study religious affiliation in real life, they usually list the groups' names by the size of the religious group/popularity of a religion in their area. For example, in the Americas and Europe the listings usually start with *Christian*. In India, one can arguably assume they start with *Hindu*, etc. This type of ordering by size is still *purposefully imposed, not an inherent one*.

### *Do it! 1.3. Nominal Variables*

Try to come up with at least three different nominal variables. Can you explain why they are nominal? Try to defend your choice in identifying the scale for these variables as nominal.

2. Of course, we could order the categories alphabetically -- just like you can order pretty much *anything* alphabetically. That would be an arbitrary decision, however, not an *inherent* order contained in the names (like that in small to big, left to right, slow to fast, less to more, etc.).

---

### 1.3.2 Ordinal Variables

As with the nominal scale, the name of this scale is indicative of its defining feature: an order. That is, the categories of an ordinal variable cannot just be ordered arbitrarily in any other way, like we can with nominal variables, no: **the categories of any ordinal variable have an inherent order to them.** Listing the categories of an ordinal variable differently would violate the intrinsic logic of their order and would make little to no sense; as well, we would lose the information contained in their order.

Think back to the variable *educational attainment* from the *Do It!* 1.2. exercise earlier. *Educational attainment* is usually measured by the educational degrees attained by an individual, so if you imagined the categories being something like *no degree*, *secondary/high school*, *Associate's*, *Bachelor's*, *Master's*, *doctorate/PhD* you are probably not alone. That is, chances are, most, if not everyone, would come up with a list *in that particular order*. Why? Because, I can hear you explaining, no degree is *the lowest* formal educational attainment one can have; it's clearly *less* than having finished secondary/high school, which in turn is *less* than having a college degree, which again is clearly *less* than achieving a Master's degree, while, finally, a PhD is the highest degree one can get in academia. Arbitrarily switching the categories in *educational*

*attainment* to be listed as, say, *Associate's*, *Master's*, *no degree*, *PhD*, *Bachelor's* makes little (rather, no) sense, and worse, it deprives us of the information about there being an intrinsically ascending order in the obtaining of the degrees (as one can only have a doctorate if they had previously finished college, which can only be done after secondary/high school).

Note that having an intrinsic order (in this case, from less to more), however, is a necessary but not a sufficient condition to identify an ordinal scale. **There is an additional requirement: a variable is ordinal only when the categories do not have a precise (numerical) value.** In other words, while we know that a Bachelor's degree is *more* than an Associate's degree, we don't know *how much* more. Having a PhD is more than a Master's degree, but again, we don't know by how much. The same goes for any of the categories. We know the order, but not the precise "distance" between one category and another. As well, the "distance" between the first category and the second one might be unequal (while still unknown) to the "distance" between the second category and the third, and so on. It is not the size of the distance that matters here, only that the distance exists and that a category is clearly less/more (or bigger/smaller, nearer/farther, etc.) than another.<sup>1</sup>

1. You might be tempted to measure the "distance" between the categories in *educational attainment* in terms of years. For example, you could say that the "distance" between *secondary/high school* and *Associate's* is two years, or that between *Associate's* and *Bachelor's* is another two years, etc. This would still be an imprecise measurement, however, because different people take different times to accomplish their degrees, not to mention that there is no way to measure the difference between *no degree* and *secondary/high school* (as *no degree* can mean anything between no



To summarize: As you can see from this example, the key feature of ordinal variables is the intrinsic logical ordering of their categories, a logic that would be lost if we were to reorder them in any other way. As well, this tells you that ordinal variables contain more information in comparison to nominal variables: namely, the ordering of the ordinal variable's categories. Ordering the categories of a variable is an additional action you can do above simply listing them. Finally, the general order is the *only* additional information: the “distances” between the categories could vary and should not be measurable/ quantifiable. If the latter is not the case, you are already moving into interval/ ratio scales territory.

#### *Do It! 1.4. Ordinal Variables*

With the risk of being repetitive, I'll ask that you try to think of three different ordinal variables. Can you explain why they should be classified as ordinal? Remember to make sure that the internal logical ordering of the categories of your variables is of the “more/less” type rather than involving precise measurement.

education -- still a sad reality in many countries -- to dropping out of school a year before graduation. As well, doctoral studies vary enormously in duration depending not only on the chosen discipline but also on the country, etc. In short, measuring the “distance” in educational attainment categories in years would vary far too much on a case by case basis to be meaningful in any way. Note, however, that you could operationalize a variable *years of schooling* measured in years but that would not be the same variable anymore (nor would it be an ordinal variable).



---

### 1.3.3 Interval and Ratio Variables

Going back to the original *Do It!* 1.2. exercise, I am sure that you found imagining the categories of *exam test scores* and *age* the easiest, as they would be simply numbers. Perhaps something like 30, 65, 72, 88, 95, etc.... points out of 100 in the former case (though I know you don't want to imagine a test score of 30 on any exam!), and, if we're imagining college students, something like 18, 19, 20, 22, 23, 24, etc.... years in the latter. Notice the major difference from the categories of the nominal and ordinal variables we discussed above: now we are working with *numbers*. Not only are the *exam scores* and *age* categories comprised of numbers (as opposed to words) but they are also ordered in *measurable* “distances”. In other words, **there is a stable/unchangeable unit by which the “distance” between any two categories can be measured: a point in the exam scores case and a year in the age case. This unit is called unit of measurement for the interval and ratio variables.**

Wait a second, you're probably thinking now — the *exam scores* above lists 30, 65, 72... as categories, and a quick calculation reveals that the “distance” between 30 and 65 is thirty-five points, while the “distance” between 65 and 72 is only seven points. Thirty-five is clearly much bigger (five times bigger to be precise) than seven: isn't that as arbitrary as the “distances” across the *educational attainment* categories above? Well, no. The difference is that for interval and ratio variables the information

contained in the categories and their “distances” from each other is not simply of the more/less, bigger/smaller, left/right, etc. kind but is readily quantifiable and measurable in precise, stable units. In practical terms, you can specify *exactly how much* smaller/bigger a category is than another (i.e., 65 points is thirty-five points more than 30 points; a 22 years-old is two years older than a 20 year-old) — unlike with ordinal variables, where we know a Bachelor’s is a bigger educational attainment than secondary/high school but there is no agreed unit to measure the “distance” precisely (as it’s neither measured in years, not in numbers of degrees).

Furthermore, my *exam scores* example lists 30, 65, 72... but I simply chose these numbers at random: I could have just as easily listed 25, 45, 70..., or 12, 54, 69..., etc. The point here is that one can have *any* number between 0 and 100 (in a conventional 100-point exam) as a *potential* score, or be a college student of *any* potential age (say, more than 5 years old),<sup>1</sup> while the categories of an ordinal variable are *fixed*, or set, during operationalization (to a usually relatively small number), and cannot potentially be anything else (unless you operationalize the variable in a different way, which would result in a new variable).

Finally, a happy corollary to the fact that interval and

1. You might think I'm joking but do look Michael Kearney up. He graduated high school at age 6 and had earned his Bachelor's degree at age 10, this making him the youngest university graduate on record. (\*January 15, 1995|**RICHARD KAHLENBERG** | The LA Times)

ratio variables' categories are comprised of numbers is our ability to perform mathematical operations on them, beyond simple comparisons — something we can do neither with nominal, nor with ordinal variables. (Exactly what kind of mathematical operations we can do with interval and ratio variables you'll see in Chapter 2.)

**To summarize, interval and ratio variables have three defining features: 1) their categories (typically called *values*) are comprised of numbers, 2) the categories follow an order inherent in the fact that there is a measurable, unit-based scale, so that we can speak of a variable's units of measurement, and 3) we can perform mathematical operations on the values (that the categories are).**

Wait though... Why did I say that interval and ratio variables are different when I keep defining them together, and in the same way? Not to worry: the difference comes next, as I saved what students usually find the trickiest part for last.

With the risk of oversimplification (and, inevitably, exaggeration), interval scales are “made-up” while ratio scales are “real”. The difference is purely conceptual: you have to know whether the scale on which the variable is measured is “artificially designed”, as it were, or whether it exists as a some sort of “objective reality”. A rule-of-thumb advise on differentiating them that you may encounter is the “existence of a true zero”: **ratio variables have a *true zero* while interval variables do not.** (Clear as mud, eh? I did say it's tricky.)

Examples usually help make this conundrum seem less of a conundrum.

### *Example 1.3. Interval Variables: Temperature*

Let's take the classic example of an interval-scale variable, *temperature*. If you go by centigrade,  $0^{\circ}\text{C}$  is, I'm sure you know, the temperature at which water freezes. If you go by Fahrenheit, however,  $0^{\circ}\text{F}$  is... well, nothing in particular; it's just equal to about  $-18^{\circ}\text{C}$ . On the other hand, if you are more scientifically-minded, you might go by Kelvin, where  $0^{\circ}\text{K}$  is the coldest-cold-and-nothing-could-ever-be-colder temperature (a.k.a., *absolute zero*), equal to  $-273.15^{\circ}\text{C}$ , or  $-459.67^{\circ}\text{F}$ .

Have you ever wondered why there are three scales of measuring temperature? From where did they come from? They were “artificially designed” (or you might say, invented) by people: Anders Celsius, Daniel Fahrenheit, and Lord Kelvin were the scientists who came up with them and whose names we use to indicate in which scale we have chosen to report temperature. Not only is a temperature of 0 degrees different in all three systems, *they don't indicate zero/nothing/absence of something.*<sup>2</sup> *does* indicate absence

of all energy, a temperature where all atoms stop moving, but it is still not an absence of *temperature*. Temperatures of 0°C or 0°F do not indicate an *absence* of temperature or *no* temperature whatsoever, they are purposefully (and one could say, arbitrarily) chosen by people as a zero-point on an human-made scale.

In a similar vein, a score of 0 points on an exam doesn't typically mean a complete absence of or no knowledge on a subject whatsoever — such a score usually simply means that the test-taker did not perform well on *that particular test*. Arguably, an easier test on the subject could be designed, and the test-taker would likely score more points.

Contrast this to our other variable from the original *Do it!* 1.2. exercise, *age*. Age of 0 years means exactly that — that we are talking about an infant who hasn't yet reached their first birthday, and thus has completed 0 years of life (pardon the awkward phrasing).<sup>3</sup>

3. Of course, we measure babies' ages in smaller units, like months, or weeks, or even days and hours -- just like we can measure any person's precise age that way. However, we *usually* don't do it for anyone who's not an infant, so I'll leave it at that. Or consider a variable for, say, income: an income of \$0 means the complete absence of income on dollars, i.e., *no income*. Both age and income are not “made up”: they exist regardless of how we measure them, and a zero on either indicates an absence of something (*time* in the former case, *dollars* in the latter). Physical attributes like height and weight work the same way.

*Do It! 1.5. Interval/Ratio Variables*

You saw it coming: Try to come up with three interval/ratio variables (in addition to the ones I listed above). Try to differentiate between the interval and ratio scales and to identify which variable goes with each. Make sure you can explain what makes each variable interval- or ratio-scale.



---

## 1.4 Level of Measurement and Operationalization Considerations

All in all, the difference between interval and ratio variables exists more on a conceptual level rather than in practical terms. As such, they are frequently grouped together in an interval/ratio category and treated the same for the purposes of statistical analysis. At this stage, while it's preferable to know the difference between them, it is still far more important to be able to differentiate interval/ratio variables from nominal and ordinal ones.

Here is proof how tricky identifying the correct level of measurement of a variable can be.

### *Watch Out!! #1 ... for Likert Scales*

Most likely, at some point you have encountered survey questions that read something like this:

“On a scale of 1 to 5, where 1 is the lowest and 5 is the highest, how much do you like ...?”

... let's say, "chocolate". It is possible that you were presented with the numbers from 1 to 5 to choose from, or that they were accompanied with phrasing of the *strongly dislike*, *dislike*, *neither like nor dislike*, *like*, *strongly like* type. Now that you know about levels of measurement, as what scale would you classify the variable *liking of chocolate*: nominal, ordinal, or interval/ratio?

Considering that the answers from which one can choose are listed as numbers, many students are tempted to classify such a variable as interval. However, the *strongly dislike*, *dislike*, *neither like nor dislike*, *like*, *strongly like* part should give you more clues. Ask yourself: is there a uniform unit that allows us to precisely measure the "distance" between *dislike* and *strongly dislike*? Or between *like* and *neither like nor dislike*? Is it even the same "distance"? We would be hard-pressed to say "yes" to any of these questions. We know that people who like chocolate like it more than those who neither like it nor dislike it but we don't know *exactly how much* more. The numbers are there to make analyzing the responses easier, and as a sort of "code" for the ranking of preferences regarding chocolate, but substantively the ranking contains only order, not precise measurement of these preferences.

Variables such as these are called **Likert scales**. As I just explained, they are ordinal by constitution (although, in some special cases — for example, when the possible responses are not five but, say, ten or more

— they can be *treated* as interval for purposes of analysis). Researchers use them usually to capture people’s preferences — but preferences are generally “fuzzy” and not fully-defined; they do not come with a build-in, measurable, uniform unit scale, despite the fact that it seems like the numbers represent one such scale.

In Chapter 2 you will see that numbers can be used to represent a lot more than actual numbers. (And you were just starting to think identifying the level of measurement is easy!)

A further word of caution: the examples I used in this chapter might leave you with the impression that you can simply *hear* the name of a variable and you should be able to identify its scale of measurement. That would be wrong. My examples are *hypothetical* and as such I *imagine* what the variables’ categories might look like. (I also ask you to imagine variables and their categories in the *Do It!* exercises.) However, variables — not hypothetical, *real* variables that we use for analysis — exist in real datasets, where they have been operationalized in one specific, concrete way.

As such, upon hearing the name of a variable, instead of *imagining* what it looks like, you should always – *always!* – *actually look* at it and its categories in the given/specific dataset of which the variable is a part. **Determining an existing variable’s scale of measurement requires**

**exploring the actual variable as it was created.** Recall that there is more than one way to operationalize a variable. Thus, the researcher/s who created some variable into which you might be looking might arguably have created it differently than you would, or differently than some other researchers might have created theirs — *even if these variables* (the different researchers' and your hypothetical one) *have the same name*.

This leads us to the question: **Can the same concept be operationalized at different levels of measurement?** The answer lies in the nature of the concept (or that of the hypothetical variable, if you prefer). Let's go back to the example of *income* from the previous section on operationalization. There I provided you with a few different ways to create income categories. One was based on a yes/no question ("Is your income below...?" a specific number), and few more ways listed several categories based on income groups ("0-19,999", "20,000-29,999",.....etc.). Additionally, we could ask people to supply their specific income, rounded to the nearest dollar. Alternatively, thinking along the lines of a survey questions, this would result in a) yes/no response, b) multiple choice answer, and lastly, c) an open-ended, respondent-supplied answer.

In this way, we can say that we can successfully operationalize the concept of *income* at three different levels of measurement: a) nominal, b) ordinal, and c) ratio, respectively. This is only possible because of the *numerical* nature of income: income is monetary, and money is countable – and expressed in *numbers*. We can *choose* to create several categories of income (out of the numbers involved), or we could *choose* to create only a binary

variable (i.e., with two categories) to indicate an income below/above some threshold. In choosing either of these, we also make the decision to forego, or lose the more specific information of the actual income of everyone we ask. Logically though, we can only forego/lose information that is otherwise potentially available: we cannot *make* information *up*.

What it all boils down to is that **we can operationalize down: from the highest level of measurement possible for a variable towards the lower ones – but never vice versa**. A concept of numerical nature, i.e., an interval/ratio variable can be operationalized *down* and created as an ordinal variable, or even further down as a nominal variable, losing potential information (actual numbers and order) along the way. A concept of ordinal nature can also be operationalized down to a nominal scale, again, foregoing the potential information of order. However, a “naturally” nominal variable cannot be operationalized as anything else but nominal: there is simply no further information available. The same goes for “naturally” ordinal variables – they cannot be operationalized as interval/ratio as the only information we can have is order, while precision and measurable, defined constant units are not possible to obtain.<sup>1</sup>

1. Beyond the original operationalization, sometimes researchers actually *recode* variables down within an existing dataset. Since they start with an interval/ratio variable, they can choose which level of measurement they want to use, and go back and forth between ordinal and nominal and back to interval/ratio. They can do this only because the information has initially been collected at interval/ratio level of detail. If the original information is collected as nominal or ordinal data, no further information cannot be accessed: *recoding up is impossible*.



---

## 1.5 Discrete and Continuous Variables

I will introduce a final useful typology by which variables can be grouped: discrete and continuous.

By definition, variables called *discrete* (note, not discreet!) have finite number of categories (i.e., “space” between them, and nothing occupies that space), while variables called *continuous* have potentially infinite number of values (i.e., it’s possible that a value exists between any two given values, in smaller and smaller — *infinite* — number of “spaces” between any two the values, to infinity). To make things easier to understand, and with more than a little risk of oversimplification, **in a very broad sense you can think of nominal and ordinal variables as discrete and of interval/ratio variables as continuous.**<sup>1</sup> For example, *hair colour*, *religious affiliation*, and *educational attainment* (as measured in educational degrees) are all discrete: they have finite number of *discrete* categories.

On the other hand, age, income, or exam scores are all continuous: a number (value) can exist between any two given values, depending on how precise you want your measurement to be. To take *age*, for example, if two people report being 20 and 22, respectively, it’s obviously possible that another person is 21. However, we need not

1. Technically speaking, in theory nominal and some ordinal variables are categorical, ordinal variables with numerical categories are discrete, and interval/ratio variables are continuous. In practice, things are less clear cut.

round to full years; between two people ages 20 and 21, a value of 21.5 (or 21 years and 6 months) is possible to exist. Further, between the ages of 21 years and 21 years and 6 months, we can have a value of 21 years and 3 months, and so on, until we are down to counting days, then counting hours, then counting minutes, then counting seconds, then milliseconds, then microseconds, then nanoseconds, etc.... The point is that, in theory, there is always a smaller number between any two numbers (which can be represented by the possibility of infinite number of digits after the decimal point). The same can be applied to income and exam scores too.

In practice, however, things are different. In sociological research (as with other similar disciplines), the data collected is *empirically* discrete, as the values collected are a finite number and are typically rounded to whole numbers: we don't bother to measure age in anything but years, income in dollars (and not cents), etc. Still, we usually call interval/ratio variables are continuous, because of the *potential* for infinite number of values.

At the same time, however, some ratio variables are truly discrete. Think, for example, about a measure called *number of children* of the respondent. Clearly, there is no possibility for an infinite number of values, just like with any "number of people"-type variable: people can only be counted in whole numbers, and the count is always finite.

All this is undoubtedly confusing, so here is a practical tip for applied research, and what you need to focus on. Regardless if a variable is discrete or continuous *in theory*,



in practice all variables you will encounter in real-life, actual datasets will be discrete. **What we do is *treat some variables as discrete, and other variables as continuous for the purposes of statistical analysis***. The rule of thumb is to make the differentiation based on the number of categories/values: ***typically nominal and ordinal variables have relatively few categories so we treat them as discrete, while interval/ratio variables typically have relatively large number of values, so we treat them as continuous***. If, however, an ordinal variable has relatively large number of categories it may be treated as continuous, and, on the flip side, if an interval/ratio variable has relatively few values it may be treated as discrete. Generally, and assuming proper justification (i.e., a large number of categories/values), the decision to treat an ordinal variable as continuous or an interval/ratio variable as discrete remains a matter of the researcher's discretion.

Finally, what is the magic number in the “relatively large number of categories/values” rule? This also depends, but from what I have seen in practice, the number is around 7-10 categories/values for most (i.e., if a variable has more categories/values that that it's treated as continuous, and if it has fewer categories/values than that it is treated as discrete).



---

## 1.6 Creating Variables

If you ever find yourself in need of creating your own variables (perhaps, in creating a questionnaire), this brief final note is for you. As well, you can learn to evaluate whether an existing variable has been operationalized properly.

**To properly create a variable, its categories need to satisfy two requirements: they need to be collectively exhaustive and mutually exclusive.** The first condition, *collectively exhaustive*, refers to the requirement that the categories cover all possible ways the variable can vary (or all possible answers to a questionnaire question) — none can be excluded. The second condition, *mutually exclusive*, adds the logical necessity that a specific variation (or an answer to a questionnaire question) can exist in one and only one category.

This is simpler than the definition makes it sound to be. The following example illustrates.

*Example 1.4. Logical Requirements to Operationalizing Variables*

Imagine you are filling out a questionnaire and one of the questions is about age, like this:

Q1. What is your age?

- a) 20-29
- b) 30-39
- c) 40-49
- d) 50-59

What if you are 18 or 19? Which answer would you pick? How about if the person filling out the questionnaire is 60 or older? As stated, the Q1 question (i.e., the way the variable *age* is operationalized by it) violates the first requirement, that of providing an exhaustive list of all possibilities. All possible variations need to be covered by the variable's categories, otherwise the variable is incomplete.

Now consider another hypothetical way to ask the same questionnaire question:

Q2. What is your age?

- a) 18-25
- b) 25-30
- c) 30-35
- d) 35-40
- e) 40-45
- f) 45-50
- g) 50+

Assuming the questionnaire is administered only to adults, Q2 provides a collectively exhaustive list of possible answers; the variable's categories are too collectively exhaustive.

They are, however, misleading as they are not mutually exclusive. Which answer do you pick if you are 25 — a) or b)? Which answer do you pick if you are 40 — d) or e)? Logically, one and the same possible variation cannot fall into two or more categories; it can only fall in *one* of the variable's categories.

Thus, one proper way to operationalize *age* is something like this:

Q3. What is your age?

- a) 18-25
- b) 26-30
- c) 31-35
- d) 36-40
- e) 41-45
- f) 46-50
- g) Above 50

See if you can spot and fix violations of the two logical operationalization requirements in the exercise below.

*Do It! 1.6. What is Wrong with These Variables' Operationalizations?*

Q4. What year in college are you?

- a) First-year
- b) Second-year

Q5. How many siblings do you have?

- a) 0
- b) 1
- c) 1-2
- d) 3-4
- e) 4 or more

Q6. How do you commute to your institution's campus?

- a) Car
- b) Public transit
- c) Bus
- d) Bike

Now that we've covered the theoretical preliminaries, go see what working with actual data is like, in Chapter 2.

---

## Chapter 2 What Data Looks Like and Summarizing Data

This chapter moves us to more practical matters, namely working with actual data. Once you get familiar with what real data sets look like and how they are organized, you will learn how to summarize the information contained within variables. We can do that through tables and through graphs. Both reflect the *distribution of a variable* (a concept which we'll discuss extensively from Chapter 3 on), which is the way the observations/data points are distributed across a variable's categories. (For example, counting how many of your friends don't have siblings, how many have one sibling, how many have two siblings, etc, and writing the information down will give you the (frequency) distribution of the variable *number of siblings you friends have*.)

We start with frequency tables, and explore the summary information contained within. We end the chapter with the way we can visually display variables (i.e., their distribution) and the discussion of what type of graph (a pie chart, a bar graph, or a histogram) is most appropriate for variables at different levels of measurement.





---

## 2.1 Data Sets and What Data "Looks" Like

By now you have learned that *variables* are tools that allow us to measure concepts and to collect information about them. As such they are comprised of information — information that varies across the *units of analysis* (the ‘things’ on which we collect information, be it people, organizations, countries, etc.). So far, we have discussed individual variables – but creating and collecting information on a single variable is uncommon. Generally, we collect information on many variables at the same time (which, in turn, allows us to analyze variables together and hypothesize about possible associations between variables).

Variables “live” in data sets (or datasets, as I prefer; both usages are common). **A *dataset* is a collection of variables that lists the information (or observations) gathered on them from the units of analysis.** As usual, I focus on analysis of people for simplicity’s sake (but do keep in mind the units of analysis can be something else.)

The best way to visualize a dataset is as a sort of a table (a.k.a a *matrix*) which summarizes the responses from every individual (in the rows of the table) on the variables in the dataset (in the columns of the table). As such, the size of a dataset depends on two things: the number of variables and the number of individuals supplying information

(a.k.a. respondents). Typically, datasets vary in size from just a handful of variables and few respondents to hundreds of variables and thousands of respondents. (Huge datasets — comprising information on millions of people — exist too; these are known as *big data*. Big data is not analyzed in the conventional ways regular datasets are, so from now on we'll leave big data aside as it's not the subject of this book.)

To start small, imagine you have just four friends at your university and you decide to list some items of information about them (say, maybe you want to compare your standing at the university with theirs, and to see differences and commonalities between you and them). You could do that in a sentence form, for example, thus: Arjun, who is twenty years old, speaks Punjabi at home and is a first year student in the Business School, has a job and his GPA is 3.6. Benjamin, on the other hand, who is 25, speaks German at home and is a third year Science student, also has a job but his GPA is lower than Arjun's at 3.2. Cecilia, who speaks Spanish at home and is a fourth year Health Sciences student doesn't have a paying job and her GPA is the highest of your friends, 4.0. Finally, Xingxing is also a first year student and is employed like Arjun but she is an Arts major, speaks Mandarin at home, and her GPA is 3.3.

Indeed, you might do that but the points of comparison might get lost as they are not easy to see: one has to read very carefully to keep track of who does what and has a GPA of how much. Instead, you could present the same information as it is in the table in Example 2.1 below.

*Example 2.1 (A) A Hypothetical Dataset of Four Friends's Characteristics*

	Age	Year at university	Employment	GPA	Major (by Faculty)	Language spoken at home
<b>Arjun</b>	20	1	yes	3.6	Business	Punjabi
<b>Benjamin</b>	25	3	yes	3.2	Science	German
<b>Cecilia</b>	22	4	no	4.0	Health	Spanish
<b>Xingxing</b>	19	1	yes	3.3	Arts	Mandarin

If you do that, what you have created is a dataset. Now imagine that instead of this contrived combination of four friends and their varying characteristics, I generalize the example like so:

*Example 2.1 (B) A Hypothetical Dataset of Four Individuals and Six Variables*

	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
Respondent #1	Response <sub>1.1</sub>	Response <sub>2.1</sub>	Response <sub>3.1</sub>	Response <sub>4.1</sub>	Response <sub>5.1</sub>
Respondent #2	Response <sub>1.2</sub>	Response <sub>2.2</sub>	Response <sub>3.2</sub>	Response <sub>4.2</sub>	Response <sub>5.2</sub>
Respondent #3	Response <sub>1.3</sub>	Response <sub>2.3</sub>	Response <sub>3.3</sub>	Response <sub>4.3</sub>	Response <sub>5.3</sub>
Respondent #4	Response <sub>1.4</sub>	Response <sub>2.4</sub>	Response <sub>3.4</sub>	Response <sub>4.4</sub>	Response <sub>5.4</sub>

In Example 2.1 (B), the respondents are the four people on whose varying characteristics we have information, and these are represented by the six variables. This, however, seems a rather cumbersome. Instead of “Variable 3”, and “Respondent 5”, and “Response<sub>4.3</sub>“, etc., a simpler way to represent all of these in a generalized way is through mathematical notation.<sup>1</sup>

So, prepare yourselves! Here comes notation:

1. A note on mathematical notation, about which, I know, many students feel quite anxious: think of notation as a type of shorthand, or a sort of simplified foreign language. It's used to simplify what you can write out in words and sentences but would be too long and not as clear. The key to notation, just like with any foreign language, is to know what the symbols mean. Keep their meaning in mind, and you can read notation as fast and as easily as your own language.

*Example 2.1 (C) A Hypothetical Dataset of Four Individuals and Six Variables 2.0*

	<b>X<sub>1</sub></b>	<b>X<sub>2</sub></b>	<b>X<sub>3</sub></b>	<b>X<sub>4</sub></b>	<b>X<sub>5</sub></b>	<b>X<sub>6</sub></b>
<b>I<sub>1</sub></b>	x <sub>11</sub>	x <sub>21</sub>	x <sub>31</sub>	x <sub>41</sub>	x <sub>51</sub>	x <sub>61</sub>
<b>I<sub>2</sub></b>	x <sub>12</sub>	x <sub>22</sub>	x <sub>32</sub>	x <sub>42</sub>	x <sub>52</sub>	x <sub>62</sub>
<b>I<sub>3</sub></b>	x <sub>13</sub>	x <sub>23</sub>	x <sub>33</sub>	x <sub>43</sub>	x <sub>53</sub>	x <sub>63</sub>
<b>I<sub>4</sub></b>	x <sub>14</sub>	x <sub>24</sub>	x <sub>34</sub>	x <sub>44</sub>	x <sub>54</sub>	x <sub>64</sub>

In Example 2.1 (C),  $I_1$ ,  $I_2$ ,  $I_3$ , and  $I_4$  are the four individuals;  $X_1$ ,  $X_2$ ,  $X_3$ ,  $X_4$ ,  $X_5$ , and  $X_6$  are the six variables; and  $x_{11}$ ,  $x_{12}$ , etc. stand for any specific characteristic/response a respondent has on a variable. More specifically,  $x_{53}$ , for example, is the characteristic that Respondent #3 has on Variable 5. Scrolling up to Example 2.1 (A) will allow you to see that  $x_{53}$  is *Health*, which is Cecilia's Major by Faculty.

*Do It! 2.1 Reading Points of Information*

In a similar vein, look up  $x_{22}$ ,  $x_{34}$ , and  $x_{61}$ . It's a simple and easy task but it will help you connect notation to what it

stands for, and to understand the logic underlying the way information is presented in datasets.

From here, it’s not difficult to extrapolate the specific dataset we had above to a general one. Thus, Example 2.1 (D) below presents a template of a typical dataset.

*Example 2.1 (D) A Hypothetical Dataset of  $N$  Individuals and  $K$  Variables*

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	...	$X_K$
$I_1$	$x_{11}$	$x_{21}$	$x_{31}$	$x_{41}$	$x_{51}$	$x_{61}$	$x_{71}$	...	$x_{k1}$
$I_2$	$x_{12}$	$x_{22}$	$x_{32}$	$x_{42}$	$x_{52}$	$x_{62}$	$x_{72}$	...	$x_{k2}$
$I_3$	$x_{13}$	$x_{23}$	$x_{33}$	$x_{43}$	$x_{53}$	$x_{63}$	$x_{73}$	...	$x_{k3}$
$I_4$	$x_{14}$	$x_{24}$	$x_{34}$	$x_{44}$	$x_{54}$	$x_{64}$	$x_{74}$	...	$x_{k4}$
$I_5$	$x_{15}$	$x_{25}$	$x_{35}$	$x_{45}$	$x_{55}$	$x_{65}$	$x_{75}$	...	$x_{k5}$
$I_6$	$x_{16}$	$x_{26}$	$x_{36}$	$x_{46}$	$x_{56}$	$x_{66}$	$x_{76}$	...	$x_{k6}$
$I_7$	$x_{17}$	$x_{27}$	$x_{37}$	$x_{47}$	$x_{57}$	$x_{67}$	$x_{77}$	...	$x_{k7}$
...	...	...	...	...	...	...	...	...	...
$I_N$	$x_{1n}$	$x_{2n}$	$x_{3n}$	$x_{4n}$	$x_{5n}$	$x_{6n}$	$x_{7n}$	...	$x_{kn}$

$N$  = number of elements in the dataset

$K$  = number of variables in the dataset

In the table above, you may think of  $N$  as the last row on the table, i.e., the last individual for whom we have information and you may think of  $K$  as the last column on the table, i.e., the last variable we have in the dataset. Both numbers can theoretically be “any positive number”, though in practice the former is usually a number up to several thousands and the latter a number up to few hundreds. The ellipses in the next-to-last row and the next-to-last column indicate that the table is truncated: there are omitted rows between the seventh and the last individuals (i.e., between  $I_7$  and  $I_N$ ), and omitted columns between the seventh and the last variables (i.e., between  $X_7$  and  $X_K$ ). (They obviously have to be omitted so that the table can fit on the page.)

Armed with this knowledge, let’s take a look at an excerpt from a real dataset. The following Example 2.1 (E) provides a snapshot of the first ten respondents and first nine variables in the *Aboriginal Peoples Survey 2012* dataset (or *APS 2012* for short)<sup>2</sup> using a software called *IBM® Statistical Package for the Social Sciences*, commonly referred to as SPSS.

*Example 2.1 (E) A Snapshot of Survey Data (APS 2012)*

2. APS 2012 is a Statistics Canada dataset which I will formally introduce in **Ch. XX**.

Snapshot of APS 2012's Data View in SPSS:

	AGE_YRSG	SEX	DIDENTG	ID_03G	ID_05G	MS_01G	DSIZHHGG	DHHTYPEG	MOB_02A
1	5	2	1	1	1	3	5	3	1
2	4	2	1	1	1	4	4	1	1
3	5	2	3	2	2	2	5	1	2
4	4	2	2	2	2	4	5	1	6
5	7	1	1	1	1	2	4	1	2
6	1	2	2	2	2	6	5	1	1
7	7	1	1	1	1	1	4	1	2
8	1	2	2	2	2	6	4	1	1
9	7	1	1	1	2	2	3	1	2
10	3	1	1	1	1	6	5	4	6

Snapshot of APS 2012's Variable View in SPSS:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns
1	AGE_YRSG	Numeric	2	0	Age group of respondent - Survey reference date	{1, Between ages 6 - 8}...	96 - 99	10
2	SEX	Numeric	1	0	Sex of respondent	{1, MALE}...	6 - 9	8
3	DIDENTG	Numeric	1	0	Aboriginal identity pop. indicator (group)	{1, Single ID - First Nations}...	6 - 9	8
4	ID_03G	Numeric	1	0	Identity - Status Indian (Registered/Treaty)	{1, Regist./Treaty Indian}...	6 - 9	8
5	ID_05G	Numeric	1	0	Identity - Member of First Nation/Indian band	{1, FN member/Indian band}...	6 - 9	8
6	MS_01G	Numeric	1	0	Marital status (respondent)	{1, Married}...	6 - 9	8
7	DSIZHHGG	Numeric	1	0	DV - Household - # of persons - Grouped	{1, One person}...	6 - 9	9
8	DHHTYPEG	Numeric	2	0	DV - Household by family/non-family type	{1, 9 fam hhold, Cple w child(s)}...	96 - 99	10
9	MOB_02A	Numeric	1	0	Reason move to current comm. - Family/Spouse	{1, Yes}...	6 - 9	9

Do It! 2.2 Understanding How Datasets Are Organized

Make sure you can connect the data snapshots from the example above with your understanding of how datasets are



organized. What do the numbers in the first (blue) columns in both images represent? (Hint: this is not a variable!) What is listed in the first (blue) row in the top image? In the top image what does 1 stand for in the first white row in column *ID\_03G*? How about the 1 in the fifth row in the *SEX* column?

Answer: *Registered/Status Indian* and *male*, respectively.

One thing you might find surprising is the obvious fact that all cell entries (i.e., the observations we have) are listed in a number format. Does that mean that all variables in this particular dataset are interval or ratio? What about any nominal or ordinal variables – do they not exist in this dataset? The answer is *no* on both accounts: the variable *SEX* (i.e., “*Sex of respondent*” as stated in *Variable View*) is nominal and the variable *AGE\_YRSG* (i.e., “*Age group of respondent...*”) is ordinal because of the hierarchical arrangement of the responses. **However, the dataset cells contain only numbers because statistical software can only analyze numerical data.**

To that effect, nominal and ordinal variables appear “in code” in datasets; i.e., **the categories of nominal and ordinal variables are assigned numerical values as labels to represent them** in the actual dataset you might be working with. Thus, the numbers in nominal and ordinal variables’ columns are not *actual numbers*, they are artificially (and in the case of nominal variables, somewhat arbitrarily) assigned to represent the words contained in the categories in order to make computer-based statistical

analysis possible. (On the other hand, interval/ratio variables' categories contain *actual numbers*. Of course, the trick then is to learn to differentiate the actual numbers from the code/ number values used as labels in the cells of a dataset.)

Therefore, you should always keep track of the code (see the Watch Out! panel below for tips on *Variable View* in SPSS which allows you to do that), and remember to refer to the categories by their proper (word-based) names — not by the artificial numerical values (i.e., code) representing them!

**Watch Out! #2**...for Making Hasty Decisions about Variables Based Only on Data View or Only on Variable View

It's tempting, but you cannot deduce *all* categories of a variable with any certainty just by looking at the snapshot in Example 2.1 (E). You cannot do that even if, instead of a snapshot, you had the real, interactive *Data View* window in SPSS in front of you. Not only you might not be able to scroll through all the data (depending on its size) but, more importantly, not all characteristics might exist among the individuals. (For example, imagine the variable *hair colour*, and say, not one respondent having red hair: then a response "red" would not be visible in *Data View*, even if such a category existed in the variable.) For the same reasons you should also not decide a variable's level of measurement based on *Data View*. (Remember, all data in the cells appears in numerical format, regardless if it's an actual number or just a value label/code!)

To explore any dataset you might end up working with and all the variables contained therein, you should always look to explore not only the *Data View* but the *Variable View* of the dataset as well (in SPSS you can toggle between Data View and Variable View easily with a click of the mouse). The *Variable View* lists all variables along with some information about them — including something which *looks like* their level of measurement, called *Measure* (it is not included in the bottom snapshot above). **The *Measure* information can be quite misleading for students so: Never trust this software-generated conclusion!**

Instead, you should always explore *both Variable View* and *Data View*. You should note the variables' respective categories (in *Variable View*, where you can click on any cell in the *Values* column for a full category listing) and the type of the observations you have in the cells in the table (in *Data View*). Then —and *only* then — reach the appropriate conclusion about the levels of measurement of the variables you have at hand.

What should guide your decision about a variable's level of measurement is what you see in the *Values* column in *Data View*. To repeat, clicking on the respective column will open up a window displaying the (nominal or ordinal) variable's categories/values along with the number label representing them in the dataset.

Again, note that reporting on the variable should be done by using its categories/values, never by the number label you see in *Variable View* standing in for them! This point will become more relevant and less abstract once we start learning what to do with variables, in Chapter 3.



---

## 2.2 Summarizing Data

Imagine a dataset containing a hundred respondents and just five variables. Such a dataset would have 500 data points and, while that may seem like a lot, a dataset of this size is considered rather small. Typically, datasets used in sociology (and other social sciences) tend to be larger. What this tells you is that there is an enormous amount of information housed within even an average dataset.

Just like a library containing thousands and thousands of books but no catalog, unless we have the means to make sense of the information – order it, systematize it, categorize it – that information is all but useless. In the previous section, I discussed exploring a dataset in SPSS's *Data View*. While that's a useful (and necessary) task to do before working with any dataset, it doesn't provide anything more than a sort of global view of the variables in it.

In order to understand any variables better and to be able to fully use the information they contain, we need tools to allow us to *zoom in* each individual variable, as it were, and to organize that information in a meaningful way.

Two of the most widely used such tools for exploring variables and presenting their information in a summarized, easy to understand way are, as you well know, *tables* and *graphs*. In the next section I start by

introducing **frequency tables**; then we will end this chapter with introducing **graphical displays**.

---

## 2.3 Frequency Tables

As usual, let's start ground-up with an example and work our way up to the concept under study. Consider the following raw (unorganized) data.

### *Example 2.2 (A) Hypothetical Raw Data on Educational Attainment*

Imagine that a group of 21 people were asked about the highest educational degree they have attained. These are their responses:

<b>Secondary/ High School</b>	<b>Bachelor's</b>	<b>Secondary/ High School</b>	<b>No Degree</b>	<b>Bachelor's</b>	<b>Didn't answer</b>
<b>Master's</b>	<b>Associate's</b>	<b>Master's</b>	<b>Secondary/ High School</b>	<b>Bachelor's</b>	
<b>Secondary/ High School</b>	<b>Secondary/ High School</b>	<b>Didn't answer</b>	<b>Didn't answer</b>	<b>Bachelor's</b>	
<b>Secondary/ High School</b>	<b>PhD</b>	<b>Bachelor's</b>	<b>Associate's</b>	<b>Associate's</b>	

What can we glean from this presentation of the information? Can we easily see which is the most frequently obtained educational degree in the group? How many people do we have of each degree? What fraction/proportion of the total are each?

Of course, we could always count — but what if I had asked you to imagine a group of 36 people? Of 72? Or 200? Or 2,000? Or more? Are you still going to painstakingly count the different responses?

You may be surprised, but the answer is “yes, if we had to”. In the past, researchers used to do a that, a lot. Nowadays of course we have computers to do it for us. SPSS can easily summarize this data but to understand the process better, we’ll start from scratch.

The most obvious way we can organize the raw data above into something less chaotic is the following:



*Example 2.2 (B) Hypothetical Data on Educational Attainment, Organized*

*Table 2.1 Educational Attainment by Frequency*

<b>Degree</b>	<b>Count (a.k.a. <i>frequency</i>)</b>
No degree	1
Secondary/High School	6
Associate's	3
Bachelor's	5
Master's	2
PhD	1
Didn't answer	3
<b>TOTAL</b>	<b>21</b>

In the most basic sense, this is a *frequency table*. It lists the different categories of a variable along with their observed count, a.k.a. *frequency*. That is, **we essentially count how many times any given category appears, i.e., we count how frequent a response is among the**

**respondents, and then indicate the number for each category/response. Frequency is usually denoted by  $f$  in statistical notation.**

Real frequency tables, however, usually contain more information than a simple count. The following few subsections provide the details, while we work our way through creating a full frequency table.

---

## 2.3.1 Relative Frequency: Adding Percentages

Simply counting the frequency of the different variable's categories (or the number of specific responses) is rarely enough. Often, we also want to know what *proportion* — or what *percentage* — of the total each category represents. This is especially important when comparing across two or more different groups. Thus we will stop on our way to frequency tables to undertake a brief side quest into *relative frequency* territory.

**Watch Out!! #3...** *for Cross-Group Comparisons Using Counted Numbers*

Imagine that researchers are conducting a study on eating habits and they have interviewed 170 people; 102 identified as men and 68 identified as women. Say that the researchers found that 17 of the men and 13 of women reported a vegan diet. Can the researchers conclude that men tend to favour vegan diets more than women do?

If you go by the actual, counted numbers reported, you may decide that yes, the researchers' conclusion is correct as

17 is more than 13, i.e., four more men than women have reported eating vegan. This, however, would be wrong. We cannot compare the two groups (men and women) directly since the groups have different sizes. That is, comparison of the numbers as counted in the two groups has little meaning since it does not take into account group size. Yes, more men report eating vegan but men in the study outnumber women by 24 to start with. Thus, **maybe we find more vegan men than women simply because there are more men than women in the study.** What we should be asking ourselves instead is whether a larger *proportion* of men eat vegan, compared to women — and the correct answer would require a comparison of the numbers **relative to group size.**

A quick calculation reveals that 17 out of 102 is actually *less* than 13 out of 68:

$$\frac{17}{102} = 0.167$$

$$\frac{13}{68} = 0.191$$

That is, **the proportion of vegan men (0.167) is smaller than the proportion of vegan women (0.191)**, so no, we cannot say that men tend to be vegan more than women do. Rather, it's the other way around: **more women than men tend to eat vegan, because vegan women are a higher proportion (i.e., the number for women is higher relative to their group size).**

To conclude, **never use numbers as counted to compare between groups** (unless they are of equal size). To make comparison possible — and meaningful — you should **always use proportions or percentages** (i.e., the numbers relative to the size of each group).

A bit more notation then: if we denote *frequency* by  $f$ , and you recall that  $N$  stands for *number* (of elements in a dataset; of people in a group, etc.), it would be easy to see that *proportion* — denoted by  $p$  — should be

$$\frac{f}{N} = p$$

.

While actual numbers represent frequency, proportions are one way of expressing **relative frequency**. You probably are more familiar with another way of expressing relative frequency — **percentages**.

In the example I used in the *Watch Out!!* #3 above, we concluded that more women than men were vegan based on the fact that the proportion of vegan women (0.191) was higher than the proportion of vegan men (0.167). In everyday life, people usually tend to use percentages to express that. **To convert proportions to percentages you only need to multiply by a 100<sup>1</sup>:**

1. After all, percent or per cent comes from the Latin "per centum", meaning "by a hundred"; i.e., whatever proportion you are expressing, standardized by a hundred.

$$\frac{f}{N}(100) = \textit{percent}$$

Thus, we get the following percentages when comparing vegan men and women from the *Watch Out!! #3* above:

$$0.167(100) = 16.7\%$$

and

$$0.191(100) = 19.1\%$$

.

That is, we could rephrase our finding and say that since only 16.7 percent of men reported being vegan while 19.1 percent of women did, clearly women are more likely to be vegan based on this particular group of respondents.

Note that **while proportions range from 0 to 1 and typically get rounded up to three digits after the decimal point** (e.g., 0.167 and 0.191), **percentages range from 0 to 100 and usually get rounded up to one or two digits after the decimal point** (e.g., 16.7% and 19.1%). Also note that **differences in percentages are expressed in *percentage points*, not in percent**: in the current example, the difference between men and women who eat vegan is  $(19.1\% - 16.7\%) = 2.4$  percentage *points* in favour of women being vegan, *not* 2.4 percent.

A final way to express relative frequencies are **ratios**, where a ratio is simply one frequency/count relative to another:

$$\frac{f_1}{f_2} = ratio$$

Using the numbers from the *Watch Out!!* #3 above, we can say that in the group of 170 respondents (102 men and 68 women), we have a men-to-women ratio of 1.5 — or, men in the study outnumber women by 1.5 to 1 since

$$\frac{f_m}{f_w} = \frac{102}{68} = 1.5$$

It's easy to see that if we want the women-to-men ratio, we only need to switch the numerator and denominator of the ratio:

$$\frac{f_w}{f_m} = \frac{68}{102} = 0.7$$

This still tells us that men outnumber women as for every 1 man there is only a “0.7 woman”. Since this type of fractions, depending on the context, can lead to an awkward phrasing (like in this case), you may choose to report a ratio in the way most apt for easier interpretation.

Relative frequencies are all nice and good, but let's go back to our main quest, the frequency table. Since we established that reported actual numbers are meaningless for comparison purposes and that we need relative frequencies to do that, it would only make sense to add a relative frequency column to our *educational attainment* Table 2.1 from Example 2.2 (B).

The percentages in Table 2.2 below have all been calculated using the steps described above: 1) obtain proportion, and 2) multiply by a 100. For example, only one of our original 21 respondents had no degree. Then the percentage of the 21 respondents with no degree is:

$$\frac{f}{N}(100) = \frac{1}{21}(100) = 0.047(100) = 4.7\%$$

The rest of the categories' percentages are calculated in the same vain.

*Example 2.2 (C) Hypothetical Data on Educational Attainment, Organized and with Relative Frequencies Added*

*Table 2.2 Educational Attainment by Frequency and Percent*



---

<b>Degree</b>	<b>Frequency</b>	<b>Percent</b>
No degree	1	4.7
Secondary/ High School	6	28.6
Associate's	3	14.3
Bachelor's	5	23.8
Master's	2	9.5
PhD	1	4.7
Didn't answer	3	14.3
<b>TOTAL</b>	<b>21</b>	<b>100.0</b>

---

This way we can easily see how the respondents are distributed across the different educational attainment categories and each category's share as a fraction of the total. If we had another group of respondents, we could easily compare between our initial group of 21 and the second hypothetical group by using the percentages above. Or can we?



---

## 2.3.2 Missing Data: Adding Valid Percentages

If you've paid attention so far, you must have noticed that three of our 21 respondents provided a "Didn't answer" response when asked about their educational attainment. Sometimes respondents may refuse to answer a question, or the question may not have been applicable to them and wasn't asked, or a response might not get recorded due to an error, etc. In short, sometimes we have a case of what is known as *missing data*.

What do we know about the educational attainment of the three individuals who, for whatever reason, didn't answer this question? Nothing.

Can we in some way infer their educational attainment? Not with the data provided in the example.

So then what do we do? How do we analyze our *educational attainment* variable?

The most frequent — and strongly recommended (especially for people just starting on their journey to research) — course of action is to simply *drop* the missing cases<sup>1</sup>. Missing cases have no part in any analysis and

1. Depending on the particular data and particular situation, and assuming strong justification, researchers experienced in data analysis may have different options, such as estimation, imputation of means, etc. These, however, are beyond the scope of this text. The safest action for students/beginners to

using them as they are would inevitably compromise conclusions — after all, we have no information on what we want to know about them, and we cannot make that information up.

Generally, how statistical software deal with missing data by default settings may vary. SPSS's default is to skip missing cases so that analysis is always based on valid cases only.

As well, SPSS provides a separate column in *Data View* indicating which values in the data stand for a missing data point. As discussed in Section 2.1 (<https://pressbooks.bccampus.ca/simplestats/chapter/2-1-data/>), you can find the *coding* of the values in the *Values* column in *Data View*. Clicking the specific cell in that column opens up a window with the values' code. There you may find several types of missing data, typically values such as “Valid skip”/“Not applicable” (the respondent had not been asked the question on which the variable is based due to a previous answer)<sup>2</sup>, “Don't know” (the respondent did not know the answer to the question), “Refusal” (the respondent refused to answer the question), “Not stated” (when the question should have been answered/ an answer should have been recorded but, for whatever reason, it hasn't been), etc.

Apart from “Not applicable”, the codes listed here are

take remains dropping any missing cases from the analysis. See <https://www.iriseekhout.com/missing-data/missing-data-methods/imputation-methods/> for a discussion.

2. For example, if a respondent has indicated previously that they didn't smoke, a subsequent question about how often they smoked would make no sense; the respondent then would be “validly skipped” from answering this subsequent question.

standard Statistics Canada codes used in all their datasets and can be found in any Statistics Canada dataset documentation<sup>3</sup>.

So given that we had three cases of missing data within our group of 21 respondents, are the percentages reported in the previous sub-section's Table 2.2 in Example 2.2 (C) *valid* to use?

**Watch Out!! #4... for Findings Based on Missing Data**

This will be a short warning but it deserves its own scary-red *Watch Out!!* reiteration: do not trust analysis and findings that include missing cases as they would be distorted and unreliable. Missing data is exactly that – *missing*. It simply does not exist. As a beginner researcher, always make sure you have dropped (i.e., excluded) any missing cases before analyzing your data and reporting any results.

Considering that Table 2.2 did include missing data in the calculation of percentages, let us correct that by modifying it and including another column, ***valid percentages***.

3. Currently, Statistics Canada uses 6, 96, 996, etc. for "Valid skip"; 7, 97, 997, etc. for "Don't know"; 8, 98, 998, etc. for "Refused"; and 9, 99, 999, etc. for "Not stated".

*Example 2.2 (D) Hypothetical Data on Educational Attainment, Organized and with Relative Frequencies and Valid Percentages Added*

*Table 2.3 Educational Attainment by Frequency, Percent and Valid Percent*

	Degree	Frequency	Percent	Valid Percent
Valid	No degree	1	4.7	5.6
	Secondary/ High School	6	28.6	33.3
	Associate's	3	14.3	16.7
	Bachelor's	5	23.8	27.8
	Master's	2	9.5	11.1
	PhD	1	4.7	5.6
	<b>Total Valid</b>	<b>18</b>	<b>85.6</b>	<b>100.0</b>
Missing	Didn't answer	3	14.3	
	Total Missing	3	14.3	
	<b>TOTAL</b>	<b>21</b>	<b>100.0</b>	

As you see in the modified Table 2.3 above, I have separated the missing cases from the valid cases (the cases for which we have educational attainment data). **Since we have only 18 valid cases, we should use only those 18 cases for any calculations and analysis — and not the**

**total of 21 cases** (which includes the missing). Thus, instead of having just

$$\frac{f}{N}(100) = \frac{1}{21}(100) = 0.047(100) = 4.7\%$$

along with the rest of the categories' percentages calculated in this way, we should calculate the categories' *valid* percentages, discarding the three missing cases, like this:

$$\frac{f}{N}(100) = \frac{1}{18}(100) = 0.056(100) = 5.6\%$$

(As usual, I only show you the calculation for the first category as the rest follow in the same way.)

Despite the fact that we do have the percentages based on missing data in the table, note that these – **the valid percentages** — **are the only percentages you should use in your analysis and report in your findings.**

*Alright, you might say now, we added percentages and valid percentages to the simple frequencies, this surely means we have a complete frequency table by now.*

Sorry, no, not yet. One thing remains.



---

## 2.3.3 Summing Up: Adding Cumulative Percentages

The thing that remains to add to our frequency table is there only for convenience's sake. It can be useful to know, for example, what percentage of the 21 people in our original group do not have graduate degrees, or what percentage of people have not gone to university, etc. Of course, in our specific *educational attainment* example it would be easy to to the quick-and-dirty calculation of adding 11.1 percent (those with Master's degrees) to 5.6 percent (those with PhD), thus finding that 16.7 percent of our respondents have graduate degrees; or adding 5.6 percent (those without a degree) to 33.3 percent (those with Secondary/High School) and finding that 38.9 percent of our respondents have not gone to university. Doing such calculations all the time, depending on the question, might get tedious, however, at best, and, at worst, it's also incorrect (hence the "quick-and-dirty" appellation).

Let's then improve on our frequency table-in-progress a final time, shall we? The version below is the final version, ta-da!

*Example 2.2 (E) Frequency Table for Educational Attainment*

Table 2.4 Educational Attainment by Frequency, Percent, Valid Percent and Cumulative Percent					
	Degree	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No degree	1	4.7	5.6	5.6
	Secondary/ High School	6	28.6	33.3	38.9
	Associate's	3	14.3	16.7	55.6
	Bachelor's	5	23.8	27.8	83.3
	Master's	2	9.5	11.1	94.4
	PhD	1	4.7	5.6	100.0
	<b>Total Valid</b>	<b>18</b>	<b>85.6</b>	<b>100.0</b>	
Missing	Didn't answer	3	14.3		
	Total Missing	3	14.3		
	TOTAL	21	100.0		

The final column I have added in our Table 2.4 is called **Cumulative Percent**. What it does is keep a sort of a

**“running total”**, adding the second category’s frequency to the first and reporting the first two categories as a fraction of the total; adding the third category’s frequency to the total of the first two and reporting the first three categories as a fraction of the total, etc. — in effect **adding each subsequent category to the total of all preceding ones, one by one, until all categories are added together.**

Note, however, that you should not add the percentages in the *Valid Percent* column to obtain cumulative percentages. Despite the quick-and-dirty trick I did before, I actually calculated the cumulative percentages based on the added categories’ frequencies, and so should you, if you have to create a frequency table from scratch.

Like this: there is one person without a degree and 6 people with secondary/high school degrees, or 7 people combined. Therefore, the cumulative percent of these two categories is obtained thus:

$$\frac{f_1 + f_2}{N}(100) = \frac{1 + 6}{18}(100) = \frac{7}{18} = 0.389(100) = 38.9\%$$

and *not* by adding 5.6 percent (the person with no degree) to 33.3 percent (the ones with secondary/high school degrees) — even if in this case, both produce the same result, 38.9 percent.

**The reason why we need to add the original frequencies and not the valid percentages themselves is rounding.** The percentages reported in the frequency table are rounded to 1 digit after the decimal point; adding

rounded numbers inevitably adds imprecision to the result, which, depending on the situation, might end up being crucial. In our case, it makes no difference but do note that the percentages reported in the *Percent* column actually only add up to 99.9 percent, not 100 percent; similarly, the percentages reported in the *Valid Percent* column actually add up to 100.1 percent rather than 100 percent. These differences, as negligible as they seem when working with a variable with few categories like the one here, can add up and become more significant in variables with numerous categories (like interval/ratio variables, for example).

You can see examples of real-data frequency tables in the next-subsection.

---

## 2.3.4 What Frequency Tables Really Look Like

Before we move on to the last section of this chapter, take a look at what frequency tables of real variables look like, using SPSS. All three variables in the tables below come from the *General Social Survey 2016* (or *GSS 2016*) (Statistics Canada 2018) which I'll formally introduce in **Chapter XX**.

*Table 2.5 Frequency Table for Sex of Respondent (GSS 2016)*

Sex of respondent					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	8782	44.8	44.8	44.8
	Female	10827	55.2	55.2	100.0
	Total	19609	100.0	100.0	

Table 2.5 shows a nominal variable, *sex of respondent*, with no missing data (thus both *Percent* and *Valid Percent* columns contain the same information).

Unlike it, Table 2.6 below shows an ordinal variable, *workplace size*, where almost half (47.4 percent) of the respondents didn't supply a valid response. In cases like this one it's imperative you only use the data as presented in the *Valid Percent* column, and not the *Percent* one.

*Table 2.6 Frequency Table for Workplace Size (GSS 2016)*

		Workplace size			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Small business	6409	32.7	62.2	62.2
	Midsized business	2165	11.0	21.0	83.2
	Large business	1732	8.8	16.8	100.0
	Total	10306	52.6	100.0	
Missing	Valid skip	9102	46.4		
	Don't know	153	.8		
	Refusal	20	.1		
	Not stated	28	.1		
	Total	9303	47.4		
Total		19609	100.0		

Table 2.7 below presents a ratio variable, *purchasing grocery store takeout dishes in the past month*, with relatively moderate number of data points missing (9.3 percent). Again, *Valid Percent* is the column at which you should be looking. As well, note that the first (blue) column lists the categories (or values) of the variable as supplied by the respondents, as it normally does. Since these consist of actual numbers, you might be tempted to see them as some sort of consecutive listing, and that would be wrong. If you look carefully, you'll see that numbers like 11, 19, 22, 23, etc. are not listed there. This is not because they are somehow "missing" but because no respondent provided such a response.

*Table 2.7 Frequency Table for Purchasing Grocery Store Takeout Dishes (GSS 2016)*

**Purchasing take out dishes from grocery stores - Past month**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	8504	43.4	47.8	47.8
	1	2533	12.9	14.2	62.0
	2	2373	12.1	13.3	75.4
	3	973	5.0	5.5	80.8
	4	1444	7.4	8.1	89.0
	5	687	3.5	3.9	92.8
	6	264	1.3	1.5	94.3
	7	66	.3	.4	94.7
	8	268	1.4	1.5	96.2
	9	13	.1	.1	96.2
	10	319	1.6	1.8	98.0
	12	114	.6	.6	98.7
	13	3	.0	.0	98.7
	14	5	.0	.0	98.7
	15	85	.4	.5	99.2
	16	11	.1	.1	99.3
	17	1	.0	.0	99.3
	18	1	.0	.0	99.3
	20	72	.4	.4	99.7
	21	1	.0	.0	99.7
	24	1	.0	.0	99.7
	25	13	.1	.1	99.8
	30	28	.1	.2	99.9
	31	5	.0	.0	99.9
	40	3	.0	.0	100.0
	43	1	.0	.0	100.0
	50	3	.0	.0	100.0
	60	1	.0	.0	100.0
	93	1	.0	.0	100.0
	Total	17793	90.7	100.0	
Missing	Valid skip	1702	8.7		
	Don't know	105	.5		
	Refusal	7	.0		
	Not stated	2	.0		
	Total	1816	9.3		
Total		19609	100.0		

Finally, note that although the *Cumulative Percent* column is less useful when we are dealing with nominal variables, it's quite handy to have when working with ordinal and especially with interval/ratio variables. Thus we can easily state that 83.2 percent of respondents work at a small or a midsize workplace and that almost 90 percent of respondents have purchased no more than 4 grocery takeout dishes in the past month.

#### *SPSS Tip 2.1: How to Request Frequency Tables*

From the *Main Menu*:

- Click *Analyze*, then *Descriptive statistics*, and then *Frequencies*;
- Select variable/s from the left-side of the window and use the arrow button to move the variable/s to the right side;
- Click *OK*.
- The *Output* window will display the selected variable/s frequency table/s.



---

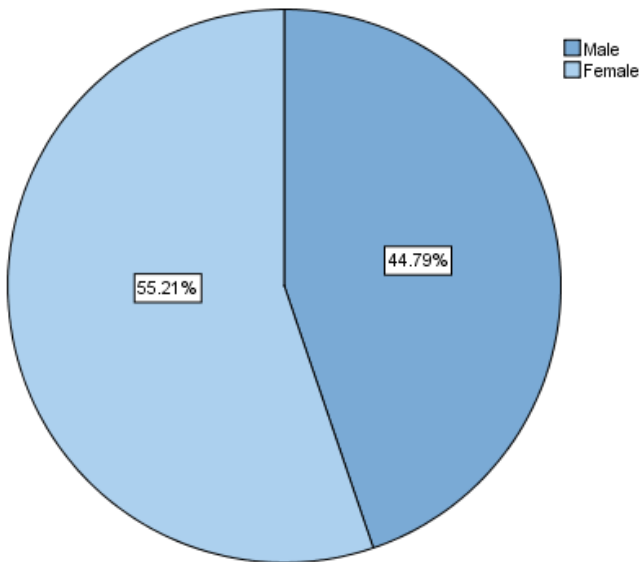
## 2.4 Graphs

A picture is worth a thousands words, they say, so in this section we will explore the most basic ways we can summarize data using graphical displays rather than tables. Unlike frequency tables which can be used to summarize variables at all levels of measurement with a a table of the same format, the types of graphs we use tend to differ depending on the variable's level of measurement. Almost all graphs in this book are produced using SPSS.

The three most basic graphs used to summarize variables are *pie charts*, *bar graphs* (or *bar charts*), and *histograms*.

**Pie charts.** You have undoubtedly encountered (and likely used) pie charts before. Fig. 2.1 below presents one such simple pie chart. The size of a slice of the “pie” corresponds to the category's size. The higher the category's frequency (and, of course, relative frequency), the larger the slice.

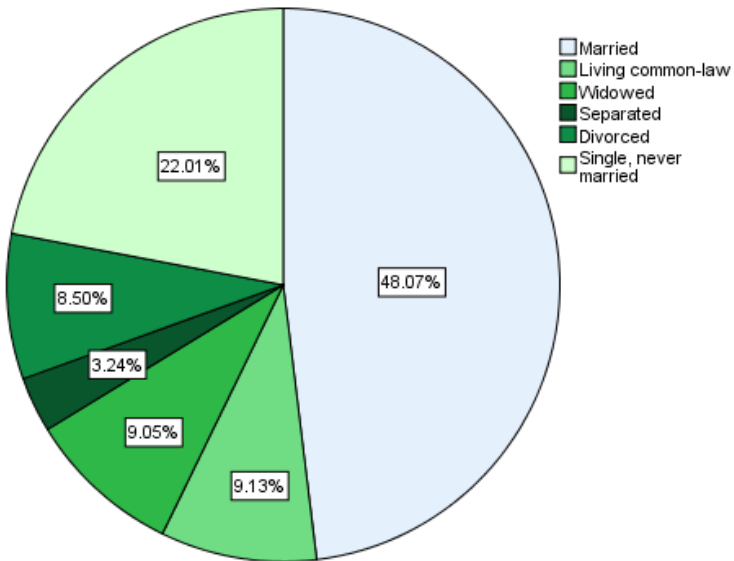
*Figure 2.1 Sex of the Respondent (GSS 2016)*



The pie chart in Fig. 2.1 corresponds to the frequency table of *sex of the respondent* in the previous section, namely Table 2.5.

Since the binary variable *sex* tends to look ‘boring’, in Fig. 2.2 below you can find a bonus pie chart for *marital status* which tends to be more colourful as it has more categories.

*Table 2.2 Marital Status of the Respondent (GSS 2016)*

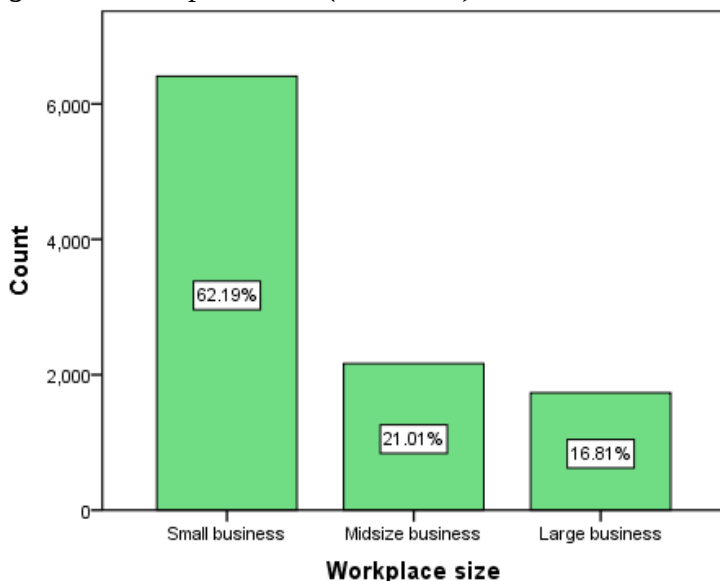


Pie charts can be used with both nominal and ordinal variables, though an argument can be made that the circular form of the pie chart may “hide” valuable insights about the order inherent in ordinal variables. As such, some prefer to use bar graphs for nominal variables *only*, and to use bar graphs for ordinal variables. Ultimately, it is a matter of preference, and both usages are correct.

You should not try to use a pie chart for an interval/ratio variable, however, as the “pie” in most cases will end up divided into far too many and far too small slices which will make “reading” the chart impossible.

**Bar graphs.** Fig. 2.3 below features a simple bar graph. The height of the bars corresponds to the size of the different categories. The higher the category’s frequency (and relative frequency), the taller the bar.

Figure 2.3 Workplace Size (GSS 2016)



This bar graph corresponds to the frequency table for *workplace size* from the previous section (Table 2.6). Note that the percentages reflected in the graph are the *valid* percentages from the frequency table.

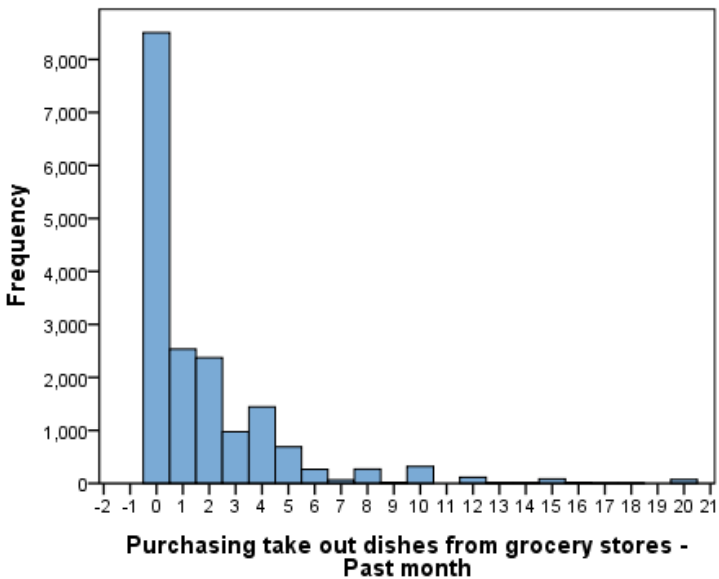
Again, using a bar graph with a nominal variable is allowed, and it's up to you whether you prefer to use a pie chart instead, since the categories of a nominal variable have no order and can be “moved around” without loss of information. However, a bar chart can present the order of a ordinal variable's categories in a more intuitive manner, so for some people bar graphs are the preferred graph of choice for ordinal variables: this way the order goes through the bars from left to right.

Like with pie charts, you shouldn't use bar graphs with

interval/ratio variables as the potential for ending up with far too many bars is quite high, making reading the graph difficult.

**Histograms.** Histograms are the graphical representations used with interval/ratio variables. Fig. 2.4 presents one such histogram. Once again, the height of each bar represents the frequency of a variable's category. In this case, the histogram corresponds to Table 2.7 from the previous section which was the frequency table of the number of takeout dishes respondents purchased in the last month.

*Figure 2.4 Purchasing Takeout Dishes from Grocery Stores in the Past Month (GSS 2016)*



At first glance, a histogram might look similar to a bar graph – albeit usually with more bars/categories. However,

the number of categories is not the only difference. Notice how the bars in the bar graph in Fig. 2.3 have space between them, while the bars in the histogram in Fig. 2.4 do not. This difference represents the difference between discrete and continuous variables: Discrete variables<sup>1</sup> have separate categories, hence the distance between the bars in the bar graph. Continuous variables (typically interval/ratio variables) have continuous categories, therefore the bars representing the categories touch each other to indicate their continuous nature (i.e., their potentially infinite number of values).

In the next two chapters you will learn how you can use these graphs in greater detail (especially the histogram). Here is how to produce them in SPSS.

#### *SPSS Tip 2.2 Basic Graphs*

##### **To get a pie chart:**

- From the *Main Menu*, click *Graphs* and then *Legacy Dialogs*;
- From the pull-down menu of *Legacy Dialogs*,

1. If you recall from Section 1.5 (<https://pressbooks.bccampus.ca/simplestats/chapter/1-5-discrete-and-continuous-variables/>), nominal and (typically) ordinal variables are considered discrete.

select *Pie*; a *Pie Charts* window will appear.

- Leave *Summaries for groups of cases* selected and click *Define*;
- Select your variable of interest from the left-hand side variable list and, using the correct arrow, move the variable into the *Define Slices* by empty space.
- You can change what the slices represent — the frequency (*N of cases*) or percentages (*% of cases*) in the top right section of the window called *Slices Represent*.
- When you are done, click *OK*. The pie chart will appear in the *Output* window.

### **To get a bar graph:**

- From the *Main Menu*, click *Graphs* and then *Legacy Dialogs*;
- From the pull-down menu of *Legacy Dialogs*, select *Bar*; a *Bar Charts* window will appear.
- Leave *Simple* and *Summaries for groups of cases* selected and click *Define*;
- Select your variable of interest from the left-hand side variable list and, using the correct arrow, move the variable into the *Category Axis* empty space.
- You can change what the slices represent — the frequency (*N of cases*) or percentages (*% of cases*) in the top right section of the window called *Bars Represent*.
- When you are done, click *OK*. The bar graph will appear in the *Output* window.

**To get a histogram:**

- From the *Main Menu*, click *Graphs* and then *Legacy Dialogs*;
- From the pull-down menu of *Legacy Dialogs*, select *Histogram*; a *Histogram* window will appear.
- Select your variable of interest from the left-hand side variable list and, using the correct arrow, move it into the *Variable* empty space.
- When you are done, click *OK*. The bar graph will appear in the *Output* window.



---

## Chapter 3 Measures of Central Tendency

Now that you have learned the preliminaries — what datasets and variables are, and how to summarize the information within a variable in tabular and graphical formats — it's time to turn to applied statistics proper. Statistics allows us to *analyze* information, i.e., to learn more than what we simply see at first glance. Thus we scrutinize the data collected in great detail to get the most out of it, in terms of both description (examining what we see) and inference (reaching evidence-based conclusions).

Aptly, we talk about *descriptive statistics* and *inferential statistics*. In the latter half of this book we will turn to inferential statistics which is devoted to inferential analysis on the basis of probability theory. We now start with descriptive statistics devoted to the descriptive analysis of variables, i.e., to learning all we possibly can about a variable and its distribution. If you recall from Chapter 2's introduction, **a variable's distribution is the way the observations/cases are distributed across the variable's categories**. The cases can be concentrated closer together or more spread out, and exploring such features of a variable's

distribution is the focus of this chapter and the next.

In addition to the visual summary of a variable which we get through graphs and which allow us to virtually *see* a variable's distribution, generally there are two further types of information we can get through descriptive analysis. They are called **central tendency** and **dispersion**.

Considering what a variable in a dataset looks like, recall that a variable has a list of observations/ cases (think, for example, of the responses collected through a survey question) where the list is size  $N$  ( $N$ , again, is the number of *elements*, in general, or *respondents* if we focus specifically on people, as we usually do). Thus, on the one hand, we talk about *typical cases*, or *where cases tend to cluster* — for example, what the most frequent response given is, if respondents tend to give similar answers, etc. — and what the “*centre*” of the variable's distribution is. Measures related to this type of information are called **measures of central tendency**. There are three of them and we explore all of them in the current chapter in turn, the mode, the median, and the mean.

On the other hand, we can also talk about how much a variable's distribution is “spread out”. That is, if a variable is called that because the responses *vary* across people, how *variable* a variable actually is – does it vary a lot or does it vary a little? Are all responses clustered around the “centre” or are they relatively dispersed? Measures related to this type of information are called **measures of dispersion**, and they are presented in the next chapter.

To summarize, **we describe variables by** providing and

exploring **1) the visual summary of their distribution (i.e., a graph), 2) their measures of central tendency, and 3) their measures of dispersion.**

There is a catch, however: **Not all measures of central tendency and dispersion are appropriate for all variables.** Just like not all graphs are appropriate for each type of variable, **whether a measure of central tendency or dispersion is applicable to a variable or not depends on the variable's level of measurement.**

I did already warn you that determining the proper level of measurement of a variable is key — without that, you can execute correctly neither descriptive, nor inferential analysis. Go back and reread Section 1.3 if necessary (<https://pressbooks.bccampus.ca/simplestats/chapter/1-3-levels-of-measurement/>) or what comes next will make little sense to you.

But enough with the boring theory — on to the the application of central tendency measures!



---

# 3.1 Mode

Central tendency is the information about the clustered-ness of a variable’s distribution; whether its observations/cases/responses tend to group together (or not) and where (i.e., in which categories/values) they tend to fall.

**There are three measures of central tendency: mode, median, and mean.** In this section, we explore the *mode*.

To find a variable’s mode, you only need a frequency table – or rather, even just the frequency column in the table (although the *Valid Percent* column will do you just as well). Here is a simple, small-*N*, real-world example.

Example 3.1 Religious Affiliation of Canadian Prime Ministers

Table 3.1 Religious Affiliation of Canadian Prime Ministers (Wikipedia 2017)

Religious affiliation	Frequency
Anglican	4
Baptist	3
Evangelical	1
Presbyterian	3
Roman Catholic	10
United Church of Canada (prev. Methodist)	2
TOTAL	23

What is the most popular religious affiliation of Canadian Prime Ministers as of 2019? Or, what religious affiliation is most frequently reported by Canadian Prime Ministers so far? In other words, what religious affiliation do Canadian Prime Ministers most have tended to have?

Surprising no one with any knowledge about Canada, the largest category among the religious denominations, or the one that Canadian Prime Ministers most frequently subscribe to — i.e., **the category with the highest frequency** — is “Roman Catholic”, with 10 of the Canadian Prime Ministers identified as such. (And are you surprised that Canada has only had Christian Prime Ministers?)

As simple as that, **the category/value with the highest frequency is called the mode of the variable.**

Alternatively, you can easily spot the mode in a graph: it would be the largest slice of the pie or the tallest column in a bar chart or a histogram.

*Do It! 3.1 Do all variables have a mode?*

Considering that the only thing you need to do to find a variable's mode is to count the frequency of each of its categories/values and indicate the one with the highest count, will it be possible to find the mode of any variable, regardless of its level of measurement? Or would the mode be a descriptive statistics applicable only to some variables depending on their level of measurement?

If by now you have a good grasp of what makes a variable nominal, ordinal, interval, or ratio (and if you do not — go back and really reread Section 1.3! (<https://pressbooks.bccampus.ca/simplestats/chapter/1-3-levels-of-measurement/>)), you should be able to easily answer the questions in the *Do It! 3.1* above. Obtaining the mode, the simplest of all measures of central tendency, does not require any calculations or complicated procedures. To identify the mode, it doesn't matter whether the categories of a variable are made of *words* or *numbers*, or if there is any order in them. All that matters is the *count* — the frequency — of responses in each category/value in order to identify *where cases tend to cluster* across the categories/values. As such, **the mode is a descriptive statistic applicable to any and all variables.**

To illustrate, let’s bring back the Example 2.2 (B) from Section 2.3:

*Do It! 3.2 Educational Attainment’s Mode*

*Table 3.2 Educational Attainment*

Degree	Count (a.k.a. frequency)
No degree	1
Secondary/High School	6
Associate’s	3
Bachelor’s	5
Master’s	2
PhD	1
Didn’t answer	3
<b>TOTAL</b>	<b>21</b>

What is the mode for educational attainment based on the 21 respondents in the example?

Looking for the largest category in Table 3.2 above, you undoubtedly already identified that the mode for



*educational attainment* is “Secondary/High School”. That is, to put this into language that even people non-trained in statistics could understand, the most frequent educational degree among the 21 respondents in the example is “Secondary/High School” as it has the highest frequency/ the largest number of cases in it, 6. (It is generally quite useful to get into the habit of translating *statistics-ese* into English when you write reports so you should practice it on all occasions.)<sup>1</sup>

And this is all there is to finding out a variable’s mode. Beyond simply counting (applicable to groups of relatively small size, as generally no one would want to count hundreds or thousands of cases by hand), the ways to obtain a mode through SPSS are listed below.

### SPSS Tip 3.1: Finding a Variable’s Mode

#### Option 1: Through a frequency table<sup>2</sup>

- Use SPSS to create a frequency table for your

1. Note that *most frequent* category does not mean that it contains the *majority* or *most* cases. Sometimes that may be the case, but it’s not necessarily so. In both examples above you can see that neither Roman Catholics nor people with Secondary/High School degrees are a majority in their respective groups (10 out of 23 and 6 out of 21, respectively). Thus, be careful when writing about a mode as being “where *most/the majority* of cases cluster” because many times the phrasing would be factually incorrect.
2. You might want to avoid this option when working with interval/ratio variables, as their frequency tables can be very, very long.

chosen variable<sup>3</sup>;

- Look for the category/value with the highest frequency (the relative frequency in the *Valid Percent* column works too);
- Report the category with the highest frequency as the mode of that variable.

### **Option 2: Directly requesting the statistic**

- From the *Main Menu*, select *Analyze*, then *Descriptive Statistics*, then *Frequencies*;
- Select your variable of choice from the left-hand side and use the arrow to move it to the right side of the window;
- Click on the *Statistics* button on the right;
- In the new window, check *Mode* in the *Central Tendency* section on your right;
- Click *Continue*, then *OK*.

Note that SPSS gives you the option to display a frequency table or not before clicking *OK* in the last step listed in the SPSS Tip above. The reason is practical: the frequency tables of interval/ratio variables can be quite long depending on the number of values they contain. As such, while identifying the mode from the frequency table of a nominal or ordinal variable is fine, it's often more practical to request SPSS to report the mode of an interval/

3. See Section 2.3.4 (<https://pressbooks.bccampus.ca/simplestats/chapter/2-3-4-what-frequency-tables-look-like/>) for the tip on how to create frequency tables in SPSS.

ratio variable directly rather than through a frequency table.

**Watch Out!! #6... for Reporting Nominal/Ordinal Variable's Modes As Given by SPSS**

One thing to keep in mind when requesting the mode directly from SPSS is that SPSS will report modes by their number labels, or code (i.e., not by the actual name of the categories). If you recall from Section 2.1 (<https://pressbooks.bccampus.ca/simplestats/chapter/2-1-data/>), datasets contain only numbers, with nominal and ordinal categories appearing in code so that the software can work with them. As such, your SPSS output will list the mode of a nominal or ordinal variable as a number, and it is your job to “translate” that number into its proper form, i.e., its the actual category.

For example, in the *Religious Affiliation of Canadian Prime Ministers* example above, going in the order the categories are listed, the categories would typically be coded in the following way: “Anglican” = 1, “Baptist” = 2, “Evangelical” = 3, “Presbyterian” = 4, “Roman Catholic” = 5, “United Church of Canada” = 6. The dataset would contain only the code (i.e., the numbers) and SPSS would report the mode as “5” in the output.

However, it is a mistake to report the code (the number label assigned to the category) instead of the actual

category's name. **You should always report the mode with its real category name.** (That is, it is up to you too look up the code — recall that you can do this through the *Values* column in SPSS's *Data View* — and find the correct name of the modal category). In this case, you should report the mode of *Religious Affiliation of Canadian Prime Ministers* not as 5 but as “Roman Catholic”. (The “5” has no real meanings, it simply indicates that Roman Catholic is the fifth category in the listing.)

I'll end this section with a final consideration regarding the mode: it is quite possible that a variable has more than one mode. After all, two (or more) categories/values might have the same frequency, so in that case we say that the variable's distribution is *multimodal* (*bi-modal* or *tri-modal* in the specific cases of two or three modes). Depending on the number of modes, it's acceptable to report only the first, while indicating that multiple modes exist for that variable. Multiple modes are usually also easy to spot in bar graphs and histograms: they appear as bars of equal height.

---

## 3.2 Median

The three measures of central tendency are all measures that tell us where typical cases fall or where cases tend to cluster. After exploring the mode in the previous section, in this section we turn to the second measure of central tendency called the *median*.

The *median* lives up to its name: it derives from the Latin root *medi*, meaning “middle”, and that’s exactly the type of information it provides. Specifically, the median divides the cases of a variable into two equal halves and identifies the case in the middle. As such, it points out the “centre” of the data in a very straightforward way — it simply reports the middle observation.

Consider, however, the following point: even in everyday life, the middle implies a beginning and an end (e.g., “in the middle of the book”); something that is in-between, a gradation from a point A to a point C, as it were. From clothes sizes (“small, *medium*, large”) to how spicy you like your Thai food (“a little, *medium*, or hot”), through turning the volume up or down while listening to music (“low, *medium*, high”), the “centre” category bisects whatever it is applied to into a smaller/larger, less/more, left/right, etc. parts. That is, to speak of *the middle* of something we need to know where it starts (e.g., the minimum) and where it ends (e.g., the maximum). Simply put, we need an *order*.

What all this should tell you is that **the median is not**

**applicable to nominal variables.** Speaking of the middle of *gender*, or the middle of *ethnicity*, or *religious affiliation*, or *hair colour*, or *degree major*, or of the middle of any other nominal variable makes no sense. After all, the order the categories of a nominal variable appear is either arbitrary or a matter of preference; nothing precludes rearranging the categories in some *other* way so that a case belonging to one category that ends up in the middle of one arrangement would not necessarily be in the middle of another arrangement. A case belonging to *any* category can easily end up being the middle one. A statistic shouldn't depend on such a chance/preference; as such **nominal variables have no median.**

On the other hand, as you know by now, ordinal and interval/ratio variables do have an inherent order arranging their categories/values. They have a “beginning” and an “end”, and therefore a “centre”. As such, **the median applies (only) to ordinal and interval/ratio variables.**

Note that while the mode applies to a *category* (reflecting the largest number of cases), the median is determined by the *case* (observation) that falls in the middle of the category-ordered listing of all cases. Thus **it's not the middle category that is the median**; depending on the size of the categories, the median *case* can belong to any category/value. **The median category/value is the one to which the middle case belongs.** Presented this way, the explanation sounds undoubtedly as clear as mud but do not despair. It will get better when we establish the manner in which we obtain the median, so trust me and read on.

*Example 3.2 (A) Three Students, Five Students, Eight Students by Year of Study, Counting*

$$N=3$$

a) Let's say we have three students at different levels of their studies: one is a first-year, the second one a fourth-year, and the third a third-year. Before we do anything else, we need to establish the correct order. We rearrange the students properly:

- (1) a first-year student
- (2) a third-year student      ← *median*
- (3) a fourth-year student

The case in the middle is Case #2, the second one on the list (as there is one student below and one student above), i.e., the third-year student. Thus we have established that the median category is "third year of study". That is, half of the students are below the third year of study and half are above (as odd as it sounds when we only have three cases).

$$N=5$$

b) What happens if I add two more students to our group, say, a first-year student and a second-year student? The order will go like this:

- (1) a first-year student

- (2) a first-year student (new)
- (3) a second-year student (new) ← *median*
- (4) a third-year student
- (5) a fourth-year student

Once again, it's easy to see that the middle case is Case #3, the third one on the list (as there are two students below and two students above), i.e., the second-year student. This time around the median category is "second-year of study". That is, half of the students are below their second year of study and half are above.

$N=8$

c) What if I complicate matters further? What if I add three more students to the group, say, two second-years and a fourth-year? Their order will be:

- (1) a first-year student
- (2) a first-year student
- (3) a second-year student      *The median is between*
- (4) a second-year student (new) ← *this case*
- (5) a second-year student (new) ← *and this case*
- (6) a third-year student
- (7) a third-year student (new)
- (8) a fourth-year student

If you go by the same logic as above, you'll quickly find that there is no "middle" student: unlike before, the students



now are an even number. The middle of the group actually falls between Cases #4 and Case #5, the fourth and the fifth cases on the list (so that four are below and four above it). Since both the fourth and the fifth students are second-year, we can conclude that, again, the median is “second-year of study”. Had the fourth and the fifth student been different years of study, we could say that the median was between their respective categories.

We could continue the same way as in Example 3.2 (A) above for larger groups too: we could arrange the cases in order of their categories/values, find the middle case (or two middle cases) and report its category/value as the median. However, you can guess that this would quickly become impractical the larger the group size gets. We need some other way of finding the median, one that generalizes across groups of any size.

Consider the following formula:

$$\frac{N + 1}{2} =$$

*“numbered position of the median case in the ordered list of cases”*

where, as usual, N is the group size.

Instead of counting, let’s apply this formula to Example 3.2 (A).

*Example 3.2 (B) Three Students, Five Students, Eight Students by Year of Study, Using a Formula*

a)  $N=3$

- (1) a first-year student
- (2) a third-year student
- (3) a fourth-year student

According to the formula,

$$\frac{N + 1}{2} = \frac{3 + 1}{2} = \frac{4}{2} = 2$$

That is, the “numbered position of the median case in the ordered list of cases” is equal to 2; the middle case is Case #2, the second one on the list, or like we established before, the third-year student.

b)  $N=5$

- (1) a first-year student
- (2) a first-year student (new)
- (3) a second-year student (new)

(4) a third-year student

(5) a fourth-year student

According to the formula,

$$\frac{N + 1}{2} = \frac{5 + 1}{2} = \frac{6}{2} = 3$$

That is, the “numbered position of the median case in the ordered list of cases” is equal to 3; the middle case is Case #3, the third one on the list, or again, the second-year student.

c)  $N=8$

(1) a first-year student

(2) a first-year student

(3) a second-year student

(4) a second-year student (new)

(5) a second-year student (new)

(6) a third-year student

(7) a third-year student (new)

(8) a fourth-year student

According to the formula,

$$\frac{N + 1}{2} = \frac{8 + 1}{2} = \frac{9}{2} = 4.5$$

That is, the “numbered position of the median case in the ordered list of cases” is equal to 4.5. Considering we have discrete numbers (after all, the cases are individuals), there is no case number 4.5. Instead, we say that the median falls between Case #4 and Case #5, the fourth and fifth cases on the list, or between two second-year students, so it is “second year of study”.

It is easy to see that we could substitute a group of any size for the  $N$  in the formula. Even when working with hundreds or thousands of cases, we can always use the formula to find the place (or which case number) bisects the variable’s distribution in two halves.

So far I only used an ordinal variable to illustrate the median. How does finding the median work for interval/ratio variables? Would it matter that interval/ratio variables have numerical values rather than qualitative categories? No, not in the least. After all, finding the median doesn’t depend on the category or value of any case in any substantive sense — only on its numbered position in the ordered list of categories/values.

There is something a bit different in the way interval/ratio variables look, however, that some people seem to find a tad more confusing when working with values rather than categories. To illustrate, I’ll give you another example.


*Example 3.3 (A) Median for Number of Siblings, Raw Data*

Imagine you talk to seven of your friends and ask them about the number of siblings they have. Let's say these are the responses you receive: 2, 1, 4, 2, 1, 0, 3. That is, two friends report having two siblings each, two friends report having one sibling each, and three of your friends report having four, zero, and three siblings each.

To find the median, the first thing we need to do is put the responses in order:

- (1) 0
- (2) 1
- (3) 1
- (4) 2
- (5) 2
- (6) 3
- (7) 4

Whether you visually identify Case #4 as the middle case (three cases below and three cases above it) or use the formula ( $\frac{N+1}{2} = \frac{7+1}{2} = \frac{8}{2} = 4$ ) to obtain the same result, it is clear that the median is “two siblings”: half of your friends in this example have fewer than two siblings, and half have two or more siblings.



What might be confusing for some people is differentiating between the numbered positions of the cases on the list and their values since both are expressed numerically. In this example I have tried to make it easier to distinguish by putting the numbered positions of the cases in brackets and the values next to them (just like the categories in the ordinal example above). Thus you can see that Case #1 has 0 siblings, Case #2 has 1 sibling, etc. Had I chosen different set of values — for example, if Case #1 had 1 sibling, Case #2 had 2 siblings, Case #3 had 3 siblings, etc. — you might have found it a bit harder. For that reason, make a mental note to keep a clear track of what is a case's value and what is its numbered position.

---

# 3.3 The Median With Frequency Tables and Other Considerations

A similar — though far more widespread confusion – may happen when working with frequency tables. Frequency tables, as you know from Section 2.3.3 (<https://pressbooks.bccampus.ca/simplestats/chapter/2-3-3-summing-up-adding-cumulative-percentages/>), list a variable’s categories/values in the first column and their frequencies in the second column. Take a look at the incomplete frequency table of the fictitious *number of siblings* variable used from before.

*Example 3.3 (B) Number of Siblings, Aggregated*

*Table 3.3 Frequency Table for Number of Siblings*

Value	Frequency
0	1
1	2
2	2
3	1
4	1
<b>Total</b>	7

Can you as easily see that one of your (imaginary) friends has zero siblings, two of your (imaginary) friends have one sibling each, another two of them have two siblings each, etc.? While Table 3.3 presents the same information as Example 3.3 (A) in the previous section does, the way the data is organized is different, so again, make sure you differentiate the variable's values (first column) from the values' frequencies (second column).

A further consideration is finding the median itself. While we saw that the mode depended only on identifying the category/value with the highest frequency (and it was therefore just a matter of finding the largest number in the *Frequency* column of a frequency table), are you able to determine the median from the partial frequency table in Example 3.3 (B) above? I would venture that the answer would be “no” for most readers.

Of course, you can find a solution to our median-finding



problem by “unpacking” the frequency column from Table 3.3 and reverting to raw (uncategorized) data again: one 0, two 1’s, two 2’s, one 3, and one 4 are 0, 1, 1, 2, 2, 3, 4. We already established (both visually and through using the position-of-the-median formula) that the middle case was Case #4, or “two siblings”. Would you like, however, to do that for the following Table 3.4?

*Table 3.4 Household Size of the Respondent (GSS 2016)*<sup>1</sup>

Household size of respondent					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	One person household	5462	27.9	27.9	27.9
	Two person household	7432	37.9	37.9	65.8
	Three person household	2803	14.3	14.3	80.0
	Four person household	2580	13.2	13.2	93.2
	Five person household	906	4.6	4.6	97.8
	Six or more person household	426	2.2	2.2	100.0
	Total	19609	100.0	100.0	

Most likely, you wouldn’t “unpack” the 19,609 cases into raw data, so we should seek some other — and more

1. Note that this variable is technically an ordinal variable. Despite the numerical values and equal “distances” (of *one person*) between the first five categories, the last category “Six or more person household” prevents us from categorizing the variable as ratio. After all, we don’t know exactly how many individuals live with any of the 426 people in that category: it could be six, or seven, or eight, etc. Thus it is not possible to say how many more persons live in the households of the respondents in the last category compared to any of the preceding categories: the “distance” is no longer *one person*. Any interval/ratio variable that has its last category truncated in this way (i.e., it has “... or more” in its label) becomes technically ordinal. Nevertheless, for heuristic purposes I will ignore the “...or more” part in this example which allows me to assume that everyone in that last category lives in a six-person household. This, in turn, allows me to pretend the variable is a ratio. However, the example works the same way regardless if the variable is truly ordinal or ratio.

generalizable — method for finding the median through frequency tables, one that would apply to  $N$  of any size.

We could, of course, use the formula to at least establish the middle case's numbered position, and then work our way through the table to identify the median.

$$\frac{N + 1}{2} = \frac{19,609 + 1}{2} = \frac{19,610}{2} = 9,805$$

That is, Case #9,805's household size will be the median household size for these almost 20 thousand respondents.

How do we find it? There are 5,462 respondents who reported living alone ("one person household") so we know that Case #5,462 does not "reach" the median yet, thus we have to count further. We take the next 7,432 respondents who reported living in two person households, but we need to add them to the 5,462 people living alone in order to obtain the second group's case number positions. After all, the case count for the 7,432 respondents does not start from 1 but from 5,463, and Case #5,463 will already be living in a two person household. So will Case #5,464, Case #5,465, etc. ... all the way up to Case #12,894 (because  $5,462 + 7,432 = 12,894$ ), which will be the last respondent living in a two person household.

However, we now see that we have "counted" too far ahead — we have jumped not to Case #9,805 but all the way to Case #12,894! We do know though that all cases between Case #4,463 and Case #12,894 live in two person households: this is enough for us to establish that Case #9,805 lives in a two person household as well.

In short, the median household size of the 19,609 respondents is two-persons household. That is, half of the respondents live in two-person or smaller households and half of them live in two-person or larger households.

*Hmm, I hear you say, this is still quite the roundabout way of getting to the median — can you do better?*

Alright, let's think of something else then. We tried adding the frequencies together until we reached the median... How about we try using percentages this time around — and more to the point, *cumulative* percentages, as they are already keeping a running total? We just need to know which percent corresponds to the middle case.

Recall, then, that the middle case splits the distribution of the cases in *two equal halves*. What percent is half of something? Of course, 50 percent. Thus it would make sense to simply look at the *Cumulative Percent* column and try to figure out where 50 percent would fall. The respondents living alone comprise 27.9 percent, so too low for the median, but the respondents living in one or two person households *added together* comprise already 65.8 percent of the total. Following the same logic as with the frequencies, the 50th percent falls within the one/two person household cumulative group. However, we know it's not within the one person household group. That means the 50th percent can only fall within the respondents living in two person household, which, again confirms what we already knew: the median household size is made up of two persons.

To generalize, if you'd rather not use the formula for the median's position and add the frequencies of a frequency

table up in order to find the median, you can always simply look for within which category/value the 50th percent would fall. That category/value will be the median one.

### *Do It! 3.3 Median Workplace Size*

Let's revisit Table 2.6 from Section 2.3.4 (<https://pressbooks.bccampus.ca/simplestats/chapter/2-3-4-what-frequency-tables-look-like/>). Can you identify the median of workplace size? And since you're at it anyway, what about the mode?

Imagine you have to tell what you have found to some of your friends who have no knowledge of statistics. How are you going to explain to them your findings about the mode and the median of *workplace size*?

*Table 2.6 Frequency Table for Workplace Size of the Respondent (GSS 2016)*

Workplace size					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Small business	6409	32.7	62.2	62.2
	Midsized business	2165	11.0	21.0	83.2
	Large business	1732	8.8	16.8	100.0
	Total	10306	52.6	100.0	
Missing	Valid skip	9102	46.4		
	Don't know	153	.8		
	Refusal	20	.1		
	Not stated	28	.1		
	Total	9303	47.4		
Total		19609	100.0		

Finally, now that you have learned what the median is and how you can find it, I will also casually mention that you can use SPSS for that. (Okay, okay. Don't throw bricks, please: it really is important to work through the examples and exercises manually so that you understand what the SPSS output tells you and so that you are able to interpret that output properly.)

#### *SPSS Tip 3.3 Finding the Median Of a Variable*

- From the *Main Menu*, select *Analyze*, then *Descriptive Statistics*, then *Frequencies*;
- Select your variable of choice from the list on the left and use the arrow to move it to the right

side of the window;

- Click on the *Statistics* button on the right;
- In this new window, check *Median* of the *Central Tendency* section on your right;
- Click *Continue*, then *OK*.
- The *Output* window will provide a small table listing the median of the selected variable.

Keep in mind that the *Watch Out!! #6* warning from Section 3.1 about the mode applies equally to the median: **for ordinal variables, SPSS will provide the median in numerical code. It is your job to “translate” the code into the actual category’s name.** In the case of *household size* SPSS supplies “2” as the median, which stands for “two person household”. Thus we say that the median household is a two-person one; we do *not* report that the median household is “2”.

**Watch Out!! #7...** for Misinterpreting the Formula for the Median

An extremely common mistake regarding the median is to take the result of  $\frac{N+1}{2}$  to be equal to the median itself. This is patently not true. Again, what the formula provides

is the *place* (or the *numbered position*) of the median once the cases have been put in their correct order:

$$\frac{N + 1}{2} =$$

“*numbered **position of the median case** in the ordered list of cases*”

Thus, once your calculation for the place of the median is done, do not forget to do the final step: check the position you have calculated and see what the *category/value* of the median case is. **You need to report only that value, not the position itself.**

**Stability of the median.** A final noteworthy observation about the median is its *stability as a measure of central tendency*. Since the median is entirely about the central position in a variable’s distribution and all it takes into account is the *order* of the cases, *not* their substantive *values*, **it’s impervious to the actual magnitude of the values**. Thus it doesn’t matter if we have a set of values like 1, 5, 20, or one like 4, 5, 6, or another like 0, 5, 9 — the median is the same for all three, even if the values in the sets are different. Whether we have a small or a large value is immaterial, all that it matters is where the value goes into the order of the variable’s cases.

You will learn why this has important implications for the central tendency in the next section, all devoted to the mean.





---

## 3.4 Mean

The third, and final, measure of central tendency is one you have undoubtedly encountered before. It is one that most people have had to calculate at least a few times in their lives, and that everyone has heard reported about one thing or another. You most likely know it by its common name, **the average**.

Recall that the measures of central tendency provide information about the typical cases, or where cases tend to centre in a variable's distribution. Thus a student's Grade Point Average (GPA) provides a measure for how well they do academically, not in one class, but *on average*, across all of them; a hockey player's points season average provides a measure of their performance on the ice not just in one game but for a whole season; a monthly average temperature gives indication of what the typical weather for a specific month is, etc. All of these averages show what is typical or expected.

**The mean of a variable** is therefore, quite simply put, **the mathematical average of the values of the variable's cases**. Reported alongside the mode and the median, it provides a fuller picture of where the cases tend to cluster, or what the typical cases are. The mode does this in the simplest way, by counting their frequency and reporting the largest one. The median does that by providing the most centrally located case in terms of order.

**Unlike the mode and the median, however, the mean takes into account the actual *values* of the cases.**

Keeping the last sentence in mind, do you think the mean will apply to all and any variables? If you have been paying attention, you would know that the answer is “no, of course not”.

Nominal and ordinal variables have categories. **Only interval/ratio variables have actual numerical values, therefore, the mean applies only to them.** After all, mathematical calculations are only possible when we have *numbers* with which to do the calculations: we cannot calculate an average of gender, or of race/ethnicity, or of religious affiliation, etc.<sup>1</sup> We could, however, calculate an average age, income, score, temperature, etc.

If you had ever calculated your GPA, you already know how to calculate the mean. I will still give you an example to strengthen your knowledge.

*Example 3.4 (A) Mean of Number of Siblings, Raw data*

1. Note that in specific cases it's possible to calculate *something like an average* for certain ordinal variables, for example, Likert-scales, to the extent that their numerical labels reflect a somewhat monotonic, stable-unit, distances. This should be done with extreme care and ample justification, however, and beginner researchers (like you) are advised against using means for ordinal variables.

If you recall our Example 3.3 (A) from the previous Section 3.2 ( <https://pressbooks.bccampus.ca/simplestats/chapter/3-2-median/> ), you imagined yourself asking seven of your friends about the number of siblings they had. We imagined the responses as follows: 2, 1, 4, 2, 1, 0, 3. We had to put these values in order to be able to find the median, but the mean works either way, whether the values are in order or not.

To calculate the average number of siblings your imagined friends have, we simply add all responses together and divide them by the total number of friends, i.e., by 7:

$$\frac{(2 + 1 + 4 + 2 + 1 + 0 + 3)}{7} = \frac{13}{7} = 1.86$$

That is, your imagined friends have 1.86 siblings on average (or not quite but closer to two, rather than one siblings on average). We could also say that the mean of *number of siblings* is 1.86.

Let's do it again, as practice makes perfect.

*Example 3.5 Textbook Prices For a Semester, Raw Data*

Depending on the courses you take in a semester, what you pay for books will vary but let's say we're interested in how much you pay for books in a typical semester. Perhaps you are very-well organized and want to finish your degree as quickly as possible so you have decided to take five courses per semester. For simplicity's sake, let's assume you were assigned one book per course. These are the books' prices: \$120, \$230, \$300, \$65, \$30. How much did you pay for a book on average?

$$\frac{(120 + 230 + 300 + 65 + 30)}{5} = \frac{745}{5} = 149$$

That is, despite the fact that some of your books were expensive (like the \$300 one), and some relatively cheap (like the \$30 one), the average price you paid for a book in that semester was \$149.

Now that we've seen how the mean works in practice, let's generalize what we did in the two examples above using proper notation. Fair warning: the formula below does *look* complicated but remember what we just did: our calculations were quite simple (adding all values, dividing their sum by their total number), and so is the formula. As usual, it simply restates what we've said in words in

a mathematical shorthand. If you know what each symbol in the shorthand stands for, you know what the formula means. So, take a deep breath:

$$(1) \quad \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N} = \frac{\sum_{i=1}^N x_i}{N} = \bar{x}$$

where  $\sum$  stands for “sum”,  $\sum_{i=1}^N$  indicates to sum all cases from the first (1) to the last ( $N$ ),  $x_i$  stands for any case with a number between 1 and  $N$ , and  $\bar{x}$  indicates the mean<sup>3</sup>, i.e., the average of all the  $x_i$ ’s. Thus, the formula basically tells you to add all values and divide by their total, just as we did in the examples.

So far, we only calculated the means for raw data, i.e., data not presented in a frequency table. Will the calculation of the mean be different if we had a frequency table instead? While the principle is the same, the fact that the values are grouped by frequency in frequency tables requires that we do a slight modification to our calculations. Here’s a small-scale illustration to demonstrate the principle before we do an example with a larger  $N$ .

*Example 3.4 (B) Mean for Number of Siblings, Aggregated Data*

2.  $\sum$  is pronounced “SIG-ma” and is the Greek letter S.

3.  $\bar{x}$  is pronounced “EX-bar”.

Arranging the raw data from Example 3.4 (A) above, we again get the following table.

*Table 3.3 Frequency Table for Number of Siblings*

<b>Value</b>	<b>Frequency</b>
0	1
1	2
2	2
3	1
4	1
<b>Total</b>	<b>7</b>

According to the formula for the mean, we need to add all values together and then divide their sum by their total number. When the values are disaggregated (i.e., raw), we can proceed to adding them up right away. However, when they are grouped by frequency, we first need to multiply each value by its respective frequency, and then add the value-times-frequency products together, before dividing them by the total number, like this:

$$\frac{\sum_{i=1}^N x_i}{N} = \frac{(0 + 1 + 1 + 2 + 2 + 3 + 4)}{7} = \frac{0(1) + 1(2) + 2(2) + 3(1) + 4(1)}{7} = \frac{13}{7} = 1.86 = \bar{x}$$

Again, the average number of siblings of these seven friends is 1.86, as previously calculated.

Now let's apply the same principle to a new, larger- $N$  example.

#### *Example 3.6 Age of Classmates, Aggregated Data*

Imagine you are doing a survey for one of your class assignments and one of the questions is about age. You aggregate the data by frequency and you get the following table.

*Table 3.5 Mean for Age of Classmates*

Value	Frequency
19	1
20	10
21	12
22	8
25	2
27	1
35	1
<b>TOTAL</b>	<b>35</b>

By the formula, we have:

$$\frac{\sum_{i=1}^N x_i}{N} = \frac{19(1)+20(10)+21(12)+22(8)+25(2)+27(1)+35(1)}{35} = \frac{19+200+252+176+50+27+35}{35} = \frac{759}{35} = 21.69 = \bar{x}$$

Or, now you know that the average age of your classmates in that class is 21.69 years, or a bit less than 22 years.



---

## 3.5 The Mean With Existing Data and Other Considerations

Let's work through some real-world data, this time from the Canadian Community Health Survey 2015-2016 (Statistics Canada 2017), a.k.a. *CCHS 15/16*, a very large dataset containing information on more than 100,000 respondents.

*Table 3.6 Number of Times the Respondent Consulted a Mental Health*

*Professional in the Last 12 Months (CCHS 15/16)*

**Consulted mental health professional - num of times - 12 mo**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	3778	3.4	24.4	24.4
	2	2851	2.6	18.4	42.9
	3	1700	1.6	11.0	53.9
	4	1426	1.3	9.2	63.1
	5	778	.7	5.0	68.1
	6	1008	.9	6.5	74.6
	7	205	.2	1.3	76.0
	8	357	.3	2.3	78.3
	9	66	.1	.4	78.7
	10	534	.5	3.5	82.2
	11	24	.0	.2	82.3
	12	2735	2.5	17.7	100.0
	Total	15462	14.1	100.0	
Missing	Valid skip	90887	82.9		
	Not stated	3310	3.0		
	Total	94197	85.9		
Total		109659	100.0		

To calculate how many times Canadians consulted a mental health professional in the last year preceding their participation in the survey based on the data above, we need to follow the principle we used in the *age of classmates* and *number of siblings* examples in the previous section.

Specifically, we need to multiply each value (1 through 12 number of times a mental health professional was seen) by its frequency, then to sum all the products together, and finally to divide the sum on the total number of respondents, 15,462 (recall that we only use valid cases for analysis and exclude the missing ones).

$$\begin{aligned}
& \frac{\sum_{i=1}^N x_i}{N} = \\
& = \frac{1(3778) + 2(2851) + 3(1700) + 4(1426) + 5(778) + 6(1008)}{15462} + \\
& + \frac{7(205) + 8(357) + 9(66) + 10(534) + 11(24) + 12(2735)}{15462} = \\
& = \frac{3778 + 5702 + 5100 + 5704 + 3890 + 6048}{15462} + \\
& + \frac{1435 + 2856 + 594 + 5340 + 264 + 32820}{15462} = \\
& = \frac{73531}{15462} = 4.76 = \bar{x}
\end{aligned}$$

That is, we have found that the respondents on average consulted a mental health professional 4.76 times over the 12 months preceding the survey.

*Do It! 3.4 How Many Times Has The Respondent Stopped Smoking for at Least 24 hrs In the Past 12 Months (CCHS 15/16)*

To save you you from calculating into the thousands, here is a variable based on a question that 99.9 percent of the respondents did not have to answer, which gives you a manageable  $N=106$ . Calculate the average number of times respondents have stopped smoking for at least 24 hrs for the 12 months preceding the survey. While you're at it, find and report the mode and median of this variable.

*Table 3.7 Number of Times Respondent Stopped Smoking In the Past Year (CCHS 15/16)*

Stopped smoking for at least 24 hours - num of times - 12 mo					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	27	.0	25.5	25.5
	2	21	.0	19.8	45.3
	3	20	.0	18.9	64.2
	4	12	.0	11.3	75.5
	5	1	.0	.9	76.4
	6	10	.0	9.4	85.8
	7	3	.0	2.8	88.7
	8	1	.0	.9	89.6
	10	4	.0	3.8	93.4
	12	2	.0	1.9	95.3
	20	2	.0	1.9	97.2
	30	1	.0	.9	98.1
	52	1	.0	.9	99.1
	95	1	.0	.9	100.0
	Total	106	.1	100.0	
Missing	Valid skip	109538	99.9		
	Don't know	11	.0		
	Not stated	4	.0		
	Total	109553	99.9		
Total		109659	100.0		

I strongly encourage you to do the above exercise yourself. Still, as usual, here is an SPSS tip on how to obtain a mean in SPSS.

*SPSS Tip 3.4 Obtaining the Mean*

- From the *Main Menu*, select *Analyze*, then *Descriptive Statistics*, and then *Frequencies*;
- Select your variable of choice from the list on the left and use the arrow to move it to the right side of the window;
- Click on the *Statistics* button on the right;
- In this new window, check *Mean* in the *Central Tendency* section on your right;
- Click *Continue*, then *OK*.
- The Output window will provide a small table listing the selected variable's mean.

---

## 3.6 Outliers

Out of the three measures of central tendency, the mean is the only one that takes into account the actual numerical values of the cases. As such, it is easily affected by the size of the values: a sequence of numbers such as “1, 5, 7, 10, 15” will produce a smaller mean than a sequence of numbers like “100, 50, 75, 130, 90”.

When all values to be averaged are of relatively comparable magnitude, the mean does a good job at reflecting the central tendency of a variable — that is why it is the most familiar and widely used measure. However, **when a variable contains an extremely small or an extremely large value (or several values) compared to the rest of the values, the mean gets easily distorted and stops reflecting the central tendency “truthfully”, as it were. Extremely small and extremely large values are called statistical outliers.**

While there is a convenient method for identifying outliers (using a concept called *interquartile range* which we will discuss in the next chapter), at this stage it is not necessary that you be so technical. You can visually identify outliers, albeit less precisely, by the “disturbance” in the general pattern of the data you observe. For example, if you have values like “1, 5, 7, 10, 15”, a value of 130 in that sequence would be considered an

outlier. Similarly, if you have values like “100, 80, 75, 130, 90”, a value of 5 would be an outlier.

Let’s calculate the means of the two sequences, first with and then without the so-called outliers and see what happens.

The first sequence is 1, 5, 7, 10, 15 and we want to see what happens when we add 130.

$$\frac{(1 + 5 + 7 + 10 + 15)}{5} = \frac{38}{5} = 7.6$$

We add 130 to the sequence:

$$\frac{(1 + 5 + 7 + 10 + 15 + 130)}{6} = \frac{168}{6} = 28$$

Both means, 7.6 and 28, are the true averages of the sequences of values as listed. However, the addition of an uncommonly large number “pulled” the mean away from the “centre” of the original data.

How truthfully does 28 represent the “centre” of a sequence where the majority of the cases’s values (in fact, five out of the six values) are 15 and below? Not that much.<sup>1</sup>

1. If you believe it's not the magnitude of the value but just its addition that causes the "pulling" of the mean, consider redoing the example with adding 18, instead of 130. Then we have  

$$\frac{(1+5+7+10+15+18)}{6} = \frac{56}{6} = 9.3$$
 The "pull" from 7.6 to 9.3 is much smaller than from 7.6 to 28. The value 9.3 reflects the central tendency of the data more truthfully than 28 does.



To demonstrate the effect of an extremely small value, we continuing with the next sequence:

$$\frac{(100 + 80 + 75 + 130 + 90)}{5} = \frac{475}{5} = 95$$

Adding a value of 5 to the sequence produces the following:

$$\frac{(100 + 80 + 75 + 130 + 90 + 5)}{6} = \frac{460}{6} = 80$$

Similarly as with the effect on the mean of the first sequence, the mean here gets “pulled”, but in the opposite direction, from 95 to 80. Both means are technically true averages of their respective values but the latter one is “artificially” low: after all, four out of the six values are the same or higher.<sup>2</sup>

What this tells you is that **the mean is an unstable measure of central tendency, prone to being affected by outliers**. Contrast this to what you know about the median: the median does not take the magnitude of the values into consideration, beyond their order. Thus, as explained in the previous Section 3.3 (<https://pressbooks.bccampus.ca/simplestats/chapter/3-3-the-median-with-frequency-tables/>), adding a value (be it extremely small or extremely large) to a sequence does not affect the median much —

2. Again, if we added a value of a comparable size to this sequence instead of 5, the mean would not be impacted as much:

$$\frac{(100+80+75+130+90+70)}{6} = \frac{545}{6} = 90.8.$$

Consider the “pull” from 95 to 80 vs. from 95 to 90.8.

unlike the mean. The median of 1, 5, 7, 10, 15 is 7 (there are two values above and two below it), and whether we add 130 or 18, it doesn't matter: it's just an additional value in the sequence.<sup>3</sup>

Since the mean is prone to being affected by outliers, while the median is not, **in some situations it is advisable to report the median as a more “valid” measure of the typical cases/”centre” of the data rather than the mean.** Specifically, watch out for reports on average income, average age, average weight, etc. where a few outliers can *skew* a variable's distribution.

**Watch Out!! #8 ... for Reports on Averages of Variables Prone to Skewing by Outliers**

Imagine a small company advertising an open position by claiming that the average salary of their employees is 100 thousand dollars per year. For simplicity's sake, let's assume the company has ten employees and these are their salaries:

*Table 3.8 Employee Salaries (Hypothetical Data)*

3. The median of 1, 5, 7, 10, 15, 18 is between 7 and 10, i.e., 8.5 (since we need the half-way distance between 7 and 10, we use the average of 7 and 10, that is  $7+10=17$  and divide it by 2 to get 8.5). The median of 1, 5, 7, 10, 15, 130 is exactly the same -- it is still half-way between the two middle values, 7 and 10, or again 8.5.

Value (in thousands)	Frequency
70	5
87.5	4
300	1
<b>TOTAL</b>	<b>10</b>

You can check for yourself what the average annual salary is:

$$\frac{\sum_{i=1}^N x_i}{N} = \frac{70(5) + 87.5(4) + 300(1)}{10} = \frac{350 + 350 + 300}{10} = \frac{1000}{10} = 100$$

or, indeed, 100 thousand dollars. However, how representative this annual salary is for the regular employee? After all, nine out of ten employees of the company get less than that. The average annual salary reported is inflated by the very high salary of one employee (perhaps the manager), a clear outlier.

Let's instead look at the median. We start by arrange the values in order:

70, 70, 70, 70, 70, 87.5, 87.5, 87.5, 87.5, 300

Using the formula for finding the position of the median, we have

$$\frac{(N + 1)}{2} = \frac{(10 + 1)}{2} = \frac{11}{2} = 5.5$$

I.e., we find that the median falls between the fifth and the sixth value in the order, or between 70 and 87.5. The halfway point between these two values is found by averaging them:

$$\frac{(70 + 87.5)}{2} = \frac{157.5}{2} = 78.75$$

which shows us that the median annual salary of the employees in that company is \$78,750. This is a lot less than the touted average of \$100,000 and a lot more reflective of what nine out of ten employees receive.

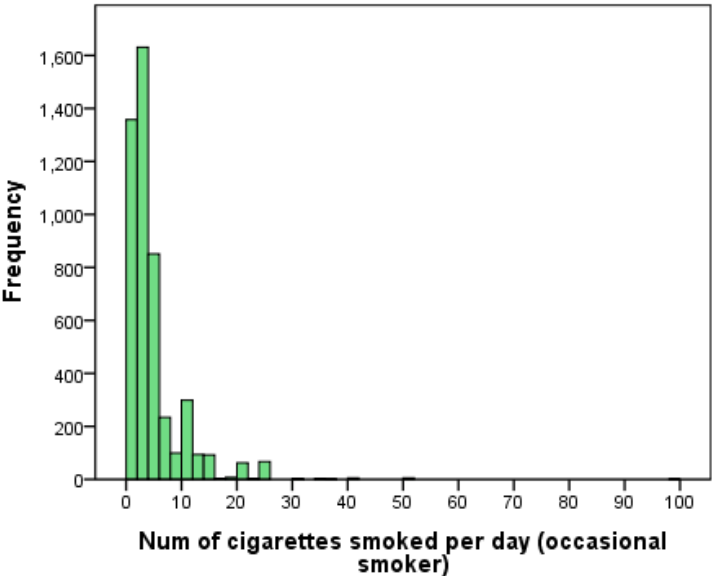
Examples like the *Watch Out!! #8* above show that relying on the mean can be tricky, and in some cases can be deliberately used to “lie with statistics” (i.e., a report might be technically correct but at the same time very misleading). Thus, **generally reporting all three central tendency measures is the way to go** and you, as a beginner researchers should do that.

Finally, you can observe a skew in the data even visually by looking at an interval/ratio variable's graphical representation, i.e., its histogram. Extremely high values tend to “pull” the mean to the right of the “centre”, i.e., with the majority of cases being relatively smaller, the few high values will produce a “tail” on the right side of the distribution (a.k.a. *positive skew*). On the other hand, extremely low values tend to “pull” the mean to the left of the “centre”, i.e., with the majority of cases being relatively larger, the few low values will produce a “tail” on the left side of the distribution (a.k.a. *negative skew*).

As well, since the median indicates the “centre” of the data better, a mean smaller than the median would typically indicate a negative/left skew, while a mean larger than the median would typically indicate a positive/right skew. When you observe a skew in the data, the median would typically be a the preferred measure of central tendency.

Observe the positive skew in Fig. 3.2 below.

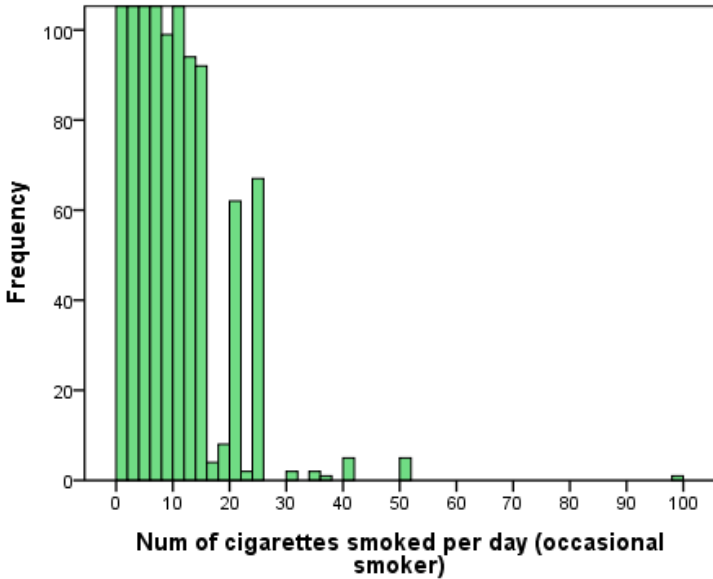
*Figure 3.1 Number of Cigarettes Smoked Per Day by Occasional Smokers (CCHS 15/16)*



The reason the numbers on the horizontal axis reach as high as 100 despite the fact that there appears to be nothing there is because there is at least one outlier case — a respondent who said they were an occasional smoker but reported smoking 99 cigarettes per day.<sup>4</sup> Thus the distribution has a long right-side “tail”, as it were, which you can better see in Fig. 3.2 providing the “zoomed-in” version of the histogram above. (The “tail” is what you will have if you trace an imaginary line through the tops of all the bars in the histogram down to the single case of 99 cigarettes per day.)

*Figure 3.2 Number of Cigarettes Smoked Per Day by Occasional Smokers (CCHS 15/16), Zoomed*

4. Whether this is to be believed is not important here, just the fact that such a value exists in the data. You will learn what is to be done about outliers in statistical analysis in Chapter 4.



In this case the median is 3 cigarettes smoked per day by an occasional smoker. The mean is 4.33, and as expected, it is larger than the median.

Similarly, an exceptionally small value compared to the bulk of the cases will produce a negatively-skewed histogram where the distribution has a “tail” but on the left of where most cases are. In that case the mean will be smaller than the median.





---

## 3.7 Central Tendency and the Levels of Measurement

This chapter introduced a lot of new concepts and terminology so a recap is in order. The three measures of central tendency — the mode, the median, and the mean — provide information about the so-called “centre of gravity” of a variable’s distribution, or where the cases tend to cluster. **The *mode* provides the most frequent category/value; the *median* provides the middle point/”centre” of the data and bisects the distribution into two equal part; and the *mean* is the mathematical average of values.**

One thing worth repeating is the caveat about the appropriateness of each of the measures of central tendency given the level of measurement of the variables at hand. Below is a quick, “cheat sheet” type of **a table summarizing which central tendency measures are appropriate for which levels of measurement.**

*Table 3.8 What Central Tendency Measures to Report for The Different Types of Variables*

	Nominal Scale	Ordinal Scale	Interval/ Ratio Scale
<b>Mode</b>	♦	♦	♦
<b>Median</b>	—	♦	♦
<b>Mean</b>	—	—	♦

In other words, the mode is appropriate for all variables, regardless of their level of measurement; the median works only with ordinal and interval/ratio variables; and the mean can be calculated only for interval/ratio variables.

I'll also restate it in terms of the variable type: **nominal variables have only a mode; ordinal variables a mode and a median; and interval/ratio variables have all three measures of central tendency.**

In terms of working with SPSS, as usual, it is *you* who makes the decision to request modes, medians, and means. You can either memorize the above Table 3.8, or, better yet, understand the logic behind each central tendency measure to know whether it's logically possible to apply it to a variable of a given scale — but in either case, SPSS will not make the decision for you.

**Watch Out!!** #9... for Trusting SPSS to Provide Only Appropriate Measures

SPSS cannot tell you the appropriate central tendency measures for a specific variable. Sometimes, if you make a mistake, depending on the mathematical procedure requested, SPSS might be genuinely unable to execute a command which will alert you to the fact that you have made an error. **However, in many cases SPSS will execute a command and will produce output, regardless of whether the command makes logical sense or not.**

To your bad luck, the measures of central tendency (and, as we will see in the next chapter, the measures of dispersion) are exactly one of these cases where SPSS will produce *any* measure of central tendency for *any* variable you ask of it. Thus, for example, if you request a mean for *race/ethnicity*, or a median for *religious affiliation*, it will execute the commands and give you what you asked for: it will produce numbers (which, if you remember, stand for the numerical labels of the categories). It will be then up to you to interpret those numbers.

This, however, would be a logical impossibility — there is no average *race/ethnicity*, nor “centre value” for *religious affiliation*. You would have made a mistake, and SPSS would have let you have your meaningless output.

This basically illustrates the saying “garbage in, garbage out”: if you input nonsense, the output will be nonsensical

too. It thus falls on you to not input nonsense and to not request measures of central tendency for variables for which they are inappropriate.

Results aside, proper communicating of findings is also very important. Even when output is produced correctly, your job is still not done: you still have to interpret the results and communicate what you have found. Considering that people in general (including in the social sciences) are variously trained in quantitative research, it is always a good idea to “translate” the more technical jargon into a more easily understandable, everyday language.

Specifically about descriptive statistics like the measures of central tendency we explored in this chapter, or the measures of dispersion in Chapter 4, the goal is to communicate your findings not only about *variables* and *measures* and *modes*, etc. but to explain what you have found in terms of *people* (or whatever units of analysis you happen to work with). Thus, “the mode of *religious affiliation* is...” becomes “the most frequently reported religious affiliation is...” or even “respondents most frequently identified as ... in terms of their religious affiliation”. (As well, getting into the habit of “translating” variable-centric jargon into people-centered statements is a good practice for your understanding of the material.)

Finally, a related issue is remembering to use the variable’s units of measurement when communicating results. To give a few examples, the median of *number*

*of siblings* is measured in “siblings”, the mean of *income* is measured in “dollars”, the mode of *age* is measured in “years”, etc. If you know the unit of measurement of the variable you describe (and you should), use it: a median age is never, say, 20; it’s 20 *years*.

With this done, we now turn to the last set of measures used to describe variables, namely measures of dispersion.



---

# Chapter 4 Measures of Dispersion

Early on in Chapter 3 we established that there are three pieces of information which helps us describe variables. Describing variables helps us to glean something from the variables' distribution beyond the raw list of observations of which it is made. In other words, through descriptive analysis we get to learn something about the cases that is not readily observable when all we have is a collection of data points.

Graphs provide a first glimpse at a variable's distribution. Measures of central tendency provide information about the typical cases, where most cases tend to cluster, or about the "centre" of the data. We now turn to measures of dispersion, the last of the three key pieces of descriptive information pertaining to variables. Measures of dispersion tell us how "spread out" a variable's cases are; they provide a "clusteredness" measure of the data, as it were, and of how *dispersed* cases are across the variable's values.

A simple illustration will make dispersion measures easier to understand. Take two sets of three numbers: "4, 5, 6" and "2, 5, 8". By now, you should be able to tell immediately that the median of both sets is 5 (each set has one value below and one above 5). You also might be able

to easily see that the mean of both sets is also 5; if not, this is how we get it:

$$\frac{\sum_{i=1}^N x_i}{N} = \frac{(4 + 5 + 6)}{3} = \frac{15}{3} = 5$$

$$\frac{\sum_{i=1}^N x_i}{N} = \frac{(2 + 5 + 8)}{3} = \frac{15}{3} = 5$$

Even if both “4, 5, 6” and “2, 5, 8” sets have the same measures of central tendency, you’d be hard-pressed to claim they are the same sets of numbers. Take a look at the image below (or just look at a ruler of your own, if you have one close by): the values of 4 and 6 are much closer to 5, than 2 and 8 are. That is, the values of our first set are more closely clustered around the “centre”, while the values of our second set are more loosely spread around it. This “clustering” vs. “spreading” is precisely what dispersion measures.



There are four commonly used measures of dispersion.<sup>1</sup>

1. A fifth measure of dispersion exists but is less commonly used. I'll introduce it only insofar as it is useful for understanding the standard deviation, the most widely used measure of dispersion.



Before we turn to each of them in turn, note what I have just demonstrated here: **it is quite possible for two variables to have the same measures of central tendency but different measures of dispersion.**

The four measures of dispersion can be divided into two groups. We begin with the simpler two, the *range* and the *interquartile range*, then turn to the more complicated (but most widely used) pair, the *variance* and the *standard deviation*.



---

## 4.1 Range

Providing the *range* for a set of values is so easy, most people don't even realize it is an actual statistical measure of dispersion. If you have ever said something to the effect of "I have friends whose ages vary between seventeen and twenty-seven" or "my scores on these exams vary from 25/100 to 95/100", etc., you have effectively been providing the range of your friends' ages or the range of your exam scores.

To give you the more technical definition, **the range of a variable is the difference between its highest and lowest values**. That is, to get the range, we simply subtract the lowest value from the highest value:

$$x_{max} - x_{min} = range$$

In the two quick examples above, the range of your friends' ages would be  $(27-17=)$  10 years, and the range of your exam scores would be  $(95-25=)$  70 points.

I'll use an older, familiar example for the longer work-through, below.

*Example 4.1 The Range for Textbook Prices Paid in One Semester*

Recall Example 3.5 from Section 3.4 (<https://pressbooks.bccampus.ca/simplestats/chapter/3-4-mean/>) where we calculated the mean price of textbooks we imagined you paid in a particular semester. The books' prices were \$120, \$230, \$300, \$65, \$30. The cheapest book (i.e., the lowest value,  $x_{min}$ ) was \$30 and the most expensive book (i.e., the highest value,  $x_{max}$ ) was \$300. Thus

$$x_{max} - x_{min} = 300 - 30 = 270 = range$$

That is, now we have found that the range of textbook prices for that semester was \$270, with prices you paid ranging between \$30 and \$300.

One thing to note here is that in order to have a difference, i.e., in order to be able to do a mathematical operation like subtraction, we need to have numerical values.

In truth, as you are about to see, *all* measures of dispersion are obtained through mathematical operations and, as such, require numerical values. Since interval/ratio variables are the only variables which contain actual numerical values, **all dispersion measures (including the range) are only applicable to interval/ratio variables.**<sup>1</sup>

1. Some people find it useful to provide *something like a range* for ordinal variables: after all, they do have a "lowest" category and a "highest" category. While technically not a statistical measure of dispersion (as no difference can be computed), it can still be useful to add a description about

A final point about the range is that it is a rather unsophisticated measure of dispersion, as you have already noticed. (Hence the very short section about it.) **By taking into account solely the highest and the lowest values, the range effectively ignores all other values**, be they more clustered or more spread out.

After all, if you recall from Section 3.6 (<https://pressbooks.bccampus.ca/simplestats/chapter/3-6-outliers/>), outliers do exist. In the presence of outliers, the range can end up being quite large, even if the majority of the observations are closely clustered. Therefore, we'd better find a dispersion measure which takes into account more than just the two extremes of a variable's distribution.

The *interquartile range* is one such measure which provides a bit more information about the variability of the distribution. Alas, the cost of this information is, of course, an increased complexity in obtaining that measure. (An ominous foreshadowing for what's to come!)

the categories ranging between the lowest and highest points, e.g., "respondents' agreement with the statement varies between "strongly disagree" and "strongly agree". Considering that the categories of nominal variables have no inherent order, nothing of the sort can be applied to them. All in all, providing a qualitative description of dispersion for ordinal variables (like the agreement one I just mentioned) is optional and, strictly speaking, not a statistical measure.



---

## 4.2 Interquartile Range

Unlike the range which focuses on the extreme ends, the **interquartile range** (frequently referred to as ***IQR***) looks into the distribution of observations around the “centre”. To that purpose, it splits the distribution into **four equal parts called *quartiles*** (from the Latin *quartus*, meaning one-fourth, i.e., a quarter), and then provides the range of the middle two parts taken together. This sounds more complicated than it actually is, so let’s turn to examples and make it better.

To begin, let me first demonstrate what all this means with a set of raw values which we can call, say, *hours worked per week*.

### *Example 4.2 Weekly Hours Worked (Raw Data)*

Imagine you have been hired as a research assistant (RA) on a research project. You have worked 20 weeks in total in the past two semesters, ten weeks in each semester (with your classes and all, you couldn’t work every week). The maximum hours per week you could work was 15, limited by the nature of your contract. You make a list of all hours

you have worked in each of the twenty weeks, and you list the twenty values *in ascending order*. Here they are:

2, 3, 3, 4, 5, 7, 7, 7, 8, 8, 10, 10, 10, 10, 12, 12, 13, 13, 13, 14

If you recall from our discussion of the median, to split a group of values into equal parts we need the values' positions in the order. You can find these in the table below:

*Table 4.1 Values and Their Positions of Hours Worked per Week*

Position	Hours Worked per week	Position	Hours Worked per Week
(1)	2	(11)	10
(2)	3	(12)	10
(3)	3	(13)	10
(4)	4	(14)	10
(5)	5	(15)	12
(6)	5	(16)	12
(7)	7	(17)	13
(8)	7	(18)	13
(9)	8	(19)	13
(10)	8	(20)	14

You might be tempted to use an intuitive method for splitting the set of twenty values given in the example into 4



equal parts (i.e., into quartiles) by simply dividing 20 by 4, which will let you have 5 values in each quartile:

2, 3, 3, 4, 5      5, 7, 7, 8, 8      10, 10, 10, 10, 12  
12, 13, 13, 13, 14,

Thus the interquartile range (or “the range of the middle two parts taken together”) of the entire set of 20 values would be the range of 5, 7, 7, 8, 8, 10, 10, 10, 10, 12.

A quick-and-dirty calculation would show that the IQR is  $(12-5=)$  7 hours. You would be correct — indeed, the interquartile range is 7 hours — but I’ll stop you nevertheless. This worked out only because I’ve chosen the numbers between the first and the second quarter of cases to be both 5, and the numbers between the third quarter and the last to be both 12. You need to read below to find out the proper method for obtaining the IQR. (The example continues further down.)

Quick-and-dirty calculations are not precise, even if they serve their purpose to give you a basic idea of what we are doing. Now that you’ve seen where this is going, let’s do everything *properly*.

First, we need to calculate the precise positions of the values that separate the quartiles. Recall how we used to split a set of values in two in order to get the position median. We used the following formula:

$$\frac{N+1}{2} = \leftarrow \text{“position of the median”}$$

We'll follow the same logic to split each of the halves in two themselves. Thus let me restate the above formula to this:

$$\frac{N+1}{2} = (N + 1)\frac{1}{2} = (N + 1)0.5 \quad \leftarrow \text{"position of the median"}$$

Since we effectively multiply  $N+1$  by 0.5 in order to split the entire set in two halves (or, to get *one half of the data*), to split the first half of the values further in two itself, we need to multiply  $N+1$  by "half of 0.5", i.e., by 0.25 (essentially getting *one quarter* of the data):

$$\frac{N+1}{4} = (N + 1)\frac{1}{4} = (N + 1)0.25 \quad \leftarrow \text{"position of the first quartile"}$$

By analogy, splitting the second half in two itself will require getting *three quarters* of the data, or to multiply  $N+1$  by "0.5 and a quarter", i.e., by 0.75:

$$\frac{(N+1)3}{4} = (N + 1)\frac{3}{4} = (N + 1)0.75 \quad \leftarrow \text{"position of the third quartile"}$$

If you follow the logic, you'll easily conclude that **the median is also *de facto* the second quartile** (i.e., *two quarters* of the data).

To restate, we have the following way to split the data into four equal parts:

The position of the first quartile,  $Q_1$ , is found through  $(N + 1)0.25$ .

The position of the second quartile,  $Q_2$  (a.k.a the median), is found through  $(N + 1)0.5$ .

The position of the third quartile,  $Q_3$ , is found through  $(N + 1)0.75$ .<sup>1</sup>

Now let's use our newfound formulas in the Example 4.2.

*Example 4.2 Weekly Hours Worked, Continued*

With  $N=20$ , we get:

$$Q_1\text{'s} \quad (N + 1)0.25 = \overset{\text{position}}{(20 + 1)0.25} = (21)0.25 = 5.25$$

$$Q_2\text{'s} \quad (N + 1)0.5 = \overset{\text{position}}{(20 + 1)0.5} = (21)0.5 = 10.5$$

$$Q_3\text{'s} \quad (N + 1)0.75 = \overset{\text{position}}{(20 + 1)0.75} = (21)0.75 = 15.75$$

Once again, do not forget that all these formulas provide the *positions* of the quartiles, *not* their respective values. To see the values, we have to look at Table 4.1 above which cross-lists the cases' positions *and* values. Since there is no

1. Obviously, we don't speak of a *fourth quartile*, as four quarters comprise the whole thing: the fourth quartile would simply be 100%, or all of the data.

Case #5.25, we know that the value we're looking for is between Cases #5 and #6 (a quarter further than #5) — but as the values of both Cases #5 and #6 are 5, we conclude that the value of the first quartile is 5.

Similarly, there is no Case #15.75 (so the value we're looking for is three quarters past the 15th case), but both Cases #15 and #16 are 12, so we conclude that the third quartile is 12.

We are still interested in the interquartile range — or the range of the two middle quarters of the data (or the middle 50 percent, so to speak). Then, since

$$Q_3 = 12 \text{ and } Q_1 = 5,$$

we have that

$$Q_3 - Q_1 = 12 - 5 = 7$$

Or, we have found that the IQR for *hours worked per week* is 7 hours per week. Or, at the mid-range, your hours worked per week varied between 5 and 12 hours per week.

Alright, but *why*, you might ask — couldn't we just have the range and be done with it?

The value added of using interquartile range is that it

takes care of outliers, so it's frequently a better measure of dispersion than range. The IQR provides the spread of the centrally located 50 percent of the data which in many situations paints a more accurate picture of how "the more typical" of the variable's cases are spread out, rather than looking at the more extreme spread provided by the range which encompasses all cases, even the clear outliers.

All in all, however, just like with choosing whether to use a median or mean, the decision which of these two measures of dispersion is the more appropriate one to be used and reported depends on the specific situation and the researcher's discretion. I would urge you, as a beginner researcher, to make a habit of reporting both the range and the interquartile range, while simultaneously discussing the effect of any potential outliers.

Instead of working with raw data, we might have frequency tables at hand. **How do we get the range and IQR from aggregated data?** For the range, simply subtract the lowest value (the one listed first in the *Values* column, of course) from the highest value (the one listed last in the *Values* column) and report the difference (in its appropriate units of measurement). For the IQR, look for the 75th percentile (i.e.,  $Q_3$ ) and the 25th percentile (i.e.,  $Q_1$ ) in the *Cumulative Percent* column, then subtract the  $Q_1$  value from the  $Q_3$  value, and again report the difference. (This is similar to how we looked for the 50th percentile for the median,  $Q_2$ , in Section 3.3 (<https://pressbooks.bccampus.ca/simplestats/chapter/3-3-the-median-with-frequency-tables/>).)

*Exercise 4.1 Range and IQR for Cigarettes Smoked per Day*

Practice your newly acquired skills to find  $Q_1$ ,  $Q_2$  (i.e., the median), and  $Q_3$  in the following table. Calculate and report the range and the interquartile range for *number of cigarettes smoked each day*.

*Table 4.2 Number of Cigarettes Smoked Per Day by Daily Smokers (CCHS 15/16)*

**Num of cigarettes smoked each day (daily smoker)**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	110	.1	.7	.7
	2	219	.2	1.4	2.2
	3	328	.3	2.1	4.3
	4	405	.4	2.7	7.0
	5	777	.7	5.1	12.0
	6	617	.6	4.0	16.1
	7	461	.4	3.0	19.1
	8	572	.5	3.7	22.9
	9	153	.1	1.0	23.9
	10	2297	2.1	15.0	38.9
	11	127	.1	.8	39.7
	12	1397	1.3	9.2	48.9
	13	390	.4	2.6	51.4
	14	95	.1	.6	52.1
	15	1637	1.5	10.7	62.8
	16	74	.1	.5	63.3
	17	121	.1	.8	64.1
	18	181	.2	1.2	65.3
	19	17	.0	.1	65.4
	20	2164	2.0	14.2	79.5
	21	8	.0	.1	79.6
	22	37	.0	.2	79.8
	23	34	.0	.2	80.1
	24	57	.1	.4	80.4
	25	2225	2.0	14.6	95.0
	26	3	.0	.0	95.0
	27	8	.0	.1	95.1
	28	13	.0	.1	95.2
	29	1	.0	.0	95.2
	30	278	.3	1.8	97.0
	31	1	.0	.0	97.0
	32	8	.0	.1	97.1
	33	5	.0	.0	97.1
	34	1	.0	.0	97.1
	35	89	.1	.6	97.7
	36	4	.0	.0	97.7
	37	24	.0	.2	97.9
	38	10	.0	.1	97.9
	40	135	.1	.9	98.8
	45	13	.0	.1	98.9

To make sure you're doing it correctly, let's quickly check your answers right away. The range is of course  $(99-1=)$  98 cigarettes per day. To find the IQR, you must have first identified  $Q_1= 10$  (since 23.9 percent of the cases make up to 9 cigarettes per day, the 25th percentile falls in the 10 cigarettes per day category) and  $Q_3 = 20$  (since 65.4 percent of the cases make up to 19 cigarettes per day, the 75th percentile falls in the 20 cigarettes per day category). Then the IQR is  $(20-10=)$  10. Thus you see the difference between range and interquartile range: while the range might leave you with the impression that cigarettes smoked per day vary by almost a hundred for daily smokers, the middle half of the cases actually only vary by 10 cigarettes.

Of course, there's also SPSS. Check below to see how to find the range and IQR (semi-) directly.

#### *SPSS Tip 4.1 Obtaining Range and Interquartile Range*

- From the *Main Menu*, select *Analyze*, then *Descriptive Statistics*, and then *Frequencies*;
- Select your variable of choice from the list on the left and use the arrow to move it to the right side of the window;
- Click on the *Statistics* button on the right;



- In this new window, check *Quartiles* from the *Percentile Values* on your top left and check *Range* (and *Minimum* and *Maximum* if you wish) from the *Dispersion* section below it;
- Click *Continue*, then *OK*.
- Range (along with the smallest and largest values, if you asked for them) will be reported in the *Output* directly.
- To obtain the IQR, simply subtract the value reported as 25th percentile from the value reported as 75th percentile.

With the range and IQR covered, we are halfway through the typically used measures of dispersion. On to the remaining two, the variance and the standard deviation.



---

## 4.3 Variance

Similarly to how the median is about the central position of a case while the mean is about the average of actual numerical values, the range and interquartile range are about positions in the overall (ordered) distribution of cases while the remaining two dispersion measures, the variance and the standard deviation, are about averaging numerical values.

Thus, like the mean, the variance and the standard deviation account for *all* cases, not just a select few. Unlike the mean, however, instead of calculating the *average of all values*, **the standard deviation and variance calculate (approximately) the average of the distances of each and every value to the mean.**

The mean is a measure of central tendency, as you know by now, and it represent a sort of “centre” of the data, *value*-wise (as opposed to *position*-wise, which is what the median is). You know that all cases’ values enter the calculation of the mean (after all, we sum all values and divide the sum on their total number to get the mean), but, at the same time, the values are *different* from the mean. (That is, either all are different, or all but one — it’s possible that one of the values is actually what the mean is, in which case the difference is zero.)

This difference, between a value of a case and the mean, is what we call *distance to the mean*.

We have to average these (by adding all of the distances of all cases's values together and dividing by their total number) to obtain the variance and the standard deviation. Once we have these dispersion measures, we'll be able to tell how *all* cases are spread out around the mean. This, in turn, gives us information about how much *variability* there is in a given variable's cases, if they are dispersed or clustered together.

You'll be glad to know that the variance and the standard deviation are calculated in almost the exact same way; the standard deviation needs just one additional mathematical operation after getting the variance. In a sense, they calculate the same thing but are expressed differently, and the standard deviation is usually considered easier to interpret.

This is all the good news I have for you at this point, I'm afraid, as what follows is a calculation process containing several steps. On the whole, it may look complicated though it really isn't; the key is to not forget what you are doing and where you are in the process. If you find yourself losing track, simply go back and start from the beginning, paying attention to what steps you go through.

**Variance.** Since we want an average of the distances of the cases from the mean, it would make sense to start with getting these distances as a Step 1. Step 2 would be to add these distances together, then Step 3 would be to divide the sum on their total number. This is easier said than done, as you shall see (ominous foreshadowing!), so I'll divide Step 1 into two sub-steps, Step 1A (getting the distances) and Step 2B (a procedure I'll keep as a mystery for now).

As usual, we'll do all this through an example. For simplicity's sake, I'll reuse Examples 4.2/4.3 from the previous section which we used to introduce the concept of IQR.

#### Example 4.4 (A) *Weekly Hours Worked*, Revisited

If you recall, we had imagined you as a research assistant (RA) on a research project and you had worked 20 weeks in total in the last two semesters, ten weeks in each semester. The maximum hours per week you could work was 15, limited by the nature of your contract.

As there are a lot of calculations to be done, to simplify our job, let's imagine further that we're interested in only one of the two semesters you had worked, and these are only the hours in the *ten* weeks of that one semester:

3, 3, 5, 7, 8, 10, 12, 12, 13, 14

Considering that **for Step 1A we need the distances of each of these ten values to the mean**, we'll calculate the mean as a preliminary requirement.<sup>1</sup>

1. Since  $N=10$  or more makes for quite the long equations if the values are listed (summed) one by one separately, from now on I will group values by frequencies in the calculations I do as a matter of principle. (I.e., instead of  $3+3$ , here I have  $(3)2$ ,

$$\begin{aligned} \frac{\sum_{i=1}^N x_i}{N} &= \\ &= \frac{(3)2 + 5 + 7 + 8 + 10 + (12)2 + 13 + 14}{10} = \\ &= \frac{(6 + 5 + 7 + 8 + 10 + 24 + 13 + 14)}{10} = \frac{87}{10} = 8.7 = \bar{x} \end{aligned}$$

Armed with the mean of 8.7 hours, we can now proceed to calculate the distance of every value to the mean (i.e., subtract the mean from each value to obtain the difference). I list the values and their respective distances from the mean in the table below.

*Table 4.3 Step 1A Calculating Distances To the Mean*

instead of 7+7+7, I would have (7)3, etc.) Coincidentally, this is exactly what we do when working with data organized in a frequency table.

$x_i$	$(x_i - \bar{x})$
3	$(3 - 8.7) = -5.7$
3	$(3 - 8.7) = -5.7$
5	$(5 - 8.7) = -3.7$
7	$(7 - 8.7) = -1.7$
8	$(8 - 8.7) = -0.7$
10	$(10 - 8.7) = 1.3$
12	$(12 - 8.7) = 3.3$
12	$(12 - 8.7) = 3.3$
13	$(13 - 8.7) = 4.3$
14	$(14 - 8.7) = 5.3$

Again, as usual,  $x_i$  is the value of each and any Case  $\#i$  (from 1 to 10), and  $(x_i - \bar{x})$  is the distance (i.e., difference) between each and any Case  $\#i$  (from 1 to 10) to the mean.

Now if we were to jump directly to Step 2 (summing all distances together) and Step 3 (dividing by the total number), we would be in trouble. You see, since the mean averages all values and provides a “centre” of the variable’s distribution value-wise, **distances of the values below the mean equal the distances of the values above the mean, albeit with opposite signs.**

That is, summing all values *below* the mean (i.e., the

negative differences) would equal the sum of all values *above* the mean (i.e., the positive differences). **As one sum is negative and the other positive (but with the same absolute value<sup>2</sup>), they cancel each other out — adding them together would result in 0, every time.** This is due to the very nature of the calculation of the mean; it's a mathematical inevitability.

Don't believe me? Try it. The sum of the distances *below* the mean is:

$$(-5.7) + (-5.7) + (-3.7) + (-1.7) + (-0.7) = -17.5$$

The sum of the distances *above* the mean is:

$$1.3 + 3.3 + 3.3 + 4.3 + 5.3 = 17.5$$

Thus, the sum of *all* distances from the mean is

$$(-5.7) + (-5.7) + (-3.7) + (-1.7) + (-0.7) + 1.3 + 3.3 + 3.3 + 4.3 + 5.3 = -17.5 + 17.5 = 0$$

Told you: *Zero. Every. Time.*<sup>3</sup>

2. The absolute value of a positive number is the number itself; the absolute value of a negative number is the number itself but without the negative sign; the absolute value of zero is zero. Absolute value is noted with two straight vertical line. For example, the absolute values of -1 and 1 are equal to each other:  

$$|-1| = |1| = 1.$$
3. If you're still not convinced and think that maybe I selected the numbers *just* so that the distances to their mean add up to zero on purpose, you are welcome to try this 'trick' with any set of numbers.



So if the sum of the distances to the mean is always zero, then what? How are we to average those distances, since dividing the sum (i.e., zero) on any  $N$  would give us zero? Are we to give up?

The thing is, the distances (below and above the mean) only cancel each other out because we consider the distances below the mean as *negative*. This, however, is a somewhat of a mathematical conceptual artifact: in real life, there is no such thing as a negative distance from one thing to another. Imagine yourself standing between two of your friends, one on your left and the other on your right. Let's assume they both stand a meter away from you: you wouldn't say that one is a negative meter away while the other is a positive meter away, would you? There are no negative and positive meters, just meters (and well, they are always positive, as distance in the physical sense always is).

Thus we are actually not interested in summing the cases' distances from the mean *as calculated* but only in their "positive version" ignoring their signs, i.e., we want their *absolute values*.

True, we could proceed with our Steps 1 and 2 using only positive distances. When done, this produces an actual dispersion measure called *mean deviation* (or *mean absolute deviation*). The mean deviation is easy to understand and quite intuitive, however (and perhaps to your chagrin), it is rarely used — specifically because we have the variance and standard deviation which are found to be much more useful (this comes into play in inferential statistics, as you will see in the latter part of this book). Due

to its unpopularity, I'll therefore skip the mean deviation — we'll have to look for another way of getting only positive numbers for our calculation of the average distance from the mean.<sup>4</sup>

Now stop and think: beside absolute values, is there another way of turning numbers positive?

If you thought of squaring, good for you! A (non-zero) number squared is a positive number:  $(-2)^2 = 2^2 = 4$ . Thus one other way of getting around our distances-summing-to-zero problem is to *square* the distances *before* adding them up! Nifty trick, eh?

Let's test how this works with our Example 4.4.

*Example 4.4 (B) Weekly Hours Worked, Revisited*

4. For the curious souls out there (all three of them), this is what the mean deviation looks like, using the numbers from Example 4.4 (A) above. As the below-the-mean sum was -17.5 and the above-the-mean sum was 17.5, ignoring the negative signs we would get
- $$5.7 + 5.7 + 3.7 + 1.7 + 0.7 + 1.3 + 3.3 + 3.3 + 4.3 + 5.3 = 17.5 + 17.5 = 35$$
- . Since  $N=10$ , by averaging the distances we get  $\frac{35}{10} = 3.5$  (the mean absolute deviation). That is, the average distance of a case's value from the mean is 3.5, or, in terms of our example, your weekly hours (which ranged from 3 to 14) on average varied by 3.5 hours from the mean of 8.7 hours, across the ten weeks you worked as a research assistant.

A reminder: what we are trying to get is a dispersion measure giving us an average distance of the cases to the mean; something to account for the variability of *all* cases, not just a few (unlike the range and IQR). To make the calculations look more orderly, I add a third column to Table 4.3 above, one with the squared distances. Thus, our mysterious **Step 1B is squaring each individual distance**.

Table 4.4 Step 1B Squaring Individual Distances

$x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
3	$(3 - 8.7) = -5.7$	$(-5.7)^2 = 32.5$
3	$(3 - 8.7) = -5.7$	$(-5.7)^2 = 32.5$
5	$(5 - 8.7) = -3.7$	$(-3.7)^2 = 13.7$
7	$(7 - 8.7) = -1.7$	$(-1.7)^2 = 2.9$
8	$(8 - 8.7) = -0.7$	$(-0.7)^2 = 0.5$
10	$(10 - 8.7) = 1.3$	$(1.3)^2 = 1.7$
12	$(12 - 8.7) = 3.3$	$(3.3)^2 = 10.9$
12	$(12 - 8.7) = 3.3$	$(3.3)^2 = 10.9$
13	$(13 - 8.7) = 4.3$	$(4.3)^2 = 18.5$
14	$(14 - 8.7) = 5.3$	$(5.3)^2 = 28.1$

We are thus ready for **Step 2: summing up the (now-squared) distances from the mean:**

$$\sum_{i=1}^N (x_i - \bar{x})^2 = (32.5)^2 + 13.7 + 2.9 + 0.5 + 1.7 + (10.9)^2 + 18.5 + 28.1 = 152.1$$

← *Sum of Squares*

As you can see above, **the sum of the squared distances from the mean is called the *sum of squares*** (sometimes indicated by *SS*).

Finally, to get the average distance from the mean we need **Step 3: to divide the sum of squares by the total number, *N***:

$$\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N} = \frac{152.1}{10} = 15.21 = \sigma^2 =$$

← *variance*

That is, the variance of your hours worked per week is 15.21, or the average of the squared distances from the mean is 15.21. (Note that we cannot say 15.21 *hours* as now we are working in squared units.)

And this is it, the *variance*. It is denoted by a small-case Greek letter  $\sigma$ , i.e.  $\sigma^2$  (SIG-ma-squared).<sup>6</sup> variance

5. It is pronounced SIG-ma, just like  $\Sigma$  which is the capital-case Greek letter *S*. and, since it's in squared units, actually  $\sigma$
6. An alternative notation for

you might encounter is **var( $x$ )** where  $x$  is the variable in question.



---

## 4.4 Variance Continued, Standard Deviation

I'm sure you'll agree the preceding section was a lot to take in. And here's the kicker: after all that, we arrived at something which we cannot easily or intuitively interpret, given the squared units. However, the variance is used a lot in statistics, for great many things. Generally, the larger the variance, the greater the *variability* of the variable, or the larger the "dispersed-ness" of the cases.

Despite the seemingly convoluted way we arrived at the variance and all the calculations and mathematical notation, what we did was actually quite simple. (No, really!)

To recap: just like we average all values by summing them up and dividing the sum on their total to get the mean, we average the distances of the values from the mean by summing them up and dividing the sum on their total. The only difference is that in order to be able to sum the distances, we need to square each of them first, or we cannot proceed.

Here are the formulas for the mean and the variance together so that you can compare:

$$\frac{\sum_{i=1}^N x_i}{N} = \bar{x} \leftarrow \text{mean}$$

$$\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N} = \sigma^2 \quad \leftarrow \text{variance}$$

Now that I have you feeling somewhat comfortable, I have a confession to make. **This above isn't the only version of the formula for variance that exists or that we will be using.**

Bear with me (and welcome back, to those who threw the reading away in disgust) — I promise to explain everything when we get to inferential statistics further in the textbook, as the explanation requires concepts and terminology we have not yet covered and which cannot be easily introduced at this point. (Hint: it deals with estimation and uncertainty.)<sup>1</sup>

$$\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1} = s^2 = \quad \leftarrow \text{variance}$$

As you can see, the modification is quite small -- **instead of dividing the sum of squares by the total number  $N$ , we actually divide it by the total *minus one*,  $N-1$ .** If it makes you feel better, dividing just by  $N$  or by  $N-1$  produces generally similar results, in terms of magnitude of the variance. We also denote this version with a regular small-case  $s^2$ .

One thing worth noting, however, is that despite the lack of proper explanation as of yet, when working with typical datasets **SPSS will produce variances by dividing the sum of squares by  $N-1$  instead of by  $N$ .**

1. If you'd like a preview, **the alternative, to-be-explained-later, formula for variance is:**



**Watch Out!! #9 ... for The Order of Operations**

When considering the formula for variance, and the steps we took to calculate it, pay special attention to the *sum of squares*. That is, we need a sum of *squares* (a.k.a., to add the squared distances from the mean together): **we *first* calculate the distances, *then* square them, and finally sum the *squared* distances up.**

A common mistake, however, is to try to calculate the distances, sum them up, *then* square the sum. As explained above, the (un-squared) distances add up to zero, and squaring the zero will not improve things. A version of this mistake is also to calculate the distances, then try to sum them and divide them by  $N-1$ , and *then* square the result. Obviously this would also be unsuccessful. To avoid these type of frustrations, try to remember the purpose of the squaring: to “turn” all distances into positive numbers. Everything else we do (summing, dividing), we do to the already squared distances.

In an effort to show you that the calculation of the variance is simple when done without the protracted explanations, take another example we have used before, *number of siblings*.

### Example 4.5 Variance for Number of Siblings

In discussing the median in Section 3.2 (<https://pressbooks.bccampus.ca/simplestats/chapter/3-2-median/>), we imagined you asked seven of your friends about the number of their siblings. These were the values we used: 2, 1, 4, 2, 1, 0, 3.

Let's produce the variance, in four simple steps, after calculating the mean; Step 1A, obtain the distances from the mean; Step 1B, square the distances from the mean; Step 2, obtain the sum of squares (i.e., sum the distances up); Step 3, divide by  $N$ .

**Preliminary step: obtain the mean.**

$$\frac{\sum_{i=1}^N x_i}{N} = \frac{2+1+4+2+1+0+3}{7} = \frac{13}{7} = 1.857 = \bar{x}$$

**Steps 1A and 1B are presented in the table below:**

*Table 4.4 Calculating Distances To the Mean and Squaring Each Distance*

$x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2	$(2 - 1.857) = 0.143$	$(0.143)^2 = 0.02$
1	$(1 - 1.857) = -0.857$	$(-0.857)^2 = 0.734$
4	$(4 - 1.857) = 2.143$	$(2.143)^2 = 4.592$
2	$(2 - 1.857) = 0.143$	$(0.143)^2 = 0.02$
1	$(1 - 1.857) = -0.857$	$(-0.857)^2 = 0.734$
0	$(0 - 1.857) = -1.857$	$(-1.857)^2 = 3.448$
3	$(3 - 1.86) = 1.143$	$(1.143)^2 = 1.306$

**Step 2, obtain the sum of squares:**

$$\sum_{i=1}^N (x_i - \bar{x})^2 = (0.02)2 + (0.734)2 + 4.592 + 3.448 + 1.306 = 10.854$$

← *Sum of Squares*

**Step 3, divide the sum of squares** (rounded down to two digits) **by  $N$** , i.e., by 7:

$$\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N} = \frac{10.85}{7} = 1.55 = \sigma^2 \quad \leftarrow \text{variance}$$

Thus, we find that your seven friends have an average of about 1.6 squared distances from the mean number of siblings 1.9 (rounded up from 1.857).

*Oh, great*, you are probably thinking now, and I can imagine the sarcasm — *we calculated something we can't even interpret properly*. I mean, it's more than a tad awkward to try to explain “an average of about 1.6 squared distances from the mean number of siblings” to anyone not versed in statistics. Maybe it would be better if we could get rid of the “squared-ness”?

You know what? *We can*. The standard deviation is here to help.

**Standard deviation.** Believe it or not, after all the steps we went through to get to the variance, calculating the standard deviation is a breeze: specifically, a breeze that turns back the squared units into *standard* units, hence the name.

See for yourself:

$$\sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}} = \sqrt{\sigma^2} = \sigma \quad \leftarrow \text{standard deviation}$$

Despite its scary looks, this is actually just the formula for variance *under a square root*. That is, **we take the square root of the variance to get the standard deviation**. That's it. Nothing more. Just a regular square root, and we're there. Cue in a sigh of relief!<sup>2</sup>

2. Note, however, that just like there is an "alternative", to-be-explained-later, formula for variance, there is an "alternative" formula for standard deviation, following the same principle regarding dividing the sum of squares by  $N-1$  instead of by  $N$ :

$$\sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}} = \sqrt{s^2} = s \leftarrow \textit{standard deviation}$$

As well, SPSS will use this  $(N-1)$  version of the formula when working with variables in a dataset.

Now that we know how to get back to standard units, let's do that for the two examples we used. We had a variance of  $\sigma^2 = 15.21$  for *hours worked per week* in the previous section and a variance of  $\sigma^2 = 1.6$  for *numbers of siblings* in the example above. Square-rooting gives us the following:

$$\sqrt{\sigma^2} = \sqrt{15.21} = 3.9$$

and

$$\sqrt{\sigma^2} = \sqrt{1.6} = 1.25$$

Now *these we can* interpret: on average, your hours worked per week deviated from the mean of 8.7 hours per week by 3.9 *hours*, and your friends deviated from the average number of siblings, 1.9, by 1.25 *siblings*.

To repeat, **the standard deviation is the square root of the variance. The standard deviation is a measure of dispersion which gives us the average deviation of the cases from the mean.** (Technically, an average of the squared distances from the mean in standard units.)

### *Do It! 4.2 Longevity of The First Fifteen Canadian Prime Ministers*

Calculate the variance and standard deviation of the longevity of the first fifteen Prime Ministers of Canada. In chronological order (starting with Macdonald and ending with Pierre Trudeau), their ages at the time of death were: 76, 70, 72, 49, 93, 94, 77, 82, 86, 75, 76, 91, 83, 75, and 80. Interpret your results (i.e., explain what you have found beyond “the standard deviation is ...”).

You can use a table like Table 4.4 to organize your calculations. (Hint: Start with calculating the mean age at death,  $\bar{x}$ , and round it up to a whole number to make your job easier.) Here  $x_i$  is age at death for each PM and  $N=15$ .

You can check your answers in this footnote.<sup>3</sup> The mean is 79 years; the sum of squares 1,717; the variance 114.5; the standard deviation 10.7 years. However, if you calculated the variance and standard deviation with  $N-1$  in the denominators, you will get a variance of 123 and a standard deviation of 11.1 years. The difference is as large as it is due to the small  $N$ . Had we been working with a real dataset of hundreds or thousands of cases, the difference between the just- $N$  and  $N-1$  versions of the formulas would have been less pronounced.

Of course, one wouldn't normally calculate variances and standard deviations by hand: we only do it so that you can understand what the measures are and what they really

provide us with, by obtaining them ourselves. Usually, however, we simply use SPSS.

*SPSS Tip 4.2 Obtaining Variance and Standard Deviation*

- From the *Main Menu*, select *Analyze*, then *Descriptive Statistics*, and then *Frequencies*;
- Select your variable of choice from the list on the left and use the arrow to move it to the right side of the window;
- Click on the *Statistics* button on the right;
- In this new window, check *Variance* and *Standard deviation* in the *Dispersion* section on the left at the bottom;
- Click *Continue*, then *OK*.
- The *Output* window will provide a table with the requested measures.
- Make sure you know how to interpret your results! (Try to use as little statistics jargon as possible.)





---

## 4.5 Summary

It sure feels like we've covered a lot! You might need a recap. You will find it below.

The measures of dispersion tell us how a variable's cases are distributed: whether they are more tightly clustered together, or more loosely spread out. After all, it's perfectly possible to have two variables with the same central tendency measures but with different measures of dispersion!

There are four measures of dispersion that are typically used: range, interquartile range (IQR), variance, and standard deviation. While the former two are simple and account for the dispersion of cases only through the positioning of a few cases in the (ordered) distribution, the latter two employ *all* cases's values to produce somewhat more complicated and comprehensive measures of a variable's spread.

The range reports the difference between the highest and the lowest values. The IQR provides the same but for the middle half of the cases. The variance calculates *something like* an average of the squared distances of all cases from the mean (in squared terms), while the standard deviation, through square-rooting the variance, provides us with an almost-average of the distances of all cases from the mean (in standard — i.e., *regular* — units). Generally,

the larger the measures of dispersion, the more *variability* the variable has.

Finally, as they all require numerical values, all measures of dispersion are applicable only to interval/ratio variables: we cannot provide dispersion measures for nominal or ordinal variables.

With this, we have the full range of measures to describe variables: we not only learned how to graph variables to see their distribution visually, but also to calculate how their cases cluster (through the three measures of central tendency, the mode, the median, and the mean) and how the cases can spread (through the four measures of dispersion, the range, the interquartile range, the variance, and the standard deviation).

We also learned that while we can graph all types of variables, the measures of central tendency and dispersion vary in their applicability depending on a variable's level of measurement. **While the mode applies to all variables, and the median to ordinal and interval/ratio variables, the mean, the range, the IQR, the variance, and the standard deviation apply *only* to interval/ratio variables.** Keep this in mind when deciding what kind of information to provide about a specific variable.<sup>1</sup>

Before we continue inching toward inferential statistics, starting with the normal curve and basic of probability in Chapter 5, here is a handy list of things you should know before proceeding further.

1. Again, do not trust SPSS to make that decision for you: it cannot and it will not.

### What You Need To Know So Far

- How to visually display a variable's distribution (i.e., how to graph variables) and the proper graph for each variable type depending on level of measurement;
- How to display a variable's distribution in a tabular format, specifically how to create and how to read frequency tables;
- What the central tendency measures are, how many and what they are, their applicability to variable types depending on level of measurement, and what methods there are to obtain them (including calculation);
- What the central dispersion measures are, how many and what they are, their applicability to variable types depending on level of measurement, and what methods there are to obtain them (including calculation);
- What outliers are and how they affect the central tendency and dispersion measures, and what makes a more appropriate measure of central tendency or dispersion in the presence of outliers.
- How to interpret graphs, frequency tables, measures of central tendency, and measures of dispersion both by using statistical jargon and *without* using statistical jargon. (You should be able to explain what any of these concepts are and what they mean to someone not trained in statistics.)

- Finally, to use proper and precise vocabulary to express yourself both orally and in writing when discussing statistics concepts — including *variables, measurement, operationalization, levels of measurement, units of analysis, units of measurement, etc.*
- **Hint/Warning: If any of the above gives you trouble, go back and reread the relevant section. Proceeding further with gaps in your knowledge will only make things worse. (There is no hope that by reading the more complicated material which follows you will suddenly learn/understand the things discussed so far!)**

---

# Chapter 5 The Normal Distribution and Some Basics of Probability

A variable's distribution, you recall, is the way the observations/cases are distributed across the variable's categories. Frequency tables, graphs, as well as measures of central tendency and dispersion all provide information about the distributions of variables.

All variables have a distribution (of course!) but some variables have a special type of distribution: one whose features and uses in statistics go beyond being simply “a variable's distribution”. We call this distribution *normal distribution*.

In the first part of this chapter I introduce the normal distribution, detailing its features that make it so special. The latter half of the chapter is devoted to a concept without which we wouldn't be able to do any statistical inference and estimation, namely statistical probability. You will learn some basics of probability theory which are necessary for us to eventually proceed to statistical inference.

You might be wondering why these two seemingly unrelated things — a variable's distribution and probability theory — are in the same chapter together. For now I will just give you a hint: probabilities have distributions too. Read on to find out more.

---

## 5.1 The Normal Distribution

You might have already heard of bell curves (or bell-shaped curves), or even normal curves. If you have, you also probably know they look similar to the one in Fig. 5.1.

*Figure 5.1 Body Mass Index of Respondents (CCHS 2015/2016)*

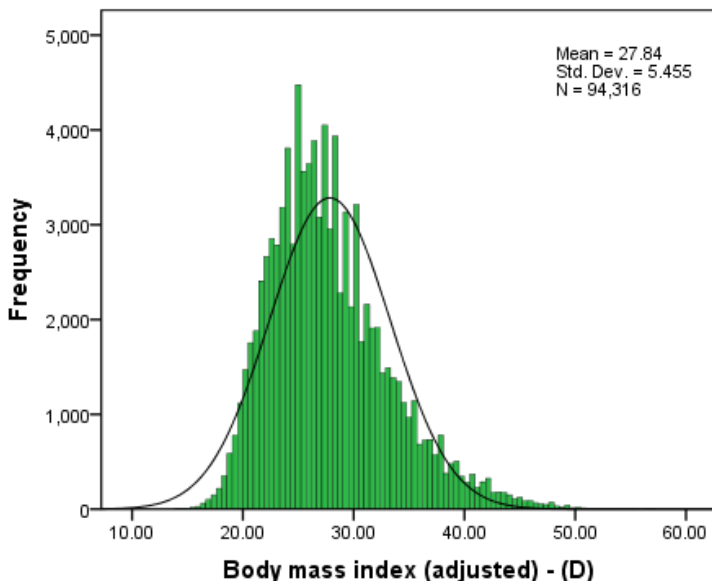


Fig. 5.1 shows a histogram with the distribution of the variable *body mass index* (or *BMI*) of respondents to the CCHS 2015/2016. Judging by the height of the bars that comprise it, the histogram illustrates the fact that most cases tend to cluster at the centre (i.e., most people's *BMI*

is average), while a decreasing number of cases end up in the “tails” of the distribution (i.e., the further their *BMI* is from the average, the fewer cases there are).

You can easily notice that the distribution (as reflected in the green bars) is not perfectly symmetric but a bit positively skewed: the right “tail” is longer than the left. Still, its shape approximates a bell well-enough (note for comparison the black curve in Fig. 5.1 which is a true bell shape). **We call this type of distribution *approximately normal*.**

A great many interval/ratio variables in the world tend to have an approximately normal distribution when plotted (true for both the social and natural sciences). That is, the majority of observations are centered in the middle of the distribution (i.e., they tend to be *average*); we find fewer observations just below and just above the average, and fewer still which are much below or much above the average.

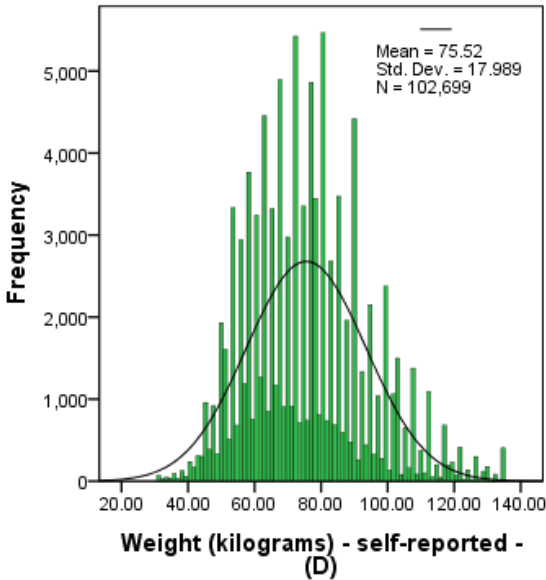
Think about height, for example. Most people are of average height (that’s why it’s called *average* height after all), some people are above and some below average, fewer people are much taller or shorter, and rather rarely are some people extremely short or extremely tall. Variables like age, or weight (which you can see in Fig. 5.2 below<sup>1</sup>) but also,

1. The reason you observe the “double” distribution -- one shorter (darker) while the other taller (lighter) -- is due to the self-reporting of weight. Most people tend to report their weight in whole numbers, and here some have done so, stating their weight as 65 kg or 85 kg, etc.; these are the tall bars. Others, however, may have reported it with grams and/or in pounds (which when converted to kilograms would produce a non-whole number weight), thus resulting in weights such as 65.35 kg or 85.75 kg, etc., leading to the short bars and to the histogram appearing like two histograms plotted on



say, test marks, or points scored per hockey game, or text messages sent per day, etc. are similar. There will be an average, and a continuous decrease in frequency the further one gets from that average.

*Fig. 5.2 Weight of Respondents (CCHS 2015/2016)*



*As fascinating as all this is, you might be thinking now, why do we care about it? It's just one type of a distribution among many.*

True, but as I already mentioned, the normal distribution is special, and not just because many variables' histograms tend to plot an approximately normal curve. To understand why, we need to start exploring the normal distribution as a

top of each other. Had the responses been rounded to the nearest whole kilogram, the histogram would have taken a regular, "single" normal-curve shape.

*theoretical* concept (or, to borrow from Max Weber, as an *ideal type*).

---

## 5.1.1 Properties of the Normal Curve

Recall that we describe a distribution via three things: its shape, its central tendency measures, and its measures of dispersion. **The perfect (i.e., theoretical) normal distribution thus has three defining features.**

First, the normal curve is **bell-shaped and perfectly symmetric** (i.e., if you bisect it in the middle, the left side will be identical to the right side).<sup>1</sup>

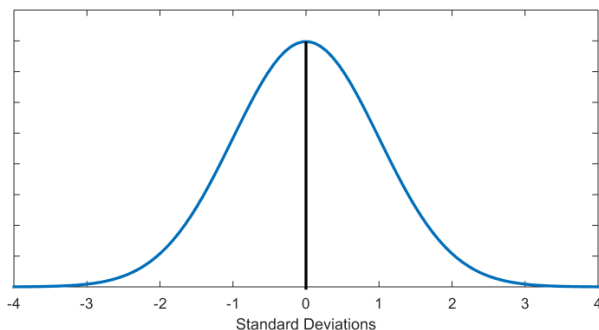
Second, the normal curve is **centered on the mean**, which also happens to be equal to its median and mode. That is, for the normal curve **all measures of central tendency fall on the same value.**

Third, **the normal curve's standard deviation tell us what percentage of observations fall within a specific distance from the mean.** When we have a normal curve, the area below the curve contains 100 percent of all observations. Then, 68 percent of all observations fall within 1 standard deviation from the mean<sup>2</sup>; 95 percent of observations fall within about 2 standard deviations from

1. It's also asymptotic to the horizontal axis line, i.e., it gets as close to it as possible in the "tails" without ever touching it. More on this after you learn about probabilities.
2. Given the symmetry, this means 34 percent fall within -1 standard deviation below and 34 percent fall within +1 standard deviation above the mean.

the mean<sup>3</sup>; and 99 percent of observations fall within about 3 standard deviations from the mean<sup>4</sup>. Fig. 5.3 illustrates.

*Figure 5.3 Normal Curve with Standard Deviations*



If you imagine Fig. 5.3 interposed on top of an approximately distributed variable's histogram, you can see what percentage of observations will fall within 1, 2, and 3 standard deviations from the mean. (Obviously, the mean is at 0, since the normal curve is centered on the midway point of the curve, and is neither below nor above itself, i.e., “the mean is 0 standard deviations away from the mean”, as awkward as it sounds.)

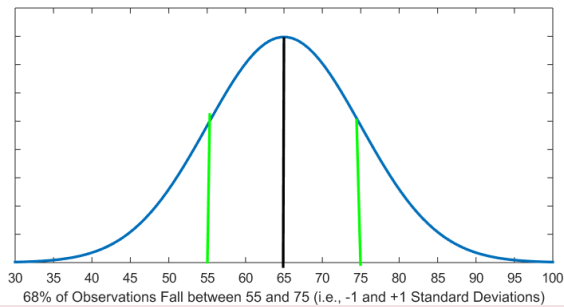
Let's make sure this makes sense to you in applied terms, through the example below.

*Example 5.1 Normally Distributed Test Scores (Hypothetical Data)*

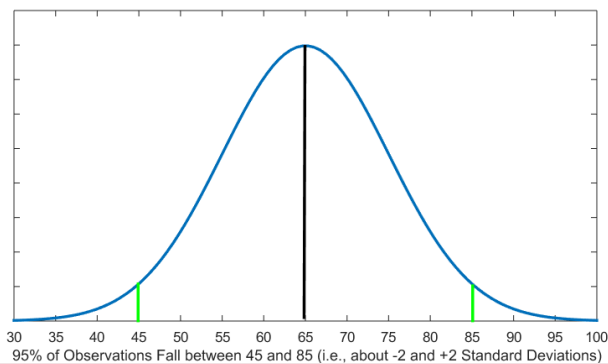
3. That is, 47.5 percent fall within about -2 standard deviations below the mean and 47.5 fall within about +2 standard deviations above the mean.
4. That is, 49.5 percent fall within about -3 standard deviations below the mean and 49.5 percent fall within about +3 standard deviations above the mean

Imagine your statistics class has taken a test. The average test score is 65 with a standard deviation of 10 and the following scores distribution. (You can imagine a histogram whose many bars follow the curve in the three Fig. 5.4 below.)

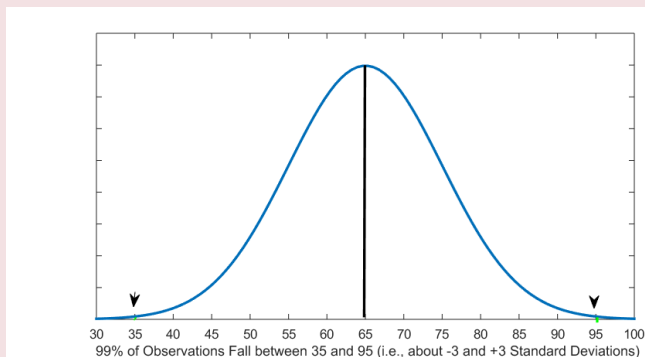
*Figure 5.4 (A) Test Scores within 1 Standard Deviation*



*Figure 5.4 (B) Test Scores within About 2 Standard Deviations*



*Figure 5.4 (B) Test Scores within About 3 Standard Deviations*



Given the properties of the normal curve, we now know that 68 percent of students in the class scored between 55 and 75 (i.e., between -1 and +1 standard deviations from the mean, and since the standard deviation is 10, then  $65 - 10 = 55$  and  $65 + 10 = 75$ ). We also know that 95 percent of students scored approximately between 45 and 85 (i.e., between about -2 and +2 standard deviations from the mean, or  $65 - 2(10) = 65 - 20 = 45$  and  $65 + 2(10) = 65 + 20 = 85$ ). Finally, we know that 99 percent of students (almost everyone!) scored approximately between 35 and 95 (i.e., between -3 and +3 standard deviations from the mean, or  $65 - 3(10) = 65 - 30 = 35$  and  $65 + 3(10) = 65 + 30 = 95$ ).

As is typical of normal distributions, the majority of scores (68

percent) are clustered in the middle (within  $-1$  and  $+1$  standard deviations) around the mean; the remaining 32 percent are split between the “tails” of the distribution, with about 16 percent in each “tail” beyond  $-1$  and beyond  $+1$  standard deviation from the mean. Only 5 percent of test scores are as far away as  $-2$  and  $+2$  standard deviations from the mean, with just 2.5 percent at the tips of each of the “tails”. And at the very, very far ends of the “tails”, beyond the  $-3$  and  $+3$  standard deviations from the mean, you have 1 percent split between them, so a minuscule 0.5 percent of students has a score below 35 and another 0.5 percent has a score above 95.

These features of the normal distribution (symmetrical, centered on the mean/median/mode, measured in standard deviations from the mean) make it very useful to work with. Simultaneously, now you can begin to see why the standard deviation is the most popular measure of dispersion, due to its unique relationship with the normal curve.

Can we find more uses of the normal distribution? Read on to find out.





---

## 5.1.2 The z-Value

In the previous section you discovered that we can “orient” ourselves about where a specific value lies along the normal distribution in relation to the average by means of the standard deviation. In Example 5.1 we saw that 68 percent of students’ test scores were between 55 and 75 (i.e., between  $-1$  and  $+1$  standard deviations from the mean), 95 percent of scores were between approximately 45 and 85 (i.e., between about  $-2$  and  $+2$  standard deviations from the mean), and that 99 percent of scores were between approximately 35 and 95 (i.e., between  $-3$  and  $+3$  standard deviations from the mean). Thus, if your score was, say, 60, you would know that it was below the mean, but within 1 standard deviation away, which wouldn’t be as bad as, say, had you scored 40, which is more than two standard deviations away from the mean.

*Hmm, do we really need standard deviations to tell us that a test score of 40 is bad news, you ask. Everyone knows that.*

In absolute terms, sure, a score of 40 (out of 100) would be considered a failing one. In relative terms, however — which is also known as grading on a curve — a score of 40 doesn’t tell you anything, unless you know the mean and the standard deviation.

To better illustrate this, imagine another set of test scores, and that on that test you get a score of 80. In absolute terms, a score of 80 (out of 100) would be quite good. What about in *relative* terms? Can you think of a situation where a score of 80 would be considered worse than a score of 40?

What if I told you that the mean in the first case (when we imagine you scored 40) was 35 with a standard deviation of 5, while the mean in the second case (when we imagined you scored 80) was 90 with a standard deviation of 2? (You might find it easier to see the point if you grab a pen and paper and simply draw a line with the mean in the middle, then add and subtract that many standard deviations away from it in each direction, above and below.)

A score of 40 (i.e.,  $35 + 5 = 40$ ) is 1 standard deviation *above the mean* of that test. A score of 80 (i.e.,  $90 - 5(2) = 80$ ) is 5 standard deviations *below the mean* of that other test. In fact, 80 is well below the even 3 standard deviations away from the mean where 99 percent of scores are; it's at the very far end of the left "tail" of the distribution, likely an outlier.

It turns out that the second test we imagined was so easy, scoring 80 on it was too low given how easy it was. On the other hand, scoring 40 on the first test we imagined was quite good given how hard it was.

This mental exercise shows you that **expressing values in terms of standard deviations** has its merits, as it **puts the values into perspective** — which allows us to make comparisons. A score/value in and of itself doesn't tell

you anything — not unless you know *where it falls in relation to the mean and how far away it is*. Now only if there was a way to express *any* value in terms of standard deviations without having to always calculate 1 standard deviation away, 2 standard deviations away, 3 standard deviations away from the mean (or to have to resort to pen and paper)...

Guess what? There is! **Expressing a value in terms of standard deviations is a process aptly called *standardization*** (as it produces scores that have a uniform, *standard* meaning allowing comparison) **and the standardized values are called *z-values* (or *z-scores*)**. We **standardize values by expressing the distance of the value from the mean in standard deviations, i.e.:**

$$\frac{\text{original score} - \text{mean}}{\text{standard deviation}} = \text{z-value}$$

Or, in proper notation, where we denote the mean by  $\mu$ <sup>1</sup>, the small-case Greek letter for *m* (from *mean*):

$$\frac{x_i - \mu}{\sigma} = z$$

Following this formula, a score of 40 when the mean is 35 and the standard deviation is 5 (i.e., when  $\mu=35$  and  $\sigma=5$ ) has a z-score of

1. The Greek letter  $\mu$  is pronounced as "MYU". The difference between using  $\bar{x}$  and  $\mu$  and the reason we use the latter here will be explained in Chapter 6.

$$\frac{x_i - \mu}{\sigma} = \frac{40 - 35}{5} = \frac{5}{5} = 1 = z$$

and a score of 80 when the mean is 90 and the standard deviation is 2 (i.e., when  $\mu=90$  and  $\sigma=2$ ) has a z-score of

$$\frac{x_i - \mu}{\sigma} = \frac{80 - 90}{2} = \frac{-10}{2} = -5 = z$$

Thus, we formally found what we already knew from before: that in the former case, the score of 40 was 1 standard deviation above the mean (i.e., its  $z = 1$ ) and the score of 80 was 5 standard deviations below the mean (i.e., its  $z = -5$ ). If this seems repetitive — after all, we reached the same conclusion without any fancy formulas — that’s only because I chose easily calculatable numbers to illustrate my point more easily. Perhaps an example with less “easy” numbers will convince you of the formula’s worth.

*Example 5.2 Average Monthly Rent for a Two-Bedroom Apartment in Vancouver*

*The Vancouver Sun* recently reported that the average monthly rent of a two-bedroom apartment in Vancouver, BC was \$2,915, at the time of writing the highest in all Canada. (REFERENCE <https://vancouversun.com/news/local-news/vancouver-two-bedroom-apartments-now-cost-close->

to-3000-report) While the standard deviation was not reported, for the purposes of this exercise we can imagine it as \$150.

What is the z-score of a family which pays \$2,630 per month for their two-bedroom condo? How about the z-score of someone who pays \$3,450 for theirs?

Of course, we could grab a pen and paper and draw the normal distribution demarcating where 1, 2, and 3 standard deviations away from the mean fall in order to see where the two listed rents are relative to the demarcations. However, using the z-score formula makes for a faster (and a more precise) answer.

In the first case, we have:

$$\frac{x_i - \mu}{\sigma} = \frac{2630 - 2915}{150} = \frac{-285}{150} = -1.9 = z$$

In the second case, we have:

$$\frac{x_i - \mu}{\sigma} = \frac{3450 - 2915}{150} = \frac{535}{150} = 3.6 = z$$

That is, the first family's monthly rent of \$2,630 is below the average but not that unusual: with a z-score of -1.9, it falls within 2 standard deviations away from the mean, which is within what

95 percent of renters in Vancouver pay for their two-bedroom apartments.

On the other hand, the second person's rent of \$3,450 is quite high: with its z-score of 3.6, it falls beyond 3 standard deviations away from the mean, i.e., it's higher than what 99 of people pay monthly for a two-bedroom apartment.

Again, we see the use of standardization and z-scores, as it allows us to put values into perspective.

Now is your turn to try.

*Do It! 5.1 Comparing Average Monthly Rent for a One-Bedroom Apartment in Vancouver, Toronto, and Montreal*

According to the *National Rent Rankings* monthly report for July 2019 by Rentals.ca (REFERENCE <https://rentals.ca/national-rent-report>), the average monthly rent for a one-bedroom apartment was \$2,028 in Vancouver, BC, \$2,259 in Toronto, ON, and \$1,231 in Montreal, QC. Assume the standard deviations are \$140 in Vancouver, \$180 in Toronto, and \$125 in Montreal.

Using z-values, compare and analyze where in the

distribution a rent of \$1,950 will put a Vancouverite, a Torontonion, and a Montrealer who all pay the same rent but in different cities.

(Answer: Vancouverite's  $z=-0.6$ , Torontonion's  $z=-1.7$ , Montrealer's  $z=5.8$ .)





---

## 5.1.3 Percentiles

Remember quartiles? We used them in Section 4.2 to find the interquartile range (<https://pressbooks.bccampus.ca/simplestats/chapter/4-2-interquartile-range/>). They would split the cases in the distribution in four equal parts (i.e., in quarters) giving us a first (1 percent to 25 percent of the data), a second (26 percent to 50 percent of the data), a third (51 percent to 75 percent of the data), and a fourth quartile (76-100 percent of the data).

What if, instead of splitting the distribution into *four* equal parts, we decided to divide it into *five*? That would be easy: Instead of having four parts, 25 percent of the data in each, we can just have five parts, 20 percent of the data in each. Like this: 1 percent to 20 percent, 21 percent to 40 percent, 41 percent to 60 percent, 61 percent to 80 percent, and 81 percent to 100 percent. This time, we call the five equal parts *quintiles* (from the Latin root “quin” like *quinctus*, meaning five).

Just as easily, we can divide the distribution into *ten* equal parts: 1 percent to 10 percent, 11 percent to 20 percent, etc. ... all the way up to the last part, 91 percent to 100 percent. Then we have ten *deciles* (from the Latin root “dec” like *decem*, meaning ten).

Following the same logic to the smallest possible whole number by which we can divide a distribution, we get

*percentiles* — a distribution divided into a hundred equal parts, 1 percent in each. It turns out percentiles can be quite useful when working with a normal distribution. (You didn't forget that's our current topic, did you?)

The key piece of knowledge you need to recall from our discussion about quartiles is that to split the distribution, we need the cases lined up in order from the lowest value to the highest (or else we wouldn't be able to speak of first, second, third or last quartiles). Applying this to the normal distribution, we might be tempted to imagine the normal curve as illustrated in Fig. 5.5 below.

*Figure 5.5 What Percentiles Do Not Look Like*

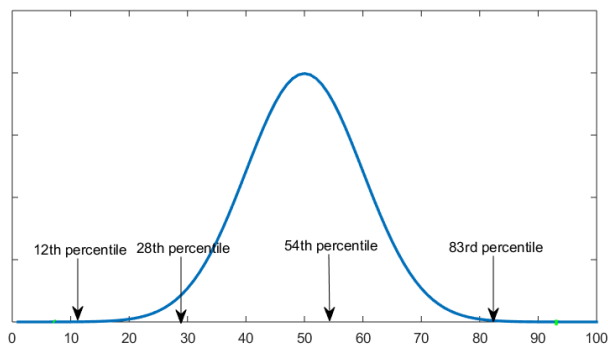


Fig. 5.5 lists the position of four randomly selected percentiles, *had the percentiles been evenly spread over the horizontal axis*. Of course, this is wrong. If we do this, we would be ignoring the *actual* distribution — you know, the blue curve on the graph. After all, we have established by now that 68 percent of observations fall in the middle, within only 1 standard deviation way from the mean, where the curve is at its highest. (Recall that the height of the curve — and the fact that it's a *curve*, not a *line* — reflects

the larger frequencies of the values around the mean, and the smaller, and smaller frequencies of the values further away from the mean, in the “tails”).

What this should tell you is that we can’t just assume the percentiles are uniformly spread — because the data is not. We need to account for the fact that that values in the middle are way more popular than the ones in the “tails”. Then how do we know what percentile a particular value has?

Again, it’s easy. We have z-scores for that. You see, every value has a z-score and the z-score reflects the percentage of cases which fall below or above that value. This is precisely the reason we know that 68 percent of the data fall within 1 standard deviation from the mean and that 95 percent of data falls within about 2 standard deviations from the mean.

Thus, with a normal distribution, you can turn any value into a z-score (as we saw in the previous section), and this z-score into a percentile. While there are z-score tables providing percentages associated with any z-value, the easiest way to find a percentile is through online calculators like this one by *Measuring U*: <https://measuringu.com/pcalcz/>.<sup>1</sup> There, you can enter a z-score (make sure you choose “one-sided”) and see what percent of data falls below it (on the normal curve on the left) and what percent of data falls above it (on the normal curve on the right). The exact percentile is the number reflecting the data “below”.

1. For that matter, you can use an online calculator to find the z-score of any value. You can try one here (provided by *Social Science Statistics*): <https://www.socscistatistics.com/tests/ztest/zscorecalculator.aspx>.

### *Do It! 5.2 Finding Percentiles Using an Online Calculator*

Using the percentile calculator linked above, you find that the percentile for  $z=1$  is 84. Explain where this result comes from. (Hint: The mean bisects the distribution in two equal halves. A  $z$ -score of 1 is of course 1 standard deviation *above* the mean.)

Answer: The area below the mean is 50 percent. To that we add the 34 percent between the mean and 1 standard deviation above the mean and get  $50+34=84$  percent. (Since 68 percent lies between  $-1$  and  $+1$  standard deviations and the normal curve is symmetrical, 34 percent fall between  $-1$  standard deviation and the mean, and 34 percent fall between the mean and  $+1$  standard deviation).

*Cool, you say (probably quite sarcastically), we now know how to find percentiles. But for what do we use them?*

I'm glad you asked. **Percentiles allow us to compare a score in relation to the rest of the data; just like  $z$ -scores, they put things into perspective.** Let's say you have 69 on a test. Turning your score into a percentile will tell you *exactly* what percent of the test-takers scored *below* you, whether it's 35 percent (then your score wouldn't be considered too impressive) or 99 percent (which would be

most impressive, seeing how you'd be in the top 1 percent of test-takers) or any other percent it might be.<sup>2</sup>

Let's make sure you understand all that, shall we?

### *Do It! 5.3 Hourly Wage*

Imagine you have applied for a job and your employer offers you \$13.5/hour. You also learn that the average hourly wage your potential employer pays to their employees is \$17.5/hour with a standard deviation of \$2.5/hour. See if this is a generous offer (after all, you would be just starting) by finding its z-score and percentile and comparing it to how the other employees of the company are fairing. (Don't forget to interpret both the percentile and the z-score.)

Answer:  $z = -1.6$ , percentile = 5.5. Only 5.5 percent of the employees in the company receive less than \$13.5/hour; almost 95 percent of the employees receive more, so no, it's not a generous offer at all.

And now that you might be starting to feel somewhat comfortable with the uses of the normal distribution, I'll pull the rug a bit from under you, as it were. Recall how

2. This is exactly what standardized tests (e.g., SAT) do to interpret individual scores. They provide percentiles so that any test-taker can find how they did *relative to others* (i.e., it provides the place of a score in the overall distribution of scores).

I started the chapter by explaining that many real-world interval/ratio variables tend to be approximately normally distributed? (That part's true.) And then we talked about where the variable's observations fall in the normal distribution? Well, there I lied. (It was necessary!)

If you think about it carefully, both statements cannot be true. On the one hand, a real-existing variable has a specific distribution — an *approximately* normal one. But would two real-existing variables have *exactly the same* approximately normal distribution? That would be unlikely, considering that different variables, in different datasets, with different number of observations, units of measurements, units of analysis, means and standard deviations, etc. cannot possibly look exactly the same if plotted on a histogram. How then do we get these *very fixed* and *very specific* numbers and percentages associated with the z-scores and the percentiles?

The thing is, everything I told you about the normal distribution, starting with its defining features and ending with the z-scores and percentiles, refers to the ideal-type, only-existing-in-theory, perfect normal distribution. All the numbers and calculations and percentages we discussed reflect the *theoretical* normal distribution; they serve as a sort of *expectation* of how a (continuous, random)<sup>3</sup> variable *is expected* to be distributed. Of course, real-existing variables generally fall short of this ideal, and therefore we call their distributions *approximately* normal.

3. I explain randomness a bit in the next section, and further in Chapter 6. For now, know that in statistics it doesn't mean "arbitrary" or "accidental" but rather "obtained in an unbiased way" (i.e., with every element having an equal chance to be picked).

I repeat: **the theoretical (perfect) normal distribution provides us with what we can *expect* the actual frequencies of the variable's values to be, in theory.** (In reality, the distribution differs from that expectation to varying degrees). It turns out, **when we work with z-scores and associated percentages and percentiles, we work with what is *expected*, not with what *is*.** (The variables' observed distributions differ but the normal — expected — distribution is always the same.)

What do we do then, with this reality versus expectation we have here? Why did we learn all we did about the normal distribution if “it isn't real”?<sup>4</sup>

This is where probability comes in. Hold the thought about the normal distribution being an expectation; we'll come back to it in the remaining sections of this chapter.

4. That said, again, some standardized tests can be designed in such a way that their test scores to be distributed normally. Thus, real-existing data *can* have a normal distribution, it's just that usually it's an approximation.





---

## 5.2 Probability Basics

Whenever we talk about **the likelihood of some future event taking place**, we talk about ***probability***. This likelihood serves as a prediction — what we can expect to happen or not happen. For example, people might mention the odds of winning the lottery, or the probability of being hit by lightning, or to discuss the fact that it's likelier to die in a car accident rather than an airplane crash, or to think that the odds of having a baby girl are the same as the odds of having a baby boy. Sociologists in particular might typically be interested in an individual's life chances, things like the probability of going to college, the probability of being unemployed, or to have a high-paying job, etc. and comparing the probabilities for any of these happening based on characteristics like race/ethnicity, gender, socioeconomic class, religion, sexual orientation, etc.

Probability is predicated on uncertainty; as the old song goes, “the future's not ours to see”. We use probabilities to manage the uncertainty, usually by quantifying it. For example, life expectancy at birth is the predicted longevity that a newborn will have (given current death rates). Or you might have even taken important decisions and made choices based on odds and likelihoods (i.e., on probabilities). An entire industry — betting and gambling — is based on the fact that we don't know what *will* happen but we nevertheless try to predict what *might* happen.

Given the dealing with uncertainty and predictions, it shouldn't be too surprising that probability is completely and entirely *theoretical*. It's an *expectation* for the future, which can't be anything but abstract. (After all if something had already happened, and has become reality, we wouldn't need to predict it or to discuss its probability of occurring.)

Let's start with an example which is familiar to absolutely everyone, usually from an early age. At some point in your life you have likely uttered the phrase "there's a fifty-fifty chance of..." Like "I didn't do too well on my last test, by now there's a fifty-fifty chance to pass the course." Or "the traffic looks bad but it might clear up; I still have a fifty-fifty chance of making it to the job interview on time." Or "this plan has a fifty-fifty chance of success." Or even "these nachos look disgusting, you have a fifty-fifty chance to get food poisoning."

A *fifty-fifty chance* of course means *an equal probability of something happening or not*. Out of two possible outcomes, either can happen with equal likelihood so it's impossible to predict in favour of any of them.

I'm sure you know that the fifty-fifty chance expression comes from the impossibility of predicting the outcome of a flipped coin: be it heads or tails. Assuming a coin cannot possibly fall on its edge, when flipped it has only two outcomes, represented by its two sides, falling as heads or as tails. Thus, the probability of its falling on a side (a 100 percent) is divided by two — giving us 50 percent chance to get heads and 50 percent chance to get tails.

The 50/50 percent is a *prediction*. The moment the coin

falls, one outcome has been realized and the prediction no longer applies because the event is no longer in the future. The distinction between the *factual* reality (the event has happened) versus the *theoretical* probability<sup>1</sup> (of the event happening) might seem trivially easy to make at this point but it's nevertheless very important. Keep it in mind, you'll need it for what's to come.

Imagine you flip a coin two times in a row. Can you predict that you'll get once heads and the other time tails? Is it possible that you get heads twice in a row? What if you flip a coin ten times? Would you get tails exactly 5 times and heads exactly 5 times? Or could you perhaps get 3 heads and 7 tails? What about 9 times heads and 1 time tails? And what if you flip a coin a hundred times? Or more?

You might have already reasoned it, or you might have even tried it at some point: it's quite possible to flip a coin and get the same side twice in a row. Or three times. Or four times. Or more. (It's even possible to flip heads ten out of ten times in a row... or even a hundred out of a hundred. In this case *possible* means that there is such a probability, as small as it is. *Possible* doesn't mean necessarily *plausible*.) How do you reconcile this with the knowledge that the probability of getting heads is 50 percent?

And that — the *probability* — is just it. We know that *theoretically* with each coin toss the coin can fall as either

1. Note that the theoretical probability is still grounded in the reality of there being only two possible outcomes. Thus predictions we base on probability are not wild, baseless guesses but a product of rational thinking and calculations.

heads or tails, and the prediction/expectation is a fifty-fifty chance. We know that *in theory*, if we flipped coins *forever*, heads and tails will average at 50 percent of the time each<sup>2</sup>. We can't flip coins forever, however, so it's possible we get a different outcomes distribution in any finite number of times we do it (but the larger the number of times, the likelier we'll be getting to 50/50 percent, or close<sup>3</sup>).

Thus there is no contradiction in *theoretically expecting* a fifty-fifty chance of flipping tails out of, say, ten tosses and *actually getting* heads 6 times and tails only 4, as I'm sure you know. The former is a probability distribution, the latter is the observed, actual frequency distribution of the cases/observations/data. Keep this thought too.

Before we continue on to something more novel and exciting than the old coin toss example, however, let formalize our discussion a bit.

2. This website provides a neat visualization of both the probability/expectation and a digital coin toss: <https://seeing-theory.brown.edu/basic-probability/index.html>. There you can try flipping the coin 100, even 1000 times, and see that the larger the number of flips, the closer you get to the fifty-fifty expectation. The same website allows you to throw a die and to pick a card out of ten consecutively numbered cards
3. You can find more on this property of large numbers in Chapter 6.

---

## 5.2.1 Working with Probabilities

**We express probabilities as proportions** (and we also denote them with  $p$ , just like we do proportions<sup>1</sup>), as this is indeed what they are:

$$p = \frac{\text{number of specific outcomes we are interested in}}{\text{number of all possible outcomes}}$$

Or, the probability of a specific outcome is the proportion of the number of such outcomes out of the number of all possible outcomes.

Thus the probability of getting heads in a coin toss is:

$$p(\text{heads}) = \frac{\text{number of heads sides of a coin}}{\text{number of all sides of a coin}} = \frac{1}{2} = 0.5$$

The same of course applies to tails:

1. If you need a reminder, the relevant part is in Section 2.3.1, here:  
<https://pressbooks.bccampus.ca/simplestats/chapter/2-3-1-adding-percentages/>

$$p(\text{tails}) = \frac{\text{number of tails sides of a coin}}{\text{number of all sides of a coin}} = \frac{1}{2} = 0.5$$

Heads and tails together exhaust all possible outcomes, so the probability that a coin will fall on any of its two sides is:

$$p(\text{heads or tails}) = \frac{2}{2} = \frac{1}{2} + \frac{1}{2} = 0.5 + 0.5 = 1$$

Now how about we extend our example to something that has more than two outcomes? With six sides, a conventional die will serve us perfectly.

Following the same logic as with the coin, the probability to throw, say, a five is:

$$p(\text{five}) = \frac{\text{number of "five" sides of a die}}{\text{number of all sides of a die}} = \frac{1}{6} = 0.167$$

The same goes for throwing a one, a two, a three, a four, or a six:

$$p(\text{one}) = p(\text{two}) = p(\text{three}) = p(\text{four}) = p(\text{five}) = p(\text{six}) = \frac{1}{6} = 0.167$$

Or, imagine you have a bowl with ten balls inside (i.e., the balls have numbers from 1 to 10). The probability of selecting each one out (without looking!) is, you guessed it, 1 out of 10, as each number appears only once and there are ten possible outcomes:

$$p(1) = p(2) = p(3) = \dots = p(10) = \frac{1}{10} = 0.1$$

While this principle applies to  $N$  of any size — so we can increase the number of outcomes as much as we want — note **the key prerequisite for the calculations to work: the outcomes must happen randomly.** A coin toss and a die throw are classical examples of random chance. But when picking balls out of a bowl we have to make sure we don't look or we might (consciously or subconsciously) *choose* one. Choosing a ball with a specific number introduces bias and thus invalidates randomness — i.e., it invalidates the principle of the outcomes having the same probability. Without this principle we cannot calculate anything: the only way to know the probability of an outcome is, in a sense, to divide the total probability, as it were, (i.e., 1) by the number of all possible outcomes, giving us equal probability for each. **We know the probability of an outcome *only if* we know how many outcomes are possible in total and they all have the same probability.** (Chapter 6 has more on the topic as it's devoted to the topic of how random selection works.)





---

## 5.2.2 Simple Probability Calculations

This section is a brief side quest which shows you how to calculate combinations of probabilities. For example, back to die rolling, what is the probability of throwing a two *or* a four?

I'm certain you already know the answer. In this case the "outcomes of interest" are two instead of one, so the probability is two out of six possible outcomes:

$$p(\text{two or four}) = \frac{\text{number of outcomes we are interested in}}{\text{number of all outcomes}} = \frac{2}{6} = \frac{1}{3} = 0.333$$

Or I could have just as easily simply added the two outcomes' individual probabilities:

$$p(\text{two or four}) = \frac{\text{number of two's}}{\text{all outcomes}} + \frac{\text{number of four's}}{\text{all outcomes}} = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = 0.333$$

And this is it: **to combine the probabilities of two outcomes which cannot happen at the same time (a.k.a. *disjoint events*<sup>1</sup>), you simply have to add them together.**

1. You can recognize disjoint event by the usage of "or": it's one *or* the other (*or* a third one, etc.). When flipping one coin, you can either get heads *or* tails; when you roll one die, you can get *only one* of its sides at a time. Hence, we *add* their probabilities.

(Recall we already used this when we started with the probability of getting heads *or* tails being 1; it's simply the probability of getting heads (0.5) added to the probability of getting tails (0.5)).

#### *Do It! 5.4 Adding Probabilities*

Since we already imagined a bowl with ten consecutively numbered balls inside, let's save ourselves the effort of imagining a new one and reuse it again. What is the probability of randomly selecting the #5 ball *or* the #7 ball *or* the #9 ball out of the ten numbered balls in our bowl?

(Answer: 0.3)

On the other hand, **combining probabilities of events that *can* happen at the same time, or that happen *one after another* in time (both a.k.a. *independent events*<sup>2</sup>)** is a tad more complicated and **requires multiplication**.

For example, the probability of throwing double two's when rolling two dice (or throwing a two with one die and then immediately throwing again another two) is:

2. Events are called independent when the outcome of one doesn't affect the outcome of the other whatsoever. (Contrast this with getting heads in a coin toss, which precludes getting tails; same with throwing any number on a die as it precludes the other numbers from being thrown.)

$$\begin{aligned} p(\text{double two's}) &= \frac{\text{number of two's (1st die)}}{\text{all outcomes (1st die)}} \times \frac{\text{number of two's (2nd die)}}{\text{all outcomes (2nd die)}} = \\ &= \frac{1}{6} \times \frac{1}{6} = \frac{1}{36} = 0.028 \end{aligned}$$

Or, if we flip a coin three times (or three coins at the same time), the probability of getting three tails is the probability of getting tails once out of one coin flip (i.e., 0.5) multiplied by the same probability and then multiplied by the same probability again (or simply  $0.5^3$ ):

$$p(\text{three tails}) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} = 0.125$$

Thus the probability of flipping three tails in a row (or three tails with three coins at the same time) is 1.25 percent.

### *Do it! 5.5 Multiplying Probabilities*

Using the same imaginary bowl with ten consecutively numbered balls inside as in the previous exercise, what is the probability of randomly selecting first the #3 ball, then the #4 ball, and then the #5 ball, *if you return the selected balls immediately back in the bowl before selecting the next one?*

(Answer: 0.001)

Now take the time to note the italicized condition at the end of the question in the exercise you just did. It's important enough to necessitate its own scary-red warning,

**Watch Out!! #10...** *for Replacement When Working with Probabilities*

What would have happened had I not specified that in the calculation in *Do It! 5.5* you should consider the selected balls being returned right after their random selection? Why, you would have tempered with the number of all possible outcomes, of course.

After all, after randomly selecting the first ball, *unless you imagine returning it back in the bowl*, there will be only  $(10-1=)$  9 balls left from which to make the second selection. Then after removing the second ball, *and again not returning it back in the bowl*, you'd have left only  $(9-1=)$  8 imaginary balls from which to select your third ball. Then, unlike the  $\frac{1}{10} \times \frac{1}{10} \times \frac{1}{10}$  you should have used above, the calculation now becomes:

$$p(\text{"3", "4", "5" balls in a row}) = \frac{1}{10} \times \frac{1}{9} \times \frac{1}{8} = 0.0013$$

The difference between this result and the one in the exercise seems small but that's only because we're working with small numbers. It's still important to understand how random selection *with replacement* differs from random selection *without replacement* and to use the correct calculations.

Before we move on using probabilities with actual data, you could use a bit more practice.

*Do It! 5.6 Adding and Multiplying Probabilities, With and Without Replacement*

Imagine you and four of your friends (let's call them Adam, Bhav, Chen, and Dila) are in a class of 25 students. Assume that it's the first time your class meets and your professor doesn't know any of you; she only has the class roster in front of her so any name she calls, she calls from the roster at random. Answer the following questions:

- What is the probability that your professor will call your name?

- What is the probability that she calls on Bhav?
- What is the probability that she calls on you, then Chen, and then Dila, one after the other? (Hint: She won't call a name twice in a row, she remembers that much.)
- What is the probability that she calls either your name or Adam's?
- What is the probability that she calls on any one of your friends?
- Your professor also needs to randomly pair up students for a group assignment; what is the probability that she selects Chen and Dila to be in the same group?

(Answers: 0.04; 0.04; 0.000; 0.08; 0.16; 0.002)

---

## 5.2.3 Probabilities with Frequency Tables

So far we've been working only with small- $N$  examples but there is no reason to think what you learned from coins and dice and balls in bowls will not apply to actual, large- $N$  data.

We already established that probabilities are proportions, and they can also be expressed in percentage terms. Conveniently enough, I had the foresight to introduce percentages (a.k.a relative frequency) as early as Section 2.3.1 (<https://pressbooks.bccampus.ca/simplestats/chapter/2-3-1-adding-percentages/>). (I am that wise.) It turns out, we can work with the percentages we find in frequency tables as easily as we can with any of the imaginary examples we did in the previous sections. I'll prove my claim with an example.

### *Example 5.3 Social Class (GSS 2016)*

Supposedly everyone thinks they're middle class and Canadians are not different. And while Table 5.1 shows that not really everyone thinks so, the majority of them do.

*Table 5.1 Respondent's Social Class (GSS 2016)*

Social class					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Upper class	233	1.2	1.2	1.2
	Upper-middle class	3321	16.9	17.3	18.5
	Middle class	12230	62.4	63.8	82.4
	Lower-middle class	2749	14.0	14.3	96.7
	Lower class	628	3.2	3.3	100.0
	Total	19161	97.7	100.0	
Missing	Don't know	312	1.6		
	Refusal	118	.6		
	Not stated	18	.1		
	Total	448	2.3		
Total		19609	100.0		

Out of all 19,161 respondents who provided a valid response when asked about their social class, what would be the probability of randomly selecting a middle-class person?

Going by the formula we've used so far, we have:

$$p(\text{middle class}) = \frac{\text{middle class } N}{\text{total } N} = \frac{12230}{19161} = 0.638$$

Or, the probability of randomly selecting a middle-class respondent from this group of people is 63.8 percent<sup>1</sup>, exactly as the *Valid Percent* column tells us.

1. In Chapter 6 we will see that this is also the probability of a randomly selected *Canadian* (out of all Canadians) to be middle class, and why that is. This of course applies to all the calculations below.



And what would be the probability of randomly selecting either an upper-class or an upper-middle-class person?

$$p(\text{upper class or upper-middle class}) = \frac{\text{upper class N}}{\text{total N}} + \frac{\text{upper-middle class N}}{\text{total N}} = \\ = \frac{233}{19161} + \frac{3321}{19161} = \frac{3554}{19161} = 0.185$$

Or, the probability of randomly selecting an upper-class or an upper-middle-class respondent is 18.5 percent, as we can well see in the Cumulative Percent column.

Finally, what would be the probability of randomly selecting (with replacement) *first* a respondent who reported being lower class *and then* a respondent who reported being upper class?

$$p(\text{lower class and upper class}) = \frac{\text{lower class N}}{\text{total N}} \times \frac{\text{upper class N}}{\text{total N}} = \\ = \frac{628}{19161} \times \frac{233}{19161} = 0.033 \times 0.012 = 0.0004$$

Or, the probability of first selecting a person who reported being lower class and then a person who reported being upper class is a minuscule 0.004 percent. (A quick-and-dirty multiplication of the valid percentages of two groups, 1.2 percent and 3.3 percent, will give you the same result.)

See, it works! Now try it on your own.

### Do It! 5.7 Marital Status (GSS 2016)

Look at Table 5.2 and answer the questions listed below.

*Table 5.2 Respondent's Marital Status (GSS 2016)*

Marital status of the respondent					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Married	9426	48.1	48.1	48.1
	Living common-law	1791	9.1	9.1	57.2
	Widowed	1775	9.1	9.1	66.3
	Separated	635	3.2	3.2	69.5
	Divorced	1666	8.5	8.5	78.0
	Single, never married	4316	22.0	22.0	100.0
	Total	19609	100.0	100.0	

- What is the probability of randomly selecting a person (out of the 19,609 people) who is living common-law?
- What is the probability of randomly selecting a person (out of the 19,609 people) who is *either* separated *or* divorced?
- What is the probability of *first* randomly selecting a person (out of the 19,609 people, with replacement) who is married *and then* one who is single?

(Answer: 0.091; 0.117; 0.106)

In passing, we can also extrapolate that since percentages and proportions are relative frequencies, and probabilities are proportions and percentages, **probability is relative frequency** too.



---

## 5.2.4 The Real Normal Distribution Is a Probability One

Now back to the normal distribution, as promised.

Recall, if you will, the distinction between discrete and continuous variables<sup>1</sup>. Flipping coins and throwing dice and selecting respondents from a small number of categories are all discrete outcomes, so their probability distributions are also discrete.

On the other hand, continuous variables (i.e., mostly interval/ratio variables) have continuous probability distributions. **The normal distribution** — whose features we discussed at length — is **one type of a continuous probability distribution**.

As well, recall that probabilities are expectations. Thus, while some continuous random variables might have an approximately normal *observed* distribution, their *probability* distribution (i.e., expected in theory) is perfectly normal — because it's theoretical.

I said it before and it bears repeating: just like a few coin flips can produce an unequal number of heads and tails despite the fact that the probabilities of getting heads or tails are both equal to 0.5 *in theory*, a variable can have

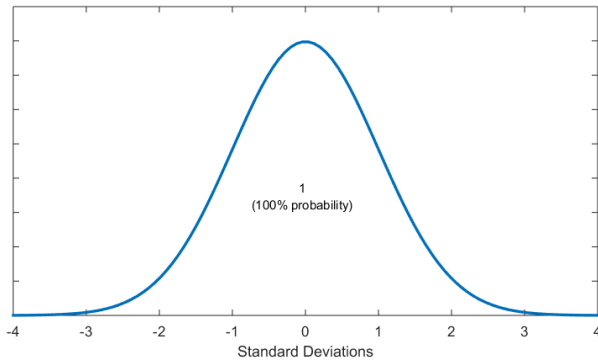
1. We discussed this in Section 1.5, here: <https://pressbooks.bccampus.ca/simplestats/chapter/1-5-discrete-and-continuous-variables/>.

an approximately normal frequency distribution while its probability distribution is theoretically normal. In short, we can *expect* some continuous variables to be normally distributed. For example, we can *expect* most people to be of average height or thereabouts, and to have few people who are much shorter or much taller, and the shortest and the tallest to be so rare as to be exceptional.

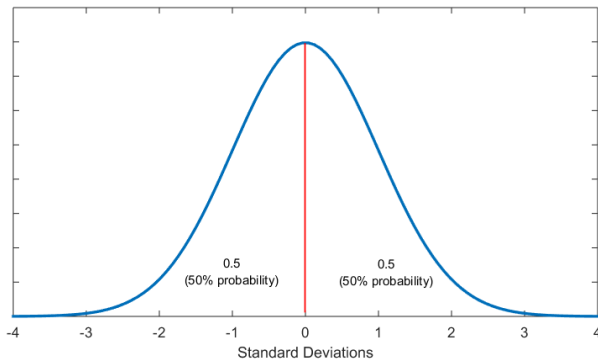
This, however, is actually *not* why the normal distribution is so important in statistics. What do we care about “some variables” and whether their distribution is normal or only approximately so? (Well, we do use that information, of course, but that’s not the point here.) **The reason the normal distribution is so valuable is because one specific very special distribution is normal — the sampling distribution**, as we will see in Chapter 6. (The sampling distribution lies at the basis of statistical inference.) But let’s not get ahead of ourselves.

After all this, you can see the normal distribution as a *normally distributed probability*. (Or, instead of a frequency distribution, it is a *relative frequency* distribution). Thus, the area under the normal curve is equal to 1 (or 100 percent, the whole probability), and it can be sectioned off, as it were, to indicate various outcomes’ probabilities. See the following set of Figures 5.6.

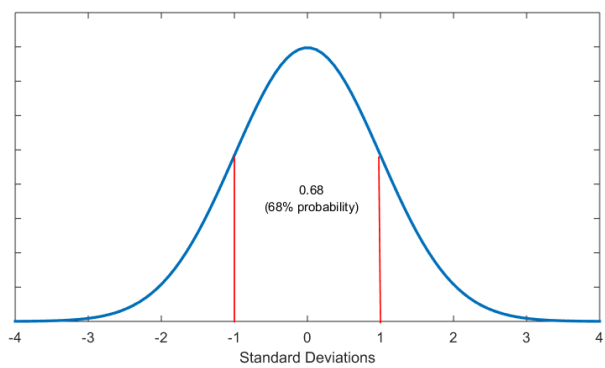
*Figure 5.6 (A) Probability of 1 (100%)*



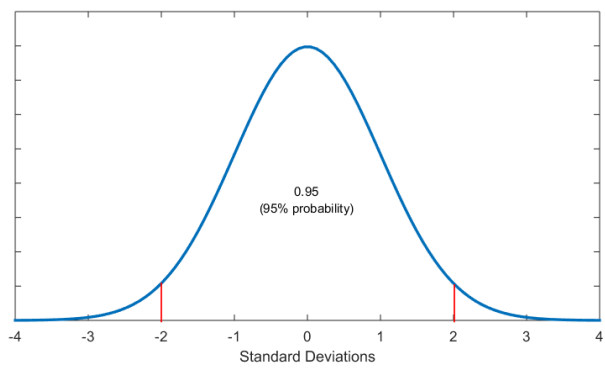
*Figure 5.6 (B) The Mean Gives Us Two Identical (Symmetric) Parts of 50% Probability Each*



*Figure 5.6 (C) 1 Standard Deviation from the Mean Sections Off 68% Probability*

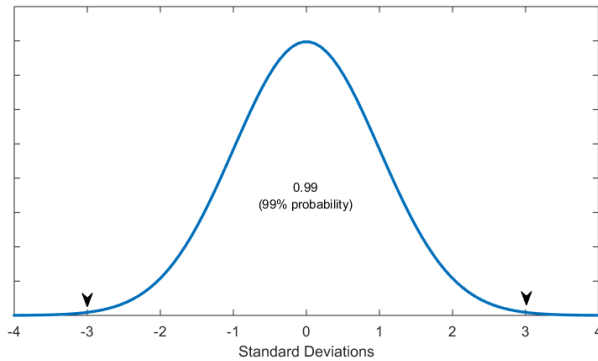


*Figure 5.6 (D) About 2 Standard Deviations from the Mean Section Off 95% Probability*



*Figure 5.6 (E) About 3 Standard Deviations from the Mean Section Off 99% Probability*





Thus, apart from what percentage of cases falls where, now we can discuss what the *probability* that a case will fall in a particular place is. Both refer to the same thing essentially but the latter indicates the *theoretical expectation* and allows us to be more precise (as empirically cases are only approximately normally distributed). Or, you can think of it like this: given the properties of the normal probability distribution, we can *expect* that much percentage of the data to be within that many standard deviations from the mean.

You'll see how the normal curve allows us to calculate probabilities through z-values in the next and (to your eternal relief) final section on the topic.



---

## 5.2.5 The Real Use of z-Values

Recall from Section 5.1.2 (here: <https://pressbooks.bccampus.ca/simplestats/chapter/5-1-2-the-z-value/>) that any value/score can be converted into a z-value, which tells us how far the value is from the mean in terms of standard deviations. Now that we know the normal curve has a bell shape reflecting probabilities (the higher the curve at any point, the bigger the probability), **any point on the horizontal axis can be seen as a z-value associated with a specific probability** — or rather, the probability below and the probability above the z-value.

You can find the z-values' probabilities listed in a Normal Distribution Table, e.g., this one: <https://www.mathsisfun.com/data/standard-normal-distribution-table.html>. Note that since the normal distribution is symmetric (i.e., the left side, below the mean, is exactly the same as the right side, above the mean), such tables usually only list probabilities between the mean and the z-score and above the z-score. This needs to be taken into account when calculating probabilities.<sup>1</sup>

Alternatively, online normal distribution calculators like this one [http://onlinestatbook.com/2/calculators/normal\\_dist.html](http://onlinestatbook.com/2/calculators/normal_dist.html) give you the option to specify which

1. To make sense of that, the linked webpage also provides an interactive tool to see all z-values with the normal curve with three options: between the mean and z, above z, and below z.

probability you need calculated based on a specific mean and standard deviations.

Let's take an example to see how this works.

#### *Example 5.4 Hockey Player Heights*

According to Hockey Graphs (REFERENCE <https://hockey-graphs.com/2015/02/19/nhl-player-size-from-1917-18-to-2014-15-a-brief-look/>), the average height of players in the National Hockey League is about 185 cm, with a standard deviation of about 5.3 cm<sup>2</sup>.

What is the probability that a new recruit (to your team of choice) will be taller than 185 cm? (Suspend disbelief and assume the recruit is randomly selected; i.e., his height (or skill) has absolutely no bearing on his selection.)

This one is easy: 185 cm is the mean, so the probability of a particular height being above the mean is 50 percent (equal to the probability of a height being below the mean). (For a visual, refer to Fig. 5.6 (B) in the previous section.)

So let's complicate matters further: What is the probability of the new recruit being taller than 198 cm?

To find it, we first need to convert the value into a z-score:

$$z = \frac{x_i - \mu}{\sigma} = \frac{198 - 185}{5.3} = \frac{13}{5.3} = 2.45$$

where of course  $x_i$  is the original value,  $\mu$  is the mean, and  $\sigma$  is the standard deviation.

Then, using a normal distribution table (e.g., the one linked above, <https://www.mathsisfun.com/data/standard-normal-distribution-table.html><sup>3</sup>), we find that the probability for a height to be above  $z=2.45$  (i.e., above 198 cm) corresponds to 0.71 percent, or less than 1 percent. (Of course, if you're curious, you'll also know that the probability of a new recruit to be shorter than 198 cm is  $(100-0.71=)$  99.29 percent.)

You can see the correspondence between the two graphs below in Fig. 5.7, one showing the height values and the other the z-scores. The area in which we are interested is beyond/above 198 cm, i.e., beyond/above  $z=2.45$ .

*Figure 5.7 (A) The Area Beyond 198 cm*

3. Or its applet, set to "z onwards".

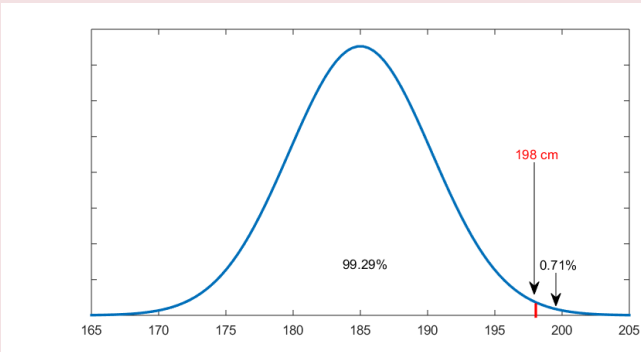
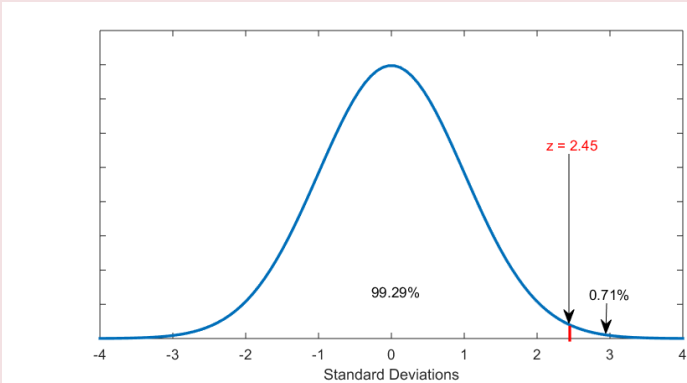


Figure 5.7 (B) The Area Beyond  $z = 2.45$

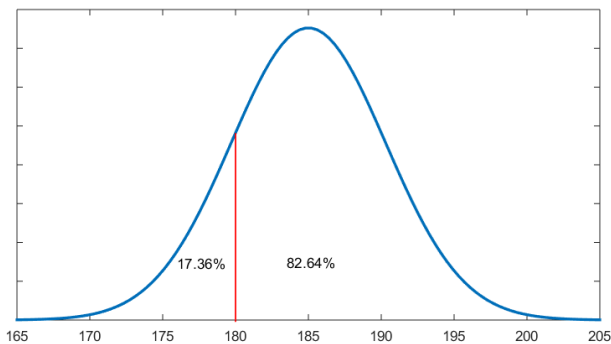


We can also ask the probability of a team recruit being shorter than 180 cm. Then:

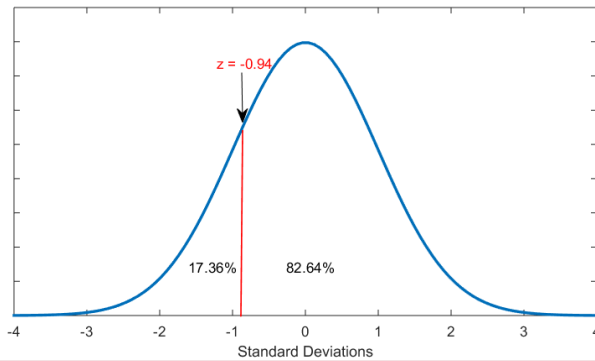
$$z = \frac{x_i - \mu}{\sigma} = \frac{180 - 185}{5.3} = \frac{-5}{5.3} = -0.94$$

Checking the normal distribution table, we find that the probability up to/below  $z = -0.94$  is 17.36 percent. Thus we have found that the probability of a recruit to be shorter than 180 cm is 17.36 percent. (Alternatively, we also know that the probability of a recruit being taller than 180 cm is  $(100 - 17.36) = 82.64$  percent.) Again, see the graphs in Fig. 5.8 below.

*Figure 5.8 (A) The Area Up To 180 cm*



*Figure 5.8 (B) The Area Up To  $z = -0.94$*



Finally, let's try finding the probability of a new recruit being between 178 cm and 188 cm. In this case we need to find two z-scores, and add the probabilities between each of the z-scores and the mean (i.e., above the lower score up to the mean, and below the higher score down to the mean).

$$z = \frac{x_i - \mu}{\sigma} = \frac{178 - 185}{5.3} = \frac{-7}{5.3} = -1.32$$

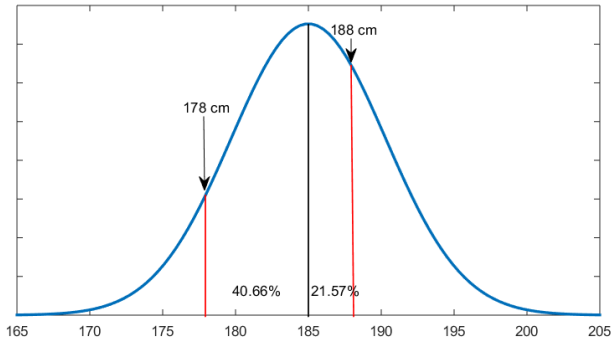
$$z = \frac{x_i - \mu}{\sigma} = \frac{188 - 185}{5.3} = \frac{3}{5.3} = 0.57$$

Using a normal distribution table we find that the probability between  $z = -1.32$  and the mean is 40.66 percent. The probability between the mean and  $z = 0.57$  is 21.57 percent. Thus, the probability that a new recruit's height will be between 178 cm and 188 cm is  $(40.66 + 21.57) = 62.23$  percent. See Fig. 5.9 below.

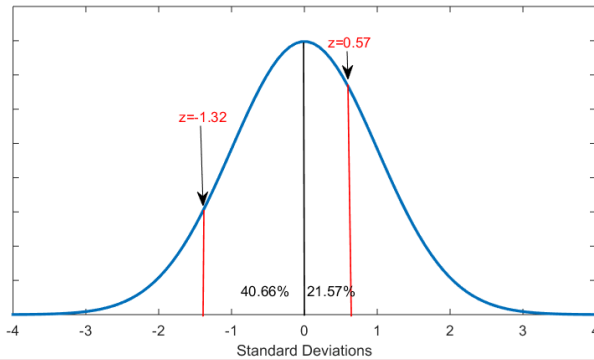
*Figure 5.9 (A) The Area Between 178 cm and 188 cm (Or*



*Rather Between 178 cm and 185 cm and Between 185 cm and 188 cm)*



*Figure 5.9 (B) The Area Between  $z = -1.32$  and  $z = 0.57$  (Or Rather Between  $z = -1.32$  and 0 and Between 0 and  $z = 0.57$ )*



Time to practice on your own!

*Do It! 5.8 Test Scores*

Imagine you learn that the average score on some test you've taken is 110 with a standard deviation 8. You still don't know your score, so you'll try to estimate some probabilities. What is the probability that you have more than 130? What about more than 95? Below 87? Between 90 and 115? Feel free to use the normal distribution table linked above. (Hint: Drawing out the normal curve centered on 110 helps.)

(Answers:  $z=2.5$ , 0.62%;  $z=-1.88$ , 96.99%;  $z=-2.88$ , 0.2%;  $z=-2.5$  and  $z=0.63$ ,  $49.38\% + 23.57\% = 72.95\%$ )

Now, with the concepts of probabilities and the normal distribution under your belt, you are finally ready to delve into statistical inference. Unfortunately for you, another theoretical chapter looms on the horizon, next. Grit your teeth and bear it, for the payoff (once we get to actually applying the theory in practice) is well worth it.

---

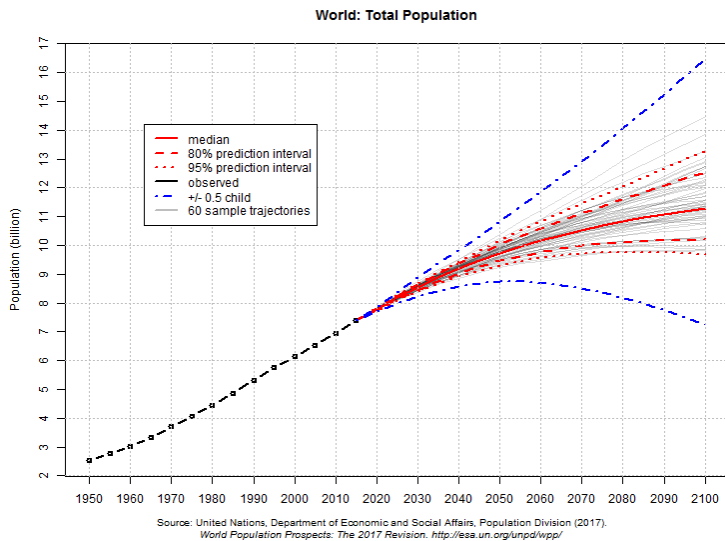
## Chapter 6 Sampling, the Basis of Inference

While describing variables is all nice and good — and useful — statistics would be rather limited if we only used it for that. In reality, descriptive statistics, while popular (consider sports statistics, for example), is only a relatively tiny part of all that statistics has to offer. The true power of statistics lies in granting us a superpower: the ability to *infer* — to know (and even to predict), within reason, things we cannot otherwise possibly know through observation alone. This part of statistics is called *inferential statistics*, and it's based on probability theory, a branch of mathematics of which you had a small taste in Chapter 5.

How do we know that life expectancy at birth is 82.3 years in Canada and 78.7 years in the United States but only 51.8 years in Sierra Leone (REFERENCE World Bank, 2016)? How can we predict, with reasonable certainty, the outcome of elections? How can we predict how many people will die of a particular cause in a specific country in a year? How do we know if most Canadians approve of immigration? Or what percentage of the Canadian work force is employed part-time? How do we predict how many people will be added to the

world population in any year, or how many people will the world have in 2100?

*Figure 6.1 World Population Projection 2100*



[<https://population.un.org/wpp/Graphs/Probabilistic/POP/TOT/>]

Fig. 6.1 above might seem complicated to you now, but soon enough you would be able to read it, as we will be covering all the concepts listed in the legend.<sup>1</sup>

While I'll leave the demography examples and projections about the future aside (as the scope of this text is quite more modest), let's take an example from closer to home and, say, talk about the attitudes to immigration

1. As it's somewhat difficult to see it on the graph, the answer to the last question -- what is the projected population of the world for 2100? -- is 11.2 bln. people (REFERENCE UN Population Division, 2017). We can do all that, and more, courtesy of inferential statistics.

in Canada. How do we know if Canadians approve of immigration? What do we mean when we even say “Canadians”? If we say “Canadians approve of immigration,” does that mean *all* Canadians do? If not, how many Canadians approve and how many disapprove?

To answer these questions we need to introduce more vocabulary than we have been using so far; vocabulary that is generally used in all sorts of research, both quantitative and qualitative, and not pertaining to statistics *per se*, though very relevant to it. In short, we have to start differentiating between a *sample* and a *population* (a term that has a more general meaning than the way we use it in everyday life), and we need to talk about *sampling*.

Following that, I’ll explain the concept of *randomness* in greater detail, which, coupled with what you now know about probability, will help us get to the *sampling distribution*. With that and the *Central Limit Theorem*, we’ll be ready. Then, and only then, we’ll be able to answer questions like *How do we know if Canadians approve of immigration?* along with any other question we might have about things/entities about which we cannot directly obtain information.

But I am getting too far ahead and too fast in my overview which, as any abstract talk, easily gets confusing. Let’s take it slowly from the beginning: samples and populations in the next section, and build from there. Be forewarned, however: what follows is indeed quite a bit theoretical and abstract, I’m afraid. (Yes, more than the last chapter, sorry.) Believe me, I wouldn’t do this to you if it weren’t necessary.



---

## 6.1 Populations and Samples

Before we start, yet another word of warning: what follows is only a brief overview of the topic of sampling and types of sampling. What I offer is enough in terms of a necessary background to statistical inference — but the main learning objective here *is* inference, *not* everything there is to know about sampling methods and their intricacies. Thus, if this is the first time you encounter the concept, you would be better served to read a thorough introduction on sampling and the benefits and downsides of the different sampling methods in virtually any one of the research methods textbooks you can find as that would give a more comprehensive treatment that I do here.

With that in mind, onward to the preliminaries: populations and samples.

In the introduction to this chapter, I asked a question: *Do Canadians approve of immigration?* How, do you think, can we go about answering it?

Presumably, the simplest way to investigate this would be *to simply ask* — imagine we contacted everyone and, indeed, simply asked them whatever version of the question we have decided on (i.e., whichever way we have operationalized our variable, *attitudes to immigration*),

noting everyone's responses. Many governments, both historically and to this day, have employed and still employ this method for gathering information.

**When we gather information from everyone in whom we are interested, we are doing a *census*.** You probably know that the Government of Canada, through Statistics Canada, conducts a census of the Canadian population every five years. (You might have even filled the form yourself, if you are of age, or seen your parents do it, otherwise.) Then, can the government (or any researcher/agency for that matter) collect information about everything it might need or want through censuses, every time the information is required?

Theoretically, it's an option. In practice, no way: it would be prohibitively expensive. You might find the reason prosaic, but any research is limited by the availability of resources, money *and* time. Asking one additional question on a questionnaire to one additional person has costs, which add up quickly the more questions and the more people are included in the study. Thus, censuses of the population are enormous undertakings reserved for collecting only *really* important (typically demographic) information, and are usually quite limited in scope.<sup>12</sup>

Given that conducting censuses for everything

1. For more information on the Canadian census program see here:  
<https://www12.statcan.gc.ca/census-recensement/index-eng.cfm>
2. Censuses of the population are so expensive, some governments cannot afford to do them (or at least not regularly) and instead rely on survey data from samples. As well, in some places censuses can be fraught with controversies due to racial/ethnic and/or religious tensions, etc. and are therefore avoided. (REFERENCE Weeks 2015).



anyone (researches, governments, etc.) might want information on is generally impractical/unfeasible, what can be done when information about a population is needed?

Here is where statistics saves the day: with probability theory and inferential statistics, we can use the next best thing to a census — *random-sample surveys*! My job in this chapter will be to convince you that you don't need to do a census of the population you want to study as long as you have a well-selected sample.

You, undoubtedly, have taken a survey at some point in your life in one form or another: a survey for which you were selected/invited or you volunteered; which included other people but definitely not *everyone*. In other words, unless we are discussing a census, surveys typically are administered to *samples* (i.e., sub-groups) of the population. However, not all surveys are created equal: those that can “substitute” for the population, as it were, rely on the just-mentioned technique of *random sampling*.

But first off, let's establish what samples and populations really are. While it's intuitive to think of *population* as the population of a country (say, 36.7 mln. Canadians), and of *sample* as a sub-group of that population (say, ten thousand Canadians), this is only a special case of the general terms *sample* and *population*. **In research, a *population* is a group encompassing everyone on whom we want information, i.e. everyone (or everything) we want to study.** Considering that we might not be studying people (recall that the units of analysis can be countries, organizations, etc.), we say that a

**population encompasses all elements under study.** This means that we could have study populations such as “countries in South America”, or “hospitals and medical clinics in Toronto”, or “departments of sociology in Canadian universities”, etc.

As well, while the elements may be people, instead of the whole population of a country, we might be interested in studying “university students in Canada,” or “early childhood educators in British Columbia,” or “dog walkers in downtown Vancouver,” or “Telus company employees,” or “dentists in Surrey, BC,” etc. All of these examples are of populations that can be defined as such by researchers interested in them.

Thus, **a sample is any sub-group of the population under study.** For example, if I decide to study “KPU students”, my study population would be defined as “everyone registered as a student at KPU”. If I select a hundred students for my study, I would have a sample of  $N=100$ .

Ultimately, again, **what the population for a particular study is depends on what the researcher wants to study.**

If we go back to the *Do Canadians approve of immigration?* example, the population under study would be, of course, “Canadians” but we have to be very careful how we define “Canadians”: Are we interested in *all* Canadians, regardless of where they live/are at the moment? (I.e., do we include ex-pats, people with dual citizenship residing abroad, Canadian tourists travelling the world, etc.?) Or do we only want to study Canadians *in Canada*? And do we want to study permanent residents

in Canada too or only people with Canadian passports? Regardless of how we want to define our study population, it has to be precise and to have objective criteria that we follow consistently.

Once a researcher has decided on and defined a study population, and collecting data on all elements of that population is considered unfeasible (and, as you will eventually see, collecting data on all elements of the population might be even undesirable as its unnecessary, even if it were feasible), the researcher needs to select a sample for their study.

**The procedure of selecting a sample is called *sampling*. There are two broad types of sampling, *non-random* and *random*, and the next section is devoted to that.**



---

## 6.2. Non-random Sampling

How do we go about selecting elements (be they individuals, organizations, etc.) for a study, once we have decided on a population? In short, how do we go about sampling?

You know by now (if only because of the title of this section) that the two broad types of sampling are non-random and random. **Statistics (specifically, inferential statistics) is based on random sampling**, therefore in what follows I disproportionately focus on that. This is not because non-random sampling is not used or isn't useful — not at all! **Non-random sampling comprises several very much valid and valuable sampling techniques, typically used in qualitative studies.** However, these are situated outside the scope of this book. As such I will do an only passing overview of non-random sampling (so that you are able to spot it and differentiate it from random sampling).<sup>1</sup>

With that in mind, I start my lopsided mini-presentation on the topic; non-random sampling first and random-sampling in the next section.

1. You would be doing yourself a favour to learn about all research (and sampling) methods available. After all, not every research question can be approached and studied from a quantitative perspective. (And, at the very least, there are study populations that can only be sampled non-randomly.) I thus very much encourage you, if you haven't already, to take an introductory course in research methods to learn all there is to learn about sampling, both non-random and random.

Professors in social science classes sometimes ask students to interview or administer surveys as part of class assignments. You might have had to do that, or you can just imagine such an assignment — so how did/would you select your subjects? Most likely you would go with what's most convenient — fellow students in your class, students that happen to be in, say, the cafeteria when you had time to do the assignment, your closest relatives or friends if you were instructed to choose non-fellow students. Any of these ways of sampling are generally classified as non-random (a.k.a non-probability) sampling.

**Non-random sampling techniques typically include** *convenience sampling* (selecting whichever elements are closest/most convenient to you), *purposive sampling* (sampling with a purpose: selecting only the most useful (e.g., most knowledgeable/ rich in information) cases as judged by the researcher, also called *judgment*, *selective*, or *subjective sampling*), *snowball sampling* (where selected few initial participants contact/invite/recruit others in their respective circles to become participants in the research), and *quota sampling* (sampling on a specific desired characteristic, e.g., specifically selecting a certain number of men and a certain number of women for a study).

As well, **any time the subjects of a study are self-selected (i.e., the study is based on people volunteering to participate), it is also considered non-random sampling.**

The one defining feature common to all non-random sampling methods is related to the probability of elements to be selected/included in the study. **If the probability of**

**the elements of the population to be included in the study is unequal — i.e., if some elements have higher probability to be in the study than others — the sampling is called *non-random*.** Non-random samples are in this sense *biased* — they focus, and select information, on some elements more than others.

The information about these specific elements might be very useful but it reflects *only* the elements from which it was collected. In other words, **such information (and studies based on it) is said to have *limited generalizability*.** To the extent that there is a claim to generalizability, the generalizability is *assumed* (perhaps by assuming the population is so uniform that any subgroup would reflect it).

A word of caution, however: The limited generalizability of non-random sampling techniques should never be taken as somehow detracting from, or invalidating, research who legitimately uses them. To take a prime example, ethnographies usually rely on non-random sampling methods, yet they typically provide a wealth of information and levels of detail that could never be achieved through a quantitative survey research alone. Thus, non-random sampling techniques should never be considered as inferior to random ones — just different, and serving different purposes.

**The purpose of random sampling, then, is to find a way for a sample to truthfully reflect — i.e., to stand in for — the population from which it is taken.** This truthful reflection — i.e., generalizability — is no longer assumed (as it is in non-random sampling), but rather it is verifiably

proven through mathematical means based on probability theory.



---

## 6.3 Random Sampling

In order to be able to use what we know about probability distributions and the normal curve and to be able to apply this knowledge in the service of inference (how exactly we do that comes later in the chapter), we need to know the probabilities of the population elements to be selected. The problem is, estimating these probabilities (every time, for each and any new study) can be way too burdensome, if not outright impossible. Consider the following example.

### *Example 6.1 Mode of Transportation of Students*

Imagine that you are interested in what mode of transportation the students in your university usually take to campus. You decide that a sample of  $N=100$  sounds reasonable. Imagine further that you don't know anything about sampling (or logic) so you decide to go to the nearest bus stop to your campus and talk to the first hundred students that happen to come by once you're there.

Arguably, if you did that, you could expect close to 100 percent of your sample to choose *bus* as their usual mode of transportation to school — after all, you have talked only to students waiting at a *bus*

*stop*. True, it's possible that some of your respondents were taking the bus only at that particular time (their car might have broken down, or they didn't feel like driving that day, etc.) but it's hardly likely this to be the case for more than a few out of the selected hundred.

So far, what you could learn from your study is that some hundred (or close to it) students in your university happen to usually take the bus to school. In and of itself, there is nothing wrong with that. The question, however, is whether you can use this information to conclude that *bus* is the usual form of transportation for students in your university *in general*. To paraphrase in the language of research: is the information regarding usual modes of transportation gathered by you from a hundred students at a bus stop generalizable to your institution's student body as a whole?

Even going by logic alone you should be able to easily see that the answer is *no, of course not*. After all, you only talked to students at a bus stop who were there specifically to take the bus, at a specific time, on a specific day. What about the students that directly went to the parking lot to take their cars, or those who went to retrieve their bikes from the bike racks, or who simply walked home? Then what about students who had no classes on the day that you went to the bus stop? Or the students that were in class at the time you were interviewing your subjects? Or the students in your institution whose classes were at a different campus and never came to the one you happened to be in?

In short, your method of collecting information had produced a *biased sample*: some elements in it (students who happened to be taking the bus at the time of your survey) had a higher chance at participating in your study than others (everyone else). The sample is biased toward bus-takers — those who you talked to had something like 100 percent chance to be in the study (and they did); other bus-takers who weren't there had a smaller but still potential chance to be in the study, and those who never take the bus had 0 percent probability to be in your study.

What's more, not only are the probabilities to be selected different for the different students, calculating the exact probability for every element in every new study and accounting for the differences would be a fool's errand, as unfeasible (or outright impossible) as collecting information on the entire population under study in the general case.

The takeaway from Example 6.1 is that in statistics we want elements to have easily known (to make calculations easy) and equal (so as to not produce bias) probabilities to be selected. Fortunately for us, random sampling (also called *probability sampling*) provides both — as the way for the probabilities to be known is based on the fact that when chosen at random, the elements have the same/equal probability of being chosen.

Recall that in a coin flip the probability of getting heads is the same as the probability of getting tails, and they are both  $\frac{1}{2}$ , one outcome out of two possible outcomes, or 0.5. The probabilities of throwing a die and getting a one, or a two, or a three, or a four, or a five, or a six are all equal,

and known:  $\frac{1}{6}$ , one outcome out of six possible outcomes, or 0.167. Similarly, the probability of selecting one person at random out of a group of thirty-five people is the same for all thirty-five people, and equal to  $\frac{1}{35}$ , or 0.028.

**Throwing dice, flipping coins, and selecting at random are all random (chance) events – there is no bias in them, as the probability of any outcome is the same as any other outcome, and easily calculatable as one out of the total number of possible outcomes.**

If we apply the same logic to sampling, we can see that the only thing we need is to make sure that our selection is random and that it applies to all elements in a population of a particular known size: then the probability of selecting an element will be always one out of the total number of elements, i.e., the total study population size.

When this condition — equal probability of elements to be selected — is met and we know that probability, we know its frequency distribution. We can thus use probability theory and its theorems and postulates which provide mathematical proof that a random (i.e., unbiased) sample reflects and *represents* the population from which it was drawn truthfully. Then and only then, whatever we learn from the sample would be generalizable to the population. (Of course, it's not *that* simple; there is more to it — like sample size — but I'll leave this for later when we get to the Central Limit Theorem).

So what would have been the best way to get a

representative answer to the question regarding usual modes of transportation for students in your institution? Theoretically, you could have obtained a list of all students from the registrar, selected your hundred at random from the full list, and contacted only the persons selected. Their responses would indicate the most popular mode of student transportation and *now*, with random sampling, *they would reflect the entire student's body*.

In practice things are more complicated: How *exactly* do you chose at random any desired number of elements from a list of all elements?<sup>1</sup> How do you even obtain a list of all elements in the first place? Even if we had one, do we put every element's name/number in a hat and pull them out one by one?

While providing details on how random sampling is done in real life is also outside the scope of this text, I can assure you several such methods exist (though pulling names out of hat isn't one of them). For a comprehensive treatment, again, I encourage you to consult a research methods textbook; for my purposes here I will just list the major ones.

*Simple random sampling* is the closest that you can get to the pulling-names-out-of-a-hat proposition, however, in this day and age it is usually done with computers using

1. The comprehensive list of all elements in a population is called a *sampling frame*. Note that in practice some sampling frames might not include all elements they purport to have. For example, using the phone book as a sampling frame for a population is a frequently used method, yet we know that some people have unlisted numbers -- or, possibly, do not have a phone -- so they are not listed in the phone book. Thus there is a difference between the population and the sampling frame for it, where the sampling frame is an approximation of, but not quite a list of the entire population.

random number generators. The same goes for *systematic random sampling* (when the selection starts at a random starting point and proceeds at a fixed interval). Then there are also *stratified random sampling* (the population is first divided into *strata* based on similar characteristics of the elements, not unlike in quota sampling but then the selection from each strata is random), and *cluster random sampling* (the population is divided into clusters — think sub-groups — and then clusters are selected at random), where the latter can be even done in several stages (called *multistage cluster random sampling*).

To conclude: ultimately, the important thing to learn here is not how the sampling is done empirically but the key difference between non-random and random sampling. **Non-random/non-probability sampling methods select elements arbitrarily at researchers' discretion, with unknown and unequal probabilities of elements to be selected; this, in turn, precludes the use of probability theory and therefore allows for only assumed (but unprovable) generalizability of the samples produced in this way.**

On the other hand, random/probability sampling methods, in selecting elements at random, ensure that elements have equal (and therefore known) probability to be chosen; this **random selection allows for the use of probability theory, the normal curve, and everything that is already mathematically proven regarding features of random variables and their probability distributions. Probability theory demonstrates that randomly selected samples (of sufficient size) are representative of and generalizable to the population from which they were drawn. Therefore, conducting**

**a census of all elements under study becomes unnecessary as long as we are able to draw a random sample (of sufficient size) of the population.**

At this point, (if you are still awake) you have probably noticed that I ask you to accept the fact that random samples are representative of their populations on my word, with little proof. While I will not go about proving this mathematically (and you'll be happier for it), I will provide the theorem on which my claims are based soon enough. First, however, we still have a few other things to cover, and the logic of inference is next.





---

## 6.4 Parameters, Statistics, and Estimators

The logic underlying statistical inference is that we want to know something about a population of interest but, since we cannot know it directly, what we do is study a subgroup of that population. Based on what we learn/know about the subgroup, we can then *estimate* (i.e., infer) things about the population. In the previous section, we already established that not any subgroup of the population will do — what we need is a *randomly* selected sample, created through one of the random sampling methods I listed (simple, systematic, stratified, and cluster). What we do is **collect data from/about elements of a *sample* (e.g., respondents) with the explicit goal of finding something and drawing conclusions about a *population*.** (Again, we can do that due to the fact that random sampling allows us to use probability theory through the normal curve.)

Saying we want to find “something” about the population of interest is hardly formal (much less precise) terminology but I wanted to get the message across before I introduced you to the proper statistics jargon. Let’s do that now.

Populations have *parameters* and samples have *statistics*. **We describe populations with their *parameters* while we describe samples with their *statistics*.** When we study something, we are interested in the parameters of the population, however, in most cases it is difficult

to collect the information to calculate them. What we do instead is **we take a random sample of the population and calculate the sample's statistics. We then use the sample statistics to estimate (i.e., infer) the population parameters.** Thus, sample statistics are also called *estimators* of population parameters.

For example, if we want to know the average age of Canadians, we could either do a census and ask everyone or simply take a nationally representative sample. Considering how expensive and time-consuming it would be to ask all 36.7 mln. Canadians (and Statistics Canada conducts the official census only every five years), we can poll a random selection of people across Canada, calculate their average age, and use *that* as an *estimate* of the average age of all Canadians<sup>1</sup>.

In this example, the average age calculated based on the people in the sample is the *statistic* which we use to *estimate* the average age of all Canadians, the population *parameter*. All measures of central tendency and dispersion describing variables based on sample data are statistics. On the other hand, if we have data from all the population when calculating measures of central tendency and dispersion, we would have parameters.

1. When people who have no statistics background learn of this, they usually protest that the information is not accurate because it's not based on *everyone*. What you will learn in this chapter is that you don't *need* everyone, and a sample is perfectly enough because random samples of sufficient size are mathematically proven to produce the best (closest, truest, most unbiased) estimates of the population parameters. To the extent that there is a difference between a statistic and the parameter it estimates, this difference is accounted for by reporting levels of certainty/confidence. More on that later.

Consider if you will, examples I have used in past chapters: whenever the example was based on actual data from a dataset, and SPSS was used, this was sample data producing statistics<sup>2</sup>. Even if we haven't used statistics in this way yet, they *can* be used to estimate things about Canadians as a whole. On the other hand, any time I have used examples using hypothetical (imaginary) data about “your friends”, “your classmates”, “hours you have worked per week”, etc. can be considered as having population data, as we imagine we have all the information about those things, and there's nothing to estimate.

A final note concerns formal notation. **To differentiate between statistics and parameters, we designate sample statistics by *Latin* letters but we denote population parameters by *Greek* letters.**

You have already seen a ready-made example for this rule: recall our discussion on variance and standard deviation. In Section 4.4 (<https://pressbooks.bccampus.ca/simplestats/chapter/4-4-standard-deviation/>) I introduced formulas for  $\sigma$  and  $\sigma^2$  and I mentioned (without much explanation) that another “version” of these exist as  $s$  and  $s^2$ . In truth, when we calculated the variance and the standard deviation with the hypothetical data in the examples, we needed the *population* standard deviation and variance (i.e.,  $\sigma$  and  $\sigma^2$ , respectively); but when we use SPSS with a dataset (i.e., sample data), we need the *sample* standard deviation and variance (i.e.,  $s$  and  $s^2$ , respectively). Here they are again:

2. All datasets used in this book are nationally representative data collected by Statistics Canada.

$$\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N} = \sigma^2 = \text{population variance}$$

$$\sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N}} = \sqrt{\sigma^2} = \sigma = \text{population standard deviation}$$

$$\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1} = s^2 = \text{sample variance}$$

$$\sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}} = \sqrt{s^2} = s = \text{sample standard deviation}$$

I'll take this opportunity to finally explain why we need the difference in the formulas (i.e., to divide by  $N-1$  in

the *sample* formulas but by  $N$  in the *population* formulas). Considering that the sample statistics *estimate* the population parameters but are arguably different from the exact parameters — i.e., some uncertainty exists, as inference is not a perfect “guess” — to assume what we obtain from a sample is exactly the same as the population would be a biased estimation. Thus, the  $N-1$  is meant to correct that bias<sup>3</sup> (which it does for the variance, and does to an extent for the standard deviation). **What we have then is that  $s$  and  $s^2$  are unbiased estimators of  $\sigma$  and  $\sigma^2$ , respectively.**

Thus it should be clear why we use the  $s$  and  $s^2$  formulas when working with datasets and SPSS — as the actual data has been collected from respondents randomly selected from a population of interest and comprising a sample of specific size. On the other hand, when we have data about everyone/everything we’re interested in (like in the small-scale examples with made-up data), we have a *de facto* population on our hands — hence the  $\sigma$  and  $\sigma^2$  formulas are appropriate. In the former case, the findings can be extrapolated to the population (acknowledging that we are dealing with inferred estimates); in the latter case, there is nothing further to extrapolate as we are calculating the parameters directly.

Another important parameter to note (as we will be using it a lot from now) on is the population mean designated by the small-case Greek letter for  $m$  (from *mean*) —  $\mu$ , which I introduced in Section 5.1.2

3. This is called *Bessel's correction*, by the name of Friedrich Bessel who introduced it.

(<https://pressbooks.bccampus.ca/simplestats/chapter/5-1-2-the-z-value/>) without giving you a reason why. Unlike the correspondence between  $s$  and  $\sigma$ , however, we don't usually denote the sample mean with an  $m$ ; as you know, we use  $\bar{x}$  instead (so that we know which variable's mean we have in mind).

Finally, when a parameter is being estimated by an estimator, it is designated by a “hat” on top: for example, if we have a sample statistic called  $a$  estimating a population parameter  $\alpha$ <sup>4</sup>, the estimated  $\alpha$  will be  $\hat{\alpha}$ , pronounced “alpha-hat”. By analogy, if a statistic  $b$  estimates a parameter  $\beta$ <sup>5</sup>, the estimated  $\beta$  will be  $\hat{\beta}$ , pronounced “beta-hat”.

Thus, the logic of inference tells us that while  $a = \hat{\alpha}$  and  $b = \hat{\beta}$  (i.e., the statistics are estimators for the parameters),  $a = \hat{\alpha} \neq \alpha$  and  $b = \hat{\beta} \neq \beta$ . **That is, the statistics (a.k.a. estimators) are not the same as the parameters.** More on this, next.

4. This is the small-case Greek letter  $\alpha$ :  $\alpha$ , pronounced “AL-pha”.

5. This is the small-case Greek letter  $\beta$ :  $\beta$ , pronounced “BAY-ta”.

---

## 6.5 The Sampling Distribution

With this section we reach a point where you will have to make a good use of your imagination and abstract thinking. Unlike our presentation and discussion of variables early on, giving real-life examples for this material becomes impossible as the sampling distribution lies firmly in the realms of abstract mathematical concepts. Yet we need it because it's the sampling distribution which makes inference possible and bridges the gap between a sample and the population from which it was taken.

Thus, as promised in my introduction to keep everything to its most necessary minimum to be understandable, below I offer as non-technical and non-mathematical explanation of what the sampling distribution is and how we use it as possible. However, this course of action has its obvious inevitable downsides: since we are skipping the actual mathematical proofs and going directly for the results of these, you will have to accept the presentation at my word. This is a hard thing to ask of anyone (*“it is what it is because I tell you so”*). My justification for doing it is because the vast majority of my students so far seem to find the alternative (*“it is what it is because of all this very long presentation of complex mathematical concepts and complicated procedures”*) even more unpalatable, without any gains in comprehensibility — and, as such, ultimately mostly useless. (Of course, if interested, you can always check other, more comprehensive books and online sources.)

Despite the dire warning about upcoming doom in the form of abstract concepts, I still begin with an example.

### *Example 6.2 Age of Classmates*

Imagine you are enrolled in a class along with 49 other students, so the total class size is 50. Let's say as a class assignment (perhaps in a research methods class) you are tasked with taking a sample of your class and administering a survey to your sample. In this sense, your class is your population of interest. For simplicity's sake, we focus on one possible question, say, *age of respondent*. You want to know the average age of the study population but, instead of asking all 50 of your classmates, you draw a random sample of them for the purposes of estimating the class's average age<sup>1</sup>.

Now despite that I still haven't said anything about sample size (but we're getting there), I'll assume that a sample size of 10 (i.e., 20 percent of the population) would sound reasonable enough to you. The random draw (with replacement) yields the following ten classmate's ages:

1. Of course, with a population of only 50 in real life you can just collect the information from everyone. I'm using a small-size example for teaching purposes only and to make calculations manageable. The principle of sampling applies equally not only to a population of 50 but of any size -- and when your study population's size is in the millions, you wouldn't attempt to survey all of them (barring the already discussed case for censuses).



19, 19, 20, 20, 20, 21, 21, 22, 23, 28

Based on these values, the average age of the sample,  $\bar{x}$ , is

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{(19)2 + (20)3 + (21)2 + 22 + 23 + 28}{10} = \frac{213}{10} = 21.3$$

I.e., your sample's average age is 21.3 years. Considering that these ten people were randomly drawn, and that they are, well, only *ten*, can we assume that the average age of your *entire* class of 50 is 21.3 years?

While this is a good — *educated* even — guess and a good starting point, **it is unlikely that, had you polled everyone in the class, your calculation would have produced exactly 21.3.** After all, polling 10 people is not the same as polling 50; in the latter case your calculation would include a lot more information than in the former. Thus, it's also reasonable to expect that there will be *some* difference between the mean based on the sample, *overlinex*, and the *true* population mean,  $\mu$ .

Then how about if you decided to draw another random sample of ten people out of your class? Would you expect to have the exact same mean of 21.3 years?

Unless you somehow end up with the exact same ten people who were in the first sample (and after Chapter 5 on probability you should know how minuscule that probability is), it is again unlikely you'd get the same mean.

We could imagine that the new, second sample's ages might look like this:

18, 19, 19, 19, 20, 20, 22, 22, 24, 25

Based on these ten new values, the average age of the second sample (let's call it  $\bar{x}_2$ ) is:

$$\bar{x}_2 = \frac{\sum_{i=1}^N x_i}{N} = \frac{18 + (19)3 + (20)2 + (22)2 + 24 + 25}{10} = \frac{208}{10} = 20.8$$

I.e., your *second* sample's average age is 20.8 years, despite it being drawn from the same population. Your two samples (of the same size) yielded two close — but still different — numbers.

As well, following the same logic, it's just as unlikely that the population mean  $\mu$  (your class's average age) is 20.8 years as it was unlikely that it's 21.3 years (the sample is still only 10 people).

What then? How can we trust a sample statistic to estimate a population parameter? It appears we need more information. Before we get to that, however, let's finally address the elephant in the room — the issue of *sample size* I have been neglecting so far.

One reason you might think the sample estimates in the Example 6.2 above differ (both from each other and from the true population mean) could be the sample size: isn't  $N=10$  just too small? The answer is of the *yes-but-no* variety: No, a sample size that's 20 percent of the population size is actually quite big for a research study of a typical, relatively large size. Yes, a sample of 10 out of population of 50 *is* way too small. And, in general, yes, the larger the sample the better. But let's unpack— and qualify — all of these three contradicting pieces of information properly.

Inferential statistics — at least the typical kind discussed in this textbook — is about estimating *relatively large* populations; luckily, quantitative social science research most commonly deals with such populations<sup>2</sup>

**The recommended sample size depends on the size of the population it will be used to estimate but at *diminishing returns*: the larger the population, the larger the sample's *absolute* size should generally be — but at the same time**

2. There is no magic number as to what constitutes a relatively large population, and therefore an adequate minimum requirement for a sample size. For the latter, I could offer 100; some suggest 30, others 50 but in truth all these are more or less arbitrary. It is a fact that having a larger sample (both in absolute and proportionate sense) puts you in a safer ground in terms of statistical inference (this has to do with probability theory, the law of large numbers, the sampling distribution, the normal curve, and the Central Limit Theorem discussed below for which to work, say, a minimum  $N=30$  is a frequently cited number). What you can take out of this is that it's better to avoid dealing with  $N<30$  (or even  $N<100$  if possible), as the tools and methods discussed in this textbook are better suited for larger sample (and population) size..

**the gains of the larger sample size diminish (to zero), the larger the population is.** In other words, **smaller populations need samples of bigger proportion to represent them correctly, while larger populations need samples of smaller (and smaller, and smaller) proportions to do so.** (This also means that even if you have larger and larger populations, there will be no gains in increasing the sample size beyond a certain point.)

In reality, no one would try *estimating* the parameters of a population as small as 50, as in most cases they can be easily obtained — not to mention that to have a meaningful estimate of a population that small, one would indeed need almost the entire sample. Sample size calculators are abundant and free online<sup>3</sup> but to give you an idea of the diminishing returns to increasing sample size, I'll just list a few. To estimate a population of 200, you'll typically need a sample of about 180<sup>4</sup>; to estimate a population of 500, you'll typically need a sample of about 380; to estimate a population of 1,000, a sample of 600 would be adequate; for a population of 2,000, a sample of about 870 would work; for a population of 5,000, a sample of 1,200 would be enough; for a population of 10,000, a sample of about 1,300 would be enough... then for a population of 50,000, a sample of about only 1,500 would suffice, and a

3. You can find one example at SurveyMoneky.com

([https://www.surveymonkey.com/mp/sample-size-calculator/?ut\\_source=help\\_center](https://www.surveymonkey.com/mp/sample-size-calculator/?ut_source=help_center)).

4. Here and on "typically" refers to a frequently used *margin of error* of  $\pm 2.5\%$ ; more on what this actually means in Section 6.6 below.

population of 100,000 would do just as well with the same number of 1,500<sup>5</sup>.

What it comes down to is that, to the surprise of many, actually **a sample size of “just” 1,500 respondents can safely and accurately estimate any population 25,000+.** This also means that a random sample of 1,500 people can statistically represent, for example, both the population of Toronto (2.7+ mln. people) *and* the population of Canada (36.7 mln. people) — however, it cannot be the *same* sample (the former needs to be drawn of Torontonians only, the latter of all Canadians<sup>6</sup>).

**Watch Out!! #11... for (Mis)Judging a Study On Its Sample Size**

The point against judging a study on its sample size alone should be clear already but it bears repeating. When people unfamiliar with statistics encounter social-scientific reports based on studies of what they consider a “too small” sample size, they tend to dismiss the findings. They tend to consider the “only 500 respondents” or “only 1000 cases” too few to accurately represent the population from which they were

5. You can also find a table summarizing sample size like this one useful:  
<https://www.research-advisors.com/tools/SampleSize.htm>.
6. In truth, researchers do want larger samples to represent Canada (or other countries' populations) but that's only to increase the *power* (defined later) of their statistical findings, not their generalizability. This desire for larger *N* is, of course, constrained by limited resources (time, money, etc.).

drawn, especially if the population is, in their view, disproportionately large.

As you should have learned by now, the generalizability of a study is more a matter of *how* the sample is drawn, not of its size (beyond a certain point). As long as the chosen sampling method is a type of random sampling, and the sample size is adequate for the population size<sup>7</sup>, the results of the study will be generalizable to the population. The actual sample size doesn't matter for that, even if it may look "too small" to some.

In any event, even if it's from a certain point on *unnecessary* as demonstrated above, as a logical inevitability, the closer the sample is in size to the population from which it is drawn, the smaller the difference between statistics and parameters should be. Even in the Example 6.2 above with its imagined, only-for-illustration-purposes population of 50, getting information from 40 of your classmates instead of the 10 we used in the example should get us an average age that is closer to the true population age (of all 50 students).

However, as a corollary, unless we obtain information from truly everyone (i.e., we do a census), **in random sampling a difference between the sample statistic and the population parameter will always exist<sup>8</sup>. This difference**

7. At the desired -- and reported -- margin of error.

8. Well, *almost* always: it is possible (though very unlikely) that a

**between the estimate (the statistic) and what is being estimated (the parameter) is called *random error*.** Random error is *inevitable* — no matter what we do, a sample will always only produce an estimate, never the “real thing”, as it were.

Going back to Example 6.2 above, we can extrapolate that when randomly drawing a sample after sample after sample (of the same size) an infinite number of times, and calculating a mean after a mean after a mean, we’ll get a long (well, *infinite*) number of means which will all be somewhat close to, but not exactly equal to, the true population mean. If you could possibly imagine this very long (infinite<sup>9</sup>) list of means as similar to a variable with a large number of observations, please do so, it helps.

This variable you imagined (made of the very large number of means that would be produced by the very large number of samples if we took them) will have a frequency distribution just like any real variable we have discussed so far. **The distribution of the variable made of the means is called the *sampling distribution of the mean*<sup>10</sup>.** However, since all this is *theoretical* (we do not take more than one sample), this distribution is not really about actual

sample will just so happen to produce the true population parameter. This will also be a result of random chance, as unlikely as it may be.

9. For ease of imagination, I’ll stick to “very long/large” from now on, but at the far back of your mind, remember it’s actually infinite.
10. I provide this definition only to make understanding the sampling distribution easier. It’s in no way the technical definition of the sampling distribution. As well, keep in mind that this “variable” made of the means is a perfectly imaginary heuristic device.

frequencies but rather about expected/relative frequencies, i.e. probabilities. As such, **the sampling distribution is a *probability distribution* — it lists a mean's *probability of occurring***<sup>11</sup>.

In a more precise phrasing, all statistics based on samples (e.g., means, medians, deviations, etc. plus many others we haven't yet encountered) have a sampling distribution, which refers to their theoretical<sup>12</sup> variability over repeated (to infinity) random samples of specific (and equal) size. What we know about the sampling distribution of sample statistics is summarized in the Central Limit Theorem, next.

11. Compare this to flipping a coin, or throwing a die: as we saw, in both cases the distribution of the *possible* outcomes (over infinite number of flips/throws) is a calculated and known probability distribution. After all, that's why we know that the probability of getting tails or heads is 0.5 *in theory*, just like it's 0.167 for throwing any of the die's six numbers *in theory* (even if calculating actual flipped/thrown frequencies in real life yields different results).
12. It is theoretical because we do not actually take multiple, much less infinite, number of samples as there is no need: courtesy of probability theory and the Central Limit Theorem, we just *know* what *would* happen if we did.



---

## 6.6 The Central Limit Theorem

Despite its scary-sounding name, the *Central Limit Theorem* (CLT) simply *describes* the sampling distribution — and simultaneously explains why, and how, we can use sample statistics (like the mean of a variable,  $\bar{x}$ , obtained through sample data) to estimate population parameters (like the true population mean of that variable,  $\mu$ ).

Recall what we use to describe a variable's frequency distribution: 1) a graph to visually display the distribution's shape; 2) measures of central tendency; and 3) measures of dispersion. In the previous section I also asked you to imagine the (entirely theoretical, i.e., *probability*) distribution of the mean (again, in theory, over infinitely repeated samples). What the CLT does then is provide information about all three of these elements (shape, central tendency, dispersion) but about the distribution of mean. **In short, the CLT describes the sampling distribution of the mean.**

The sample size plays an important role: the CLT applies to “large  $N$ ”, and is stated for “as the sample size grows”, bringing us back to the point that the larger the  $N$ , the better for inference it is.

Specifically, the CLT states that with random sampling, as  $N$  increases (i.e., for large  $N$ ), the shape, central tendency, and the dispersion (of the sampling distribution) of the mean,  $\bar{x}$ , will be the following:

1. The distribution of  $\bar{x}$  will approach normal distribution in shape. (That is, the sampling distribution is a bell-shaped curve.)
2. The mean of the sampling distribution<sup>1</sup> (denoted as  $\mu_{\bar{x}}$ ) will become the population mean,  $\mu$ . (That is,  $\mu_{\bar{x}} = \mu$ .)
3. The standard deviation of the sampling distribution (denoted as  $\sigma_{\bar{x}}$ ) is called *the standard error*, and is related to the population standard deviation,  $\sigma$ , by the formula  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$ .

This may seem like a lot to take in (what with all the jargon, notation, and all) but it really *is* simply a description of a distribution. The next paragraph clarifies each of the CLT's points in turn.

As brief as it is, the CLT is conveniently packed with all sorts of useful information: The sampling distribution is normal in shape — so we can apply all we know about the normal distribution to it (for example, that it's bisected by its mean). Hence, the sampling distribution is *centered* on the population mean. Finally, according to the formula for the sampling distribution's standard deviation (a.k.a the standard error), as the sample size  $N$  grows, the standard error becomes smaller<sup>2</sup> — so the distribution will be less variable/spread out, and thus the estimates will be closer to the parameters<sup>3</sup>

1. You can think of it as "the mean of the means", or the mean of the hypothetical variable *mean*.
2. After all,  $N$  is in the denominator.
3. On the flip side, the larger the original variables's dispersion, the larger the

To summarize, the sampling distribution provides us with a bridge between sample statistics (i.e., estimators) and population parameters (i.e., the estimated). **The CLT provides a description of the sampling distribution: by giving us information about an estimator (in hypothetical repeated sampling), it decreases the uncertainty of the estimation since now we can calculate how close the statistic is to the parameter.**

I say *estimator* and *statistic*, not *mean*, because CLT (or a version thereof) applies to all statistical estimators, as they all have a normal distribution with increasing sample size. The latter is noteworthy because it's true regardless of the shape of the original variable's distribution (in the population): a variable might not be normally distributed but its mean (and other statistics) always is.<sup>4</sup>

If you are wondering about the connection between random sampling and the normal distribution, the following video might help:

standard error and the smaller the original variable's dispersion, the smaller the standard error (as  $\sigma$  is in the numerator)..

4. Many variables tend to be approximately normally distributed in the population. The point I'm emphasizing here is that even when they are not, the statistics of these variables based on random sample data *are* normally distributed. This relates to our discussion of how large  $N$  should be: if the original variable's distribution in the population is close to normal to start with, a smaller  $N$  will be fine. On the other hand, if a variable is not normally distributed in the population (or is too widely dispersed/has a lot of outliers, as reflected in  $\sigma$ ), a relatively large  $N$  will be needed to ensure the normality of the sampling distribution.



*A YouTube element has been excluded from this version of the text. You can view it online here:*

<https://pressbooks.bccampus.ca/simplestats/?p=99>

The video above uses a *Galton board* to demonstrate the connection between randomness and normal curves by showing that balls falling randomly end up distributed approximately into a bell-shaped curve — with the majority in the centre, fewer to the sides, and fewer yet in the “tails”. You can think of a sample mean as one of these balls (all other balls are the means of other samples of the same size). Thus, what we see is that the majority of means would fall in the centre, fewer to the sides, and fewer still in the tail ends. However, since we do not have many means at all but only one, produced by one sample, we are dealing with a probability distribution. In turn, this tells us that the highest probability is the mean to

fall in the centre region, with smaller probability to be to the sides but still close to the centre, and a further decreasing probability the farther it gets from the centre, just like with any probability normal curve<sup>5</sup>.

If you still find all this hopelessly abstract (as I'm sure most do), you can see exactly how we use the CLT for inference in the example below. (Unfortunately, your relief to be back to examples will be premature at this point: we have more necessary theory to cover ahead. On the bright side, we are more than half-way in the chapter so cheer up, the end is near.)

As a heads-up, here's the rationale of what we'll do: In order to explain inference about populations based on samples, we'll reverse-engineer it. That is, we'll start with "knowledge" about the population and, based on the CLT, we'll "infer" the sample statistic. At the end we'll see that following the same logic (but in reverse) we can easily do the opposite — to estimate the population parameter through a sample statistic — which is exactly what we want to do in the first place.

### *Example 6.3 Price of Statistics Textbooks*

5. Of course, in the video you see an *approximation* of a normal curve; after all, this is a finite, not infinite, number of balls. That is why the perfectly normal distribution is only a theoretical concept.

Let's say that university students on average spend \$250 for a statistics textbook, with a standard deviation of \$100 — i.e., we assume to know the population parameters:

$$\mu = 250 \text{ and } \sigma = 100$$

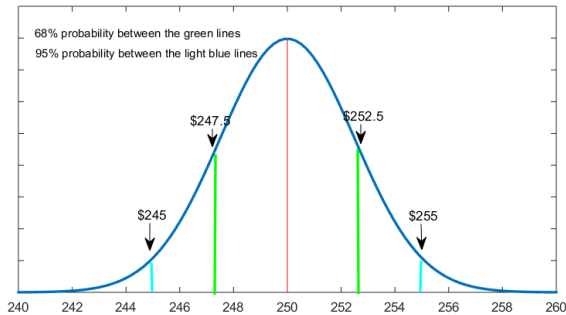
We draw a random sample of  $N=1,600$  students. We want to know the probability for that sample to have a specific mean price paid for statistics textbooks.

To get that probability, we first need the standard error,  $\sigma_{\bar{x}}$ :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{100}{\sqrt{1600}} = \frac{100}{40} = 2.5$$

Next, we can draw the sampling distribution: bell-shaped, centered on  $\mu$ , and with a (standard deviation called) standard error of \$2.5. Applying what we know about the normal distribution in terms of the probability under the curve, we get the following Fig. 6.1.

*Figure 6.1 The Sampling Distribution of the Mean Price of Statistics Textbooks*



That is, we see that 68% of the sample mean prices of statistics textbooks (in hypothetical repeated sampling) would fall between \$247.5 and \$252.5<sup>6</sup> (i.e., within 1 standard error away from the mean, denoted with green in Fig. 6.1) and 95% of the sample means will fall approximately between \$245 and \$255<sup>7</sup> (i.e., within about 2 standard errors away from the mean, denoted with blue in the graph).

Since this is just a heuristic way to *imagine* the sampling distribution, we can restate our finding more correctly: a single, one-off sample mean will fall between \$247.5 and \$252.5 68 percent of the time, and between approximately \$245 and \$255 95 percent of the time.

Or, even *more* precisely, we have a 68 percent probability that the average paid price for statistics books obtained from a random sample of 1,600 students will be between \$247.5 and \$252.5, and a 95 percent probability that it will be approximately between \$245 and \$255. This means that we

6. That is,  $250 - 2.5 = 247.5$  and  $250 + 2.5 = 252.5$ .

7. That is,  $250 - 2(2.5) = 250 - 5 = 245$  and  $250 + 2(2.5) = 250 + 5 = 255$ .

have a 95 percent chance that the sample mean,  $\bar{x}$ , will fall within \$10 (i.e.,  $\pm\$5$ ) of the population mean,  $\mu$ .

Quite good as far as predictions go, eh?

Of course, we rarely would have the population mean to go by, and we would *never* need to estimate a statistics — usually, it's the other way around. But the sampling distribution is the same, as we still go by the CLT: With large  $N$ , it is still a normal curve. With large  $N$ , the sample mean,  $\bar{x}$ , is still approaching the true population mean,  $\mu$ . And, with large  $N$ , the formula for the standard error is still the same,  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$ . For statistical inference, we need only follow the logic presented in Example 6.3 above (albeit in reverse).

However, there is one thing we normally do *not* have in order to proceed: the population standard deviation,  $\sigma$ . We typically use the sample standard deviation,  $s$ , as a substitute, even if this does increase the uncertainty of the estimates<sup>8</sup>

Then, finally, here is **how inference works**, in one paragraph: **we use sample statistics to estimate population parameters** — i.e., the statistics we calculate based on random sample data act as statistical estimators

8. We have a way to account for that, however, as we will see in Section 6.6 on the *t-distribution* below and the concept of *degrees of freedom*..



for what we truly want to know, the unknown population parameters. **We do that by the postulates of the Central Limit Theorem** which describe the sampling distribution, the bridge between the statistics and the parameters. By the CLT, we have **the sampling distribution as normal**. Again, by the CLT, **we can center the sampling distribution on the sample mean, and calculate the sampling distribution's standard error using the sample standard deviation**. By applying the properties of the normal probability distribution to the sampling distribution, **we then produce population estimates**. Ta-da!

I will end this section with an example to illustrate the full process from the beginning to the end.

#### *Example 6.4 Average Annual Income*

Imagine you are interested in the average annual income in a medium-size city. You randomly select  $N=1,600$  people living in that city and ask them about their annual income. You then calculate the mean of the resulting variable as \$50,000, and the standard deviation as \$12,000. I.e.,

$$\bar{x} = 50,000 \text{ and } s = 12,000$$

*As a first guess, you could say that the average annual income in the city is \$50,000. However, since we know this is an estimate, and random error exists, you can do better:*

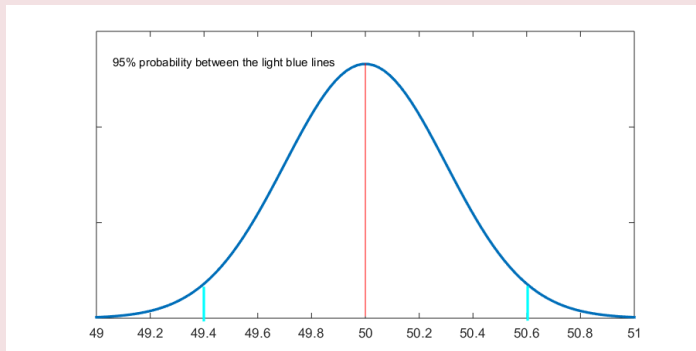
you can also provide information about how certain you are about your estimate along with some margins for error.

To do that, you need to draw the sampling distribution of the mean. Following the CLT, you draw the sampling distribution as a normal curve centered on \$50,000. At this point, you also need information about the sampling distribution's dispersion, i.e., its standard error. You substitute the  $s$  you do know for the  $\sigma$  you don't<sup>9</sup>:

$$\hat{\sigma}_{\bar{x}} = s_{\bar{x}} = \frac{s}{\sqrt{N}} = \frac{12000}{\sqrt{1600}} = \frac{12000}{40} = 300$$

Fig. 6.2 shows the resulting sampling distribution.

*Figure 6.2 Average Annual Income*



Based on the figure above (and following the same logic as in the previous Example 6.3), we find that the average annual income of the city's population will be between

9. Recall that a "hat" over a symbol indicates it being estimated.

\$49,400 and \$50,600 with 95 percent probability<sup>10</sup>. That is, we can be 95 percent confident that the city's average annual income will be within \$1,200 of the sample average of \$50,000, or, that the city's average annual income is  $\$50,000 \pm \$600$ , with 95 percent certainty. (Don't worry, all this talk of *confidence* and *certainty* will be explained in the next section.)

You should be able to appreciate that this “average annual income of  $\$50,000 \pm \$600$ ” is a much more qualified and precise statement than simply assuming the population average is the same as the sample average (which it is likely not). **Now you know how much potential variability the population mean has, with a specific (and quite high!) level of certainty.**

This is no way trivial, and the best “guess” you can offer as an estimate of the population mean. No other research method using sample data is able to produce a closer level of generalizability of the sample findings to the level of population, much less with the mathematical, probability-theory-backed evidence offered by random sampling. This is what statistical inference does, and now you even know how and why it works! In the next section, you can try it for yourself.

We are almost but not quite done with this abstract monster of a chapter. There is a light at the end of the tunnel — what is left is tying some loose ends, formally introducing a concept we're already using (psst, that's the

10. We get these bounds (i.e., within about 2 standard errors away from the mean) through  $50,000 - 2(300) = 50,000 - 600 = 49,400$  and  $50,000 + 2(300) = 50,000 + 600 = 50,600$ .

*confidence* I mentioned above), and providing some final details on inference in the next section — and then we are good to go: we can start on some real research and working with variables again in Chapter 7!

---

## 6.7 Confidence Intervals

In our discussion on statistical inference so far, I have used not one type of estimators but *two*, without bringing your attention to it. Probability theory and the Central Limit Theorem describing the sampling distribution of statistics provide us with two types of estimators, called *point estimators* and *interval estimators*.

**A single sample statistic which estimates a population parameter** — and which offers a “best guess” for that parameter — **is a *point estimator***. We have worked with several point estimates by now: the sample mean  $\bar{x}$  is a point estimate of the population mean  $\mu$  while the sample standard deviation  $s$  is a point estimate (which we can note as  $\hat{\sigma}$ ) of the population standard deviation  $\sigma$ .

Similarly, I’ll add another useful point estimate, of the sample *proportion*. Imagine we are interested in studying unemployment. We take a random sample which reveals that, say, 10 percent of the sample respondents report being unemployed. Thus, we have the sample proportion  $p$  as 0.1 and we can use that proportion as a point estimate of the proportion of the population which is unemployed. We denote population proportions by the small-case Greek letter  $p$  which is  $\pi$ <sup>1</sup>. In other words, the sample

1. Pronounced PAI, as you probably already know from the mathematical constant  $\pi=3.14$ . While we use the letter  $\pi$  for both population proportions and the mathematical constant, context provides enough clues to differentiate them.

proportion  $p$  serves as a point estimate of the population proportion  $\pi$ .

You'll be happy to know that you are also already familiar with the other, *interval*, type of statistical estimators. As their name suggests, **interval estimators, called *confidence intervals*, provide not just one number as a best guess but a whole set of plausible values for the population parameter.**

If you recall Examples 6.3 and 6.4 from the previous section, you'll recognize that we already calculated confidence intervals. In Example 6.4 on the average annual salary, we found a range of values within which the average annual salary of the city population was estimated to fall. Specifically, the average annual salary of the random sample was \$50,000 and we were able to estimate with 95% certainty that the average annual salary of the city population would fall between \$49,400 and \$50,600. This range of values between \$49,400 and \$50,600 is in effect a confidence interval (a 95% confidence interval, to be precise). The actual numbers "bracketing" the interval are called *error bounds*; the interval itself is between, and including, the *lower error bound* and the *upper error bound*.

Up until now, we calculated the confidence interval in a fast and easy way as I wanted to get the point of the logic underlying statistical inference across. At this time, however, we need to get more technical and precise about it.

First, let's revisit how we did it in the previous section to

refresh your memory; then I'll show you the *more* correct way to do it. (Before you panic, know that what we did before was not incorrect; we just used rounded numbers to make calculations easier/faster.)

This is the information about the sample mean and standard deviation we had from Example 6.4 *Average Annual Income* (without the dollar signs for clarity of presentation):

$$\bar{x} = 50000$$

$$s = 12000$$

$$N = 1600$$

Our starting point is the sample mean (which, according to the CLT approximates the population mean, with large  $N$ ). In order to calculate a confidence interval around the sample mean  $\bar{x}$ , we first need to get the standard error  $\sigma_{\bar{x}}$ , given by the CLT-based formula:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

We don't know the population standard deviation  $\sigma$  but we estimate it with its point estimator  $s$ , so we get:

$$\hat{\sigma}_{\bar{x}} = s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

Substituting  $s$  and  $N$  in the formula gives us the following:

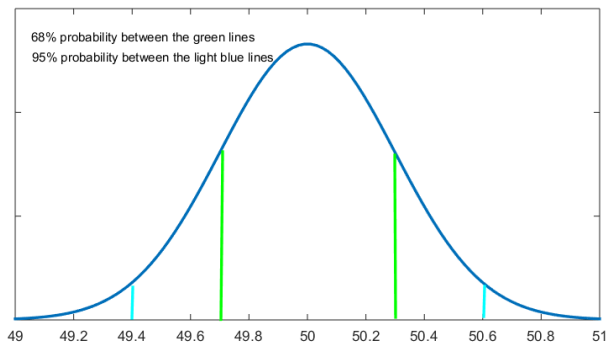
$$s_{\bar{x}} = \frac{12000}{\sqrt{1600}} = \frac{12000}{40} = 300$$

Now, by the CLT, we have everything we need for the

sampling distribution: its mean (as estimated by the sample mean  $\bar{x}$ , its standard deviation (i.e., the standard error  $\sigma_{\bar{x}}$ ), and its shape as a normal curve. From Section 5.2.4 (<https://pressbooks.bccampus.ca/simplestats/chapter/5-2-4-the-real-normal-distribution/>), we know the probabilities under the normal curve, and that 68 percent of cases<sup>2</sup> fall within 1 standard deviation from the mean while 95 percent of cases fall within about 2 standard deviations from the mean.

The resulting graph was presented in Fig. 6.2 in the previous section. Here it is again, this time with the 68 percent demarcations included.

*Figure 6.3 Average Annual Income (in thousands of dollars), Revisited*



2. Here you can imagine the cases as the hypothetical means over infinite sampling.



However, to calculate a confidence interval we don't need to draw the sampling distribution every time; we just need to keep in mind what it represents in terms of probabilities.

From our discussion of Example 6.4 in the previous section and now, we can easily deduce the basic formula for calculating a confidence interval:

- for a 68% confidence interval around the mean, we would have

$$\circ \bar{x} \pm 1 \times \hat{\sigma}_{\bar{x}}$$

- for a 95% confidence interval around the mean, we would have

$$\circ \bar{x} \pm 2 \times \hat{\sigma}_{\bar{x}}$$

We could even add the 99% confidence interval, encompassing values within about 3 standard deviations away from the mean:

- for a 99% confidence interval, we would have

$$\circ \bar{x} \pm 3 \times \hat{\sigma}_{\bar{x}}$$

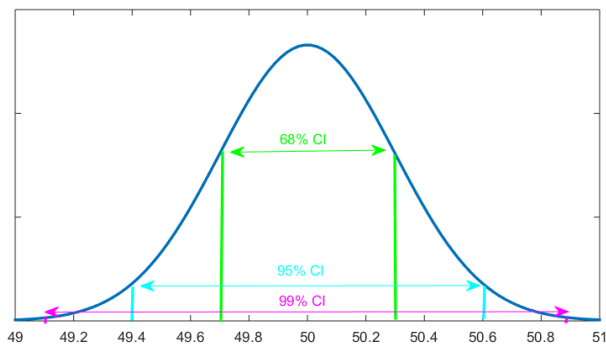
Using the data from Example 6.4 for illustration, we then have the following confidence intervals (CI):

- 68% CI:  $\bar{x} \pm 1 \times \sigma_{\bar{x}}$   
 $= 50000 \pm 1 \times 300 = 50000 \pm 300 = (49700; 50300)$
- 95% CI:  $\bar{x} \pm 2 \times \sigma_{\bar{x}}$   
 $= 50000 \pm 2 \times 300 = 50000 \pm 600 = (49400; 50600)$

- 99% CI:  $\bar{x} \pm 3 \times \sigma_{\bar{x}}$   
 $= 50000 \pm 3 \times 300 = 50000 \pm 900 = (49100; 50900)$

Fig. 6.4 illustrates these confidence intervals.

*Figure 6.4 Confidence Intervals for Average Annual Income*



That is, we find that the average annual income for the city population is between \$49,700 and \$50,300 with 68% certainty; it is between \$49,400 and \$50,600 with 95% certainty; and it's between \$49,100 and \$50,900 with 99% certainty. Alternatively, we could report that the average annual income of the city population is  $50,000 \pm \$300$  with 68% confidence;  $50,000 \pm \$600$  with 95% confidence; and  $50,000 \pm \$900$  with 99% confidence.

The  $\pm \hat{\sigma}_{\bar{x}}$  (i.e., the plus and minus the estimated standard error) represents the *margin of error* for the specific confidence interval.

Now that you understand the principle of calculating

confidence intervals, let's start doing it with greater precision, as we normally would in real-life research.

Even if I used “1, 2, 3 standard deviations/errors away from the mean” in the calculations so far, this is a quick-and-easy rounding only *approximating* the real formula for confidence interval. From Section 5.2.5 (<https://pressbooks.bccampus.ca/simplestats/chapter/5-2-5-the-real-use-of-z-values/>), we know that the probabilities under the normal curve are associated with specific z-values.

If you check the z-values (standard normal distribution) table<sup>3</sup>, you'll actually see that the precise z-values associated with 95% probability<sup>4</sup> and 99% probability<sup>5</sup> are 1.96 (almost but not quite 2) and 2.58 (almost but not quite 3), respectively.

Now you know that even if the z-value associated with 68% probability<sup>6</sup> is indeed 1, the other two confidence intervals we have used so far need to be recalculated properly:

- 68% CI:  $\bar{x} \pm 1 \times \hat{\sigma}_{\bar{x}}$   
 $= 50000 \pm 1 \times 300 = 50000 \pm 300 = (49700; 50300)$

3. Like the one here <https://www.mathsisfun.com/data/standard-normal-distribution-table.html>.
4. Since the distribution is symmetric, recall that the table only gives you *half* the probability (i.e., between the mean and the z-score). Thus for 95% (i.e., the two sides together), you need to check  $(95/2=)$  47.5%.
5. By analogy, for 99% (i.e., the two sides together), you need to check  $(99/2=)$  49.5%.
6. By analogy, for 68% (i.e., the two sides together), you need to check  $(68/2=)$  34%.

- 95% CI:  $\bar{x} \pm 1.96 \times \hat{\sigma}_{\bar{x}}$   
 $= 50000 \pm 1.96 \times 300 = 50000 \pm 588 = (49412; 50588)$
- 99% CI:  $\bar{x} \pm 2.58 \times \hat{\sigma}_{\bar{x}}$   
 $= 50000 \pm 2.58 \times 300 = 50000 \pm 774 = (49226; 50774)$

To interpret, we find that we can be 95% certain that the average annual income of the population is between \$49,412 and \$50,588. As well, we find that we can be 99% certain that the average annual income is between \$49,226 and \$50,774.

Furthermore, although going by “1, 2, 3 standard deviations/errors” makes intuitive sense, in reality would you be happy to learn anything “with 68% certainty”? Sixty-eight percent certainty is hardly certain at all! (As such, it is pretty much never used outside of teaching.)

On the other hand, while the 95% and 99% confidence intervals are the most widely used and useful ones, there is no need to restrain yourself, should you choose to calculate *any* confidence interval you wish.

**The general formula for a confidence interval is thus:**

- **Any % CI:**  $\bar{x} \pm z \times \sigma_{\bar{x}}$

To calculate this, you need to choose the level of certainty you want; once you have the probability, (divide it by two and) check its corresponding z-value, then multiply it by the standard error to get the margins of error with the desired probability level of certainty.

For example, I might want the 90% CI (not as popular as the other two but still a relevant confidence interval that has its uses).

I check for the z-value associated with 90% probability<sup>7</sup> in a z-distribution table and I find that it's 1.65. Then, for the example used above, I would get:

- 90% CI:  $\bar{x} \pm 1.65 \times \hat{\sigma}_{\bar{x}}$   
 $= 50000 \pm 1.65 \times 300 = 50000 \pm 495 = (49505; 50495)$

Or, I can be 90% certain that the average annual income of the population of that city is between \$49,505 and \$50,495.

By analogy, you can thus produce *any* confidence interval with *any* level of certainty you want.

A bit more on confidence intervals in the next section.

7. By analogy, for 90% (i.e., the two sides together), you need to check  $(90/2=)$  45%.



---

## 6.7.1 Additional Confidence Intervals Considerations

**Precision vs. certainty.** One thing you you might have noticed from the calculations in the examples in the previous section is that **the more certainty you get, the larger your confidence interval becomes** (or vice versa: the smaller the interval, the less precise your estimate):

Based on the annual income details from Example 6.4, we had

- between \$49,700 and \$50,300 with 68% confidence;
- between \$49,505 and \$50,495 with 90% confidence;
- between \$49,412 and \$50,588 with 95% confidence; and
- between \$49,226 and \$50,774 with 99% confidence.

Of course, who wouldn't want both more precise *and* more certain estimates? Unfortunately there simply is no way to have our cake and eat it too: As you can see above, the more confident in our estimate we get, the more the error bounds of the confidence intervals spread out wider. There is a trade-off between precision and confidence. **The more precise our estimate, the less certain we are of**

**it; the more confident we are in our estimate, the less precise our “guess” is.**

Logically, this makes a lot of sense: imagine the population parameter as a target and estimation as throwing a dart at it. The smaller the target, the more precise you’ll have to be but also the less confident of hitting it. At the same time, increasing the target size will accommodate less precise “shots” while simultaneously increasing the certainty of the target being hit.

**And why can’t we have a 100% CI?** The non-technical answer is simply because a statistical estimator is based on a sample drawn from a population of interest: as long as you don’t have data from your *entire* population, there will always be a possibility for random error (and uncertainty).

The more technical answer lies in the characteristics of the normal probability distribution. Specifically, the normal curve *approaches* but never *reaches* the horizontal axis; the probability in its “tails” is not bound — i.e., a probability for *any* z-value exists, no matter how small or large, and it never reaches 0. Thus, a 100% confidence interval would result in  $-\infty$  to  $+\infty$ , i.e., it would be virtually *infinitely* large, to accommodate the perfect certainty. Logically, no bound, finite interval can provide 100% certainty by the nature of statistical *inference* itself. (Since, at 100%, it would stop being inference altogether: we will have no need to *estimate*, as we would *know*.)

**The effect of sample size on confidence intervals.** Let’s also consider the effect of sample size on the precision and level of certainty of confidence intervals. In Section 6.5 (<https://pressbooks.bccampus.ca/simplestats/>)



[chapter/6-5-the-sampling-distribution/](#)) I attempted to convince you that increasing the sample size beyond a specific (large) number becomes not only unfeasible in a world of limited resources but also statistically pointless. Let's see if I could further support my claim by the effect of sample size on the standard error.

To recall, we find the standard error in the following way:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

where we estimate  $\sigma$  (the standard deviation of the population) with  $s$  (the standard deviation of the sample) to get

$$\hat{\sigma}_{\bar{x}} = s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

We already established that a larger  $N$  would result in a smaller standard error (as  $N$  is in the denominator). Given the formula for calculating confidence intervals, a smaller standard error should in turn lead to smaller intervals (i.e., to more precise estimates) *at a fixed level of certainty*. The question is — how much smaller?

#### *Example 6.5 The Effect of Sample Size on Confidence Intervals*

Going back to our *Average Annual Income* (Example 6.4) specifications, we had that

$$N = 1600$$

$$\bar{x} = 50000$$

$$s = 12000$$

We had also already calculated its 95% CI:

$$\begin{aligned} \bullet \text{ 95\% CI: } \bar{x} \pm 1.96 \times \hat{\sigma}_{\bar{x}} \\ = 50000 \pm 1.96 \times 300 = 50000 \pm 588 = (49412; 50588) \\ . \end{aligned}$$

What would happen if we increased the sample size to, say,  $N=10,000$ ?

As usual, we start with calculating the standard error:

$$\hat{\sigma}_{\bar{x}} = s_{\bar{x}} = \frac{s}{\sqrt{N}} = \frac{12000}{\sqrt{10000}} = \frac{12000}{100} = 120$$

Then, the new 95% CI would be

$$\begin{aligned} \bullet \text{ 95\% CI: } \bar{x} \pm 1.96 \times \hat{\sigma}_{\bar{x}} \\ = 50000 \pm 1.96 \times 120 = 50000 \pm 235 = (49765; 50235) \\ . \end{aligned}$$

To be sure, the larger- $N$  confidence interval *is* smaller; we did gain precision. But consider these numbers for what they actually are, in *actual dollar* terms, had this been a real-life research instead of a hypothetical example. With a sample of  $N=1,600$  we found that, with 95% certainty, the average annual income for the population is between \$49,412 and \$50,588. We now find that had we a sample

of  $N=10,000$ , the average annual income of the population would be between \$49,765 and \$50,235.

The precision “gain” between the two sample sizes is \$353 on each error bound; i.e., our estimate of average annual income of the population becomes  $\pm\$353$  more precise (a total “gain” of \$706). At the same time, consider that surveying a sample size of  $N=10,000$  would cost more than *six times* more than surveying one of  $N=1,600$  (as 10,000 is 6.25 times more than 1,600). Would this be worth it, to only be able to improve your estimate by \$350, give or take, on both sides, when the actual sums we are dealing with are in the tens of thousands dollars magnitude?

Most people would agree that \$49,412 to \$50,588 is precise enough, and that there’s no need to waste six times more resources on such a relatively insignificant gain in precision when it comes to average annual income<sup>1</sup>.

Bear in mind, however, that had we been discussing effectiveness of a life-saving medical treatment instead of average annual income, our preferences regarding the trade-off between precision and cost would most likely be different. Thus, the actual value of increasing sample size cannot be judged solely on statistics grounds: what

1. To demonstrate the effect of sample size only, this example keeps the other conditions (i.e., the sample mean and standard deviation) the same.

Arguably, however, a larger  $N$  would have a mean and a standard deviation “truer” to the population. To the extent that a larger sample ends up with a smaller standard deviation, the standard error would be further reduced, and the confidence interval would be even narrower, thus gaining more precision. Still, the point of the effect of sample size *per se* remains.

is considered a small/insignificant change in precision for one thing may very well be a large and worthy change in another context. Still, in social science research there's rarely a need for such increasing precision of inference no matter the costs, even if larger samples are generally preferred<sup>2</sup>

2. Large sample sizes are very useful for gaining *power* in detecting associations between variables, as you'll see in the remaining chapters..

---

## 6.7.2 Confidence Intervals for Proportions

Just like we may like to know the population mean of something (like the average annual income above), we might want to know the population *proportion* of something else (like, say, the proportion of Canadians working part time). Population proportions are, like population means, parameters that can be estimated.

**The principle of estimating a population proportion through a confidence interval is the same as estimating the mean — we need a standard error for creating error bounds around the sample statistic (in this case, the proportion).**

The question, however, is *how* to calculate the standard error of a proportion. After all, the CI formula requires the use of a standard deviation; a standard deviation that proportions do *not* have (as the dispersion measures we studied are only applicable to interval/ratio data, if you recall from Section 4.4 (<https://pressbooks.bccampus.ca/simplestats/chapter/4-4-standard-deviation/>)). Thus, calculating the mean and the standard deviation of an interval/ratio variable is all well and good but what do we do with proportions, considering that they relate to *categories*, not numerical values?

In fact, there *is* a way to measure dispersion in a binary distribution (i.e., where there are only two categories/

outcomes, e.g., employed vs. unemployed, women vs. men, undergraduate vs. graduate students, heads vs. tails, approval vs. disapproval, yes vs. no, success vs. failure, etc.). Unlike interval/ratio variables (which usually have an approximately normal — and *continuous* — distribution), such a binary distribution is a *discrete* distribution.

Since the standard deviation is off the table, here is an example to demonstrate the logic underlying the measurement of variability of proportions.

#### *Example 6.6 Variability Through Clothing*

Imagine you have a friend who is partial to the colour black so much so that they always wear a monochromatic, all-black outfit. Then one day you notice your friend is wearing a single article of a different colour, say, dark purple. Arguably, that's more variability than wearing all-black, but the outfit will still be predominantly black. Then on the next day, there are two pieces of purple amid all the black, then three, then four, and so on. At what point would your friend's outfit stop being "predominantly black" and would become "predominantly purple"? And what would happen eventually, if the exchanging-black-for-purple trend continues?

The answer to the latter question is obvious: the end point of such a trend would be for the outfit to become monochromatic again, this time all-purple. Now think about

variability. At what point was there the greatest and at what point was there the least amount of variability in your imaginary friend's outfit?

To make it easier, let's add a numerical aspect to what we have imagined, and say that your friend's outfit consisted of 10 articles of clothing (and accessories) to start with, and then your friend swapped a black article for a purple article on each successive day, for ten days straight after that. Table 6.1 illustrates.

*Table 6.1 Black and Purple Articles of Clothing*

	Black Articles	Purple Articles
Initial state	10	0
Day 1	9	1
Day 2	8	2
Day 3	7	3
Day 4	6	4
Day 5	5	5
Day 6	4	6
Day 7	3	7
Day 8	2	8
Day 9	1	9
Day 10	0	10

Again, on what day(s) would your friend’s outfit be the least and the most variable in terms of colour? Looking at Table 6.1, it’s not difficult to spot that the least variable were your friend’s initial (all-black) outfit and what they wore on Day 10 (all-purple), both consisting of a single colour. There is a slight variability on Days 1 and 9 (when there was a *single* article of different colour); then more variability on Days 2 and 8 (when there were *two* articles of different colour); then even more variability on Days 3 and 7 (when your friend had *three* different-coloured articles); and yet even more variability on Days 4 and 6 (when there were *four* articles of different colour).



The outfit was most variable on Day 5, when it was half-black and half-purple, neither colour predominating.

Going by “half-black and half-purple”, let’s restate the information in Table 6.1 in terms of proportions, as this will help us generalize the logic without the constraint of an actual count (of 10 articles of clothing, or anything else).

*Table 6.2 (A) Black and Purple Articles of Clothing: Proportions*

	Black Articles	Purple Articles
Initial state	1	0
Day 1	0.9	0.1
Day 2	0.8	0.2
Day 3	0.7	0.3
Day 4	0.6	0.4
Day 5	0.5	0.5
Day 6	0.4	0.6
Day 7	0.3	0.7
Day 8	0.2	0.8
Day 9	0.1	0.9
Day 10	0	1

One convenient way to quantify what we found in terms of the least and the largest variability is through multiplying the proportions in the two columns, like so:

*Table 6.2 (B) Black and Purple Articles of Clothing: Variability*

	Black Articles	Purple Articles	Variability
Initial state	1	0	1(0)= <b>0</b>
Day 1	0.9	0.1	0.9(0.1)= <b>0.09</b>
Day 2	0.8	0.2	0.8(0.2)= <b>0.16</b>
Day 3	0.7	0.3	0.7(0.3)= <b>0.21</b>
Day 4	0.6	0.4	0.6(0.4)= <b>0.24</b>
Day 5	0.5	0.5	0.5(0.5)= <b>0.25</b>
Day 6	0.4	0.6	0.4(0.6)= <b>0.24</b>
Day 7	0.3	0.7	0.3(0.7)= <b>0.21</b>
Day 8	0.2	0.8	0.2(0.8)= <b>0.16</b>
Day 9	0.1	0.9	0.1(0.9)= <b>0.09</b>
Day 10	0	1	0(1)= <b>0</b>

That is, starting from zero, variability is the highest at precisely the half-and-half point, when neither outcome/ category (in our example, neither *colour*) predominates.

Now we are ready for the formula to measure the

dispersion of a proportion. I demonstrate it by restating Table 6.2 (B), by designating black as 1 and purple as 0, and taking black as the colour of interest (i.e., all proportion will be expressed in terms of black).

*Table 6,2 (C) Black and Purple Articles of Clothing: Generalized*

	<b>Black Articles</b>	<b>Non-black Articles</b>	<b>Variability</b>
<b>Initial state</b>	1	0	$1(0)=0$
Day 1	0.9	(1-0.9)	$0.9(1-0.9)=\mathbf{0.09}$
Day 2	0.8	(1-0.8)	$0.8(1-0.8)=\mathbf{0.16}$
Day 3	0.7	(1-0.7)	$0.7(1-0.7)=\mathbf{0.21}$
Day 4	0.6	(1-0.6)	$0.6(1-0.6)=\mathbf{0.24}$
Day 5	0.5	(1-0.5)	$0.5(1-0.5)=\mathbf{0.25}$
Day 6	0.4	(1-0.4)	$0.4(1-0.4)=\mathbf{0.24}$
Day 7	0.3	(1-0.3)	$0.3(1-0.3)=\mathbf{0.21}$
Day 8	0.2	(1-0.2)	$0.2(1-0.2)=\mathbf{0.16}$
Day 9	0.1	(1-0.1)	$0.1(1-0.1)=\mathbf{0.09}$
Day 10	0	(1-0)	$0(1-0)=\mathbf{0}$

And there you have it in the Table 6.2 (C) above, the

formula for calculating variability for a proportion (i.e., for a discrete binary variable). Since we denote sample proportions with  $p$  and population proportions with  $\pi$ , **the variability of a proportion is given by multiplying the proportion of the outcome we're interested in by 1 minus the proportion** (i.e., on the *other* outcome's proportion) — that is, we have  $p(1-p)$  for samples and  $\pi(1-\pi)$  for populations.

Technically speaking, this variability is the proportion's *variance*:

$$\sigma^2 = \pi(1 - \pi) = \text{variance of the proportion}$$

As usual, to get the proportion's *standard deviation*, we need a square root of the variance:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\pi(1 - \pi)} = \text{standard deviation of the proportion}$$

With this, we are finally ready to get back to calculating a confidence interval for a proportion, as we now have everything we need to calculate its standard error. If you recall, the formula for the standard error was:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

Substituting the standard deviation of the proportion, we get:

$$\begin{aligned} \sigma_p &= \frac{\sigma}{\sqrt{N}} = \frac{\sqrt{\pi(1-\pi)}}{\sqrt{N}} = \sqrt{\frac{\pi(1-\pi)}{N}} \\ &= \text{standard error of the proportion} \end{aligned}$$

Of course, when we don't have the population standard deviation, we estimate it with the sample standard deviation — i.e., we need to substitute  $p$  for  $\pi$ :

$$\hat{\sigma}_p = \frac{\sigma}{\sqrt{N}} = \frac{\sqrt{p(1-p)}}{\sqrt{N}} = \sqrt{\frac{p(1-p)}{N}}$$

= estimated standard error of the proportion

Following our true and tested formula for confidence intervals (i.e., the sample statistic  $\pm z \times$  the standard error), we ultimately get **the confidence interval for a proportion**:

- **Any % CI:**  $p \pm z \times \hat{\sigma}_p = p \pm z \times \sqrt{\frac{p(1-p)}{N}}$

As with the mean, we can calculate a confidence interval with *any* preferred level of certainty by substituting with the  $z$ -value associated with that probability. For example, the 95% confidence interval for the proportion would be:

- 95% CI:  

$$p \pm 1.96 \times \hat{\sigma}_p = p \pm 1.96 \times \sqrt{\frac{p(1-p)}{N}}$$

If you find all this too technical and abstract, the following example should help.

*Example 6.7 Part-Time Workers in Canada, Age 25-54*

Let's say we want to know what proportion of Canadian workers work part-time, and that we are especially interested in what Statistics Canada calls "the core ages" 25 to 54 (REFERENCE Statistics Canada, 2017 [<https://www150.statcan.gc.ca/n1/pub/71-222-x/71-222-x2018002-eng.htm>]).

We conduct a survey of  $N=1,600$  Canadian individuals aged 25-54 and find that 12 percent of our respondents work part-time. As usual, we want to estimate the proportion of *all* Canadians aged 25-54 who work part time.

We start with calculating the standard error:

$$\hat{\sigma}_p = \sqrt{\frac{p(1-p)}{N}} = \sqrt{\frac{0.12(0.88)}{1600}} = \sqrt{\frac{0.106}{1600}} = \frac{0.325}{40} = 0.008$$

Then, a 95% confidence interval for the proportion would be:

- 95% CI:

$$p \pm 1.96 \times \hat{\sigma}_p = p \pm 1.96 \times 0.008 = 0.12 \pm 0.016 = (0.104; 0.136)$$

Thus we estimate with 95% certainty that (i.e, 95% of the time such a study is undertaken it will find that) between 10.4% and 13.6% of the Canadian workers aged 25-54 work part-time. Alternatively, we can say with 95% certainty that

12%  $\pm$  1.6 percentage points of Canadian workers aged 25-54 work part time.

As there is a lot to take in here, a second example is in order.

#### *Example 6.8 Women in Managerial Positions*

Let's say a large, nationally-representative study of  $N=10,000$  finds that women in Canada occupy 36 percent of managerial positions. [REFERENCE <https://www.expertmarket.com/female-managers>] What would be the estimate for Canada as a whole?

The estimated standard error of the proportion would be:

$$\hat{\sigma}_p = \sqrt{\frac{p(1-p)}{N}} = \sqrt{\frac{0.36(0.64)}{10000}} = \sqrt{\frac{0.230}{10000}} = \frac{0.48}{100} = 0.005$$

As in the previous examples, the 95% confidence interval for the proportion would be:

- 95% CI:

$$p \pm 1.96 \times \hat{\sigma}_p = p \pm 1.96 \times 0.005 = 0.36 \pm 0.01 = (0.35; 0.37)$$

That is, we can estimate with 95% certainty (or, 95% of the time such a study is undertaken it will find) that between 35% and 37% of managerial positions in Canada are occupied by women. Alternatively, we can say with 95% certainty that women occupy  $36\% \pm 0.01$  percentage points of managerial positions in Canada<sup>1</sup>.

If you find this a bit too precise to believe, note the quite large sample size of  $N=10,000$ . As established above, confidence intervals based on large  $N$  and around proportions indicating not very strong variability (after all, the sample statistics indicated that managerial positions are predominantly occupied by men) tend to have small standard errors (due to the relatively small numerator (the variability) and the large denominator (the sample size)).

1. In this chapter I have presented the most commonly used interpretation of confidence intervals, and the one most frequently taught to introductory statistics students. I should point out, however, that this is one of those instances (of which I spoke in the introduction to this book) where the reality is a bit different than what is being taught. The interpretation presented here is easier to understand and follows a logic that is more intuitive to students than what confidence intervals *really* tell us. Briefly, the range of plausible values we find are just that -- values that the population *could* have, as we haven't ruled them out yet, and 95% (or 99%) of the time such studies will not be able to rule these plausible values out (REFERENCE van der Zee, 2017 [How (Not) To Interpret Confidence Intervals, in the hyperlink]). This, technically speaking, is somewhat different than the "95% (or 99%) certainty that the population mean/proportion *will be* between the calculated error bounds" version we usually work with. If you'd like to go down that particular rabbit hole, go here: <http://www.timvanderzee.com/not-interpret-confidence-intervals/>. For everyone else, the interpretation of confidence intervals presented so far in this chapter should be enough.



Now finally it's your turn to try, first with means...

*Do It! 6.1 Average Height of NHL Players, In Inches This Time*

Let's say that a random sample of  $N=900$  past and present players in the National Hockey League finds that the average height of players is 73 inches, with a standard deviation of 3 inches. What can you say about the average height of NHL players as a whole? Construct a 95% and a 99% confidence intervals for the average height of NHL players.

Answer: (72.8; 73.2) and (72.7; 73.3), respectively.

... And now with proportions.

*Do It! 6.2 Paying Off Student Debt Within Three Years After Graduation*

Let's say that a sample of  $N=1,600$  finds that only 34 percent of Canadians with a bachelor's degree have paid off their student loans within three years after graduation. Can you estimate the rate for all Canadians with a bachelor's degree? Construct both a 95% and a 99% confidence interval for that rate.

Answer: (31.7; 36.3) and (31; 37), respectively.

To summarize, confidence intervals allow us to estimate population parameters with a specific level of precision and certainty. We construct them based on the idea of the (normally distributed) sampling distribution of the mean (or the proportion) using CLT's postulates: centering the interval on the sample mean (or proportion) and taking that many times the standard error below and above the mean (or proportion). The "how many times the standard error" (i.e., the z-score) determines the interval's confidence (i.e., certainty in terms of probability) level.

Before we move on to variable associations (along with further uses of confidence intervals in statistics inference; you didn't think it was just this, did you?), let's finally address the glaring omission in my presentation so far: How come we can simply use the sample standard deviation  $s$  instead of the population standard deviation  $\sigma$  in calculating the standard error? I left that explanation for last, in the next section.

---

## 6.8 The t-Distribution

If, having reached this chapter's final section, after all we had been through, random sampling, sampling distribution, CLT, parameters, estimates, statistics, confidence intervals, you are now groaning in dismay — *why is there even more to this topic??*<sup>1</sup> *always* more. Much, much more; it's not a matter *if* but of *how much* something is left out. — take heart, this is a short explanation I kept for last, through a brief introduction of new concept.

If you recall, when we needed to calculate the standard error of the mean (or proportion) in the previous few sections, I simply replaced the *unknown* population standard deviation  $\sigma$  with the *known* sample standard deviation  $s$  in the formula. This is what I did:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \text{standard error of the mean}$$

Substituting in  $s$  for  $\sigma$  we had:

$$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{N}} = \text{estimated standard error of the mean} = s_{\bar{x}}$$

Similarly, for the proportion we had

$$\sigma_p = \frac{\sigma}{\sqrt{N}} = \frac{\sqrt{\pi(1-\pi)}}{\sqrt{N}} = \sqrt{\frac{\pi(1-\pi)}{N}} = \text{standard error of the proportion}$$

1. As a general principle, in introductory texts such as this there is

and substituting the known sample proportion  $p$  for the unknown population proportion  $\pi$  in calculating the proportion's variability, we ended up with:

$$\hat{\sigma}_p = \frac{\sigma}{\sqrt{N}} = \frac{\sqrt{p(1-p)}}{\sqrt{N}} = \sqrt{\frac{p(1-p)}{N}} = \text{estimated standard error of the proportion}$$

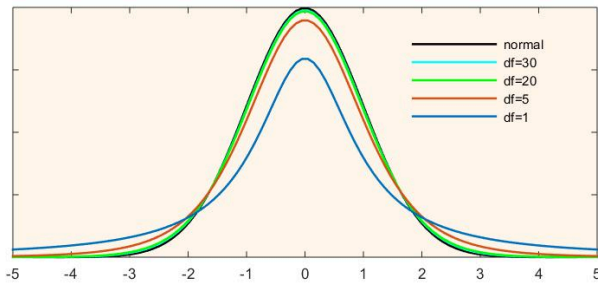
But why can we do that?

The more observant of you might have noticed that I swept the explanation for this change under the carpet and simply moved on — but why should the variability of the population be the same as the sample?

In truth, they are not — or rather, they *might* be; there's just no way to know. That is, by using the sample statistics to estimate the variability of the population, we introduce more *uncertainty* in the calculation. When we do that, we actually move away from using the normal distribution and its associated z-values. What we end up using is something similar, called the *t-distribution*<sup>2</sup>: an entire set of bell-shaped curves, accounting for each and every sample size  $N$ . Figure 6.5 illustrates.

*Figure 6.5 The Normal vs. the t-Distribution*

2. Also called the *Student's t-distribution*, after the pseudonym of William Gosset who introduced it to statistics (along with many other concepts). Due to contractual obligations, William Gosset used to publish under the name of "Student" (Pagels, 2018). Here you can find more about his curious case: <https://medium.com/value-stream-design/the-curious-tale-of-william-sealy-gosset-b3178a9f6ac8>.



The *t*-distribution provides a separate bell-shaped curve for each possible sample size, thus helping us “ground”, as it were, the estimation in the reality of an actual sample of a specific size.

The accommodation of the sample size is done through the concept of *degrees of freedom* (commonly abbreviated to *df*). The degrees of freedom represent the number of values in a statistical calculation that are free to vary. In the case of the *t*-distribution, the degrees of freedom are  $N-1$  as one degree of freedom is reserved for estimating the mean, and  $N-1$  degrees remain for estimating the variability. Unlike with *z*-values, where each *z*-value represents a specific probability under the normal curve, the probabilities associated by *t*-values are calculated based on its degrees of freedom.

Still, none of this explains why I was able to shamelessly switch from using the *z*-distribution to the *t*-distribution, without any change to the standard error and confidence interval calculations in the examples in the previous sections. If *z*-values and *t*-values (and their associated probabilities) are different, shouldn't the calculations differ too?

Before I reassure you that all is well (and it is), let's revisit what z-values actually represent. From Chapter 5 you know that the z-value is the distance between a case and the mean, expressed in terms standard deviations (i.e., standardized):

$$z = \frac{x_i - \bar{x}}{s}$$

The reason we were able to use  $z=1$ ,  $z=1.96$ , and  $z=2.58$  in the calculations of the 68%, 95%, and 99% confidence intervals, respectively, was because the sampling distribution is a normal distribution (per the Central Limit Theorem). That is, the z-value in this case is the distance between the sample mean (the “case” in the sampling distribution) and the population mean (“the mean of means”, the mean of the sampling distribution), expressed in standard errors (the “standard deviation” of the sampling distribution):

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

3

Now what about  $t$ ? By substituting the sample standard deviation for the population standard deviation, we end up with the *estimated* standard error. In turn, substituting the *estimated* standard error for the standard error in the formula for the z-value above, we get the  $t$ -value, the

3. where  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$ .

distance between the sample mean and the population mean, expressed in *estimated* standard errors:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

4

Compare the two formulas for the  $z$ -value and the  $t$ -value above. As similar as they look, the  $t$ -value is more “uncertain” than the  $z$ -value, and comes with the aforementioned specification of degrees of freedom. Given specific degrees of freedom, the shape of the  $t$ -distribution curve changes, and thus the probabilities associated with each  $t$ -value change too.

Finally, for the drum roll: The reason I was able to work with  $t$ -values instead of  $z$ -values in the calculations of confidence intervals in the previous section without acknowledging it is due to the sample sizes I chose for my examples. See, **the biggest difference between the  $z$  and the  $t$  happens with small  $N$  (especially  $N < 30$ ). The larger the  $N$ , the closer and closer the  $t$ -distribution approaches the  $z$ -distribution.**

You can see this in Figure 6.5 above: as the degrees of freedom increase, the shape of the distribution becomes more and more normal, so much so that the  $t$ -distribution at  $df=30$  is already rendered invisible in the figure, its light blue colour overridden by the normal distribution’s black. And from  $N=100$  on, **the  $t$  converges so**

4. Where  $s_{\bar{x}} = \frac{s}{\sqrt{N}}$ .

**fast to  $z$ , the  $t$ -distribution curve becomes** our old, familiar, beloved **normal curve!** (Okay, maybe “beloved” applies just to me.)

Given that in the confidence interval examples in the few preceding sections I used only large  $N$ 's ( $=900$  and above), the probabilities associated with the  $t$ -value at  $N-1$  degrees of freedom ( $=899$  and above) were the same as those associated with the  $z$ -values: 68% for  $t=z=1$ , 95% for  $t=z=1.96$ , 99% for  $t=z=2.58$ . (Hence I left them out of the discussion at that time to properly explain here.)

*Hmm, much ado about nothing*, I can imagine you saying at this point. If the  $t$ -distribution and the  $z$ -distribution are no different at larger  $N$ , why even bother with the  $t$  (beyond any small- $N$  uses)? And as unsatisfying the answer “I’ll explain later” is, I’m afraid I have no choice but to resort to it, again. Briefly, it has to do with something called a  *$t$ -test for significance* which we will be using soon enough for hypothesis testing in Chapter 7, next.

For now, what you should take away from this section is that **the  $t$ -distribution exists, and it is what we actually use for estimation (and not  $z$ !), given a specific sample size.** As well, remember that **for  $N=100$  and above,  $t$  converges to  $z$  so you can readily apply any probabilities you associate with  $z$  to  $t$  with  $N-1$   $df$ .** (Regarding the latter, **do not forget to always specify the degrees of freedom for whatever  $t$  you might have. A  $t$ -value *always* comes with  $df$  attached as it’s meaningless/undefined without them.**)



---

## 6.10 Summary [EMPTY]



---

# Chapter 7 Variables

## Associations

Statistical inference is hardly only a matter of estimating single variable means and proportions, and of constructing confidence intervals around them. Rather, quantitative sociologists (and other social scientists), like all scientists trying to explain the world around them, study *associations between variables*. Does class attendance affect students marks? Are male professors praised more highly in student evaluations than female professors? Are children of more educated parents more likely to earn post-secondary degrees? Does abstinence-only sex education lead to higher teen pregnancy (and abortion) rates? Are rich people more likely to vote? Are religious people more likely to espouse more socially conservative values? Does playing violent video-games increase incidence of real-life violence and crime? Does race/ethnicity affect one's educational attainment and/or income?

All of these questions reflect variable *associations*. Every time we hypothesize that two characteristics are related, or think that something *causes* change in another, every time we ask *why* something is the way it is and what makes it to be that way, we already speak the language of variable associations, even without acknowledging it as such.

While we can use various research methods to provide answers to these questions, quantitative analysis can shed a unique light on them due to its grounding in probability theory and the generalizability that stems from it. Of course, like any research method, using statistics for inference particularly in the social sciences has its problems and limitations. Thus, we have to be very careful in not overstating conclusions and to always qualify our findings based on the specific way we have operationalized our variables (i.e., exactly how we have measured a concept), as well as depending on our sample size, the statistical assumptions we've made, the uncertainty we're dealing with, etc., etc.

While most real-life research involves many variables at the same time, examining *multivariate* associations like that are beyond the scope of this book. Instead, in the remainder of this textbook I focus on *bivariate* associations — associations between two variables. Still, keep in mind that while this is a necessary first step when just entering the world of variable associations, this hardly ever (rather never) reflects reality in any way: the social world is too complex for there to only be one and *only one* cause to something we observe and that we're trying to explain. I'll remind you of this fact frequently as one of the biggest mistakes you could probably make with inference is to assume that the variable on which you have chosen to focus is the *only* one associated with (or worse, affecting) another variable of interest.

In short, from now on we work with two variables in

order to understand how associations work in principle, not because inference based on two variables reflects reality (neither in general, nor in real-life research).

The chapter starts with introducing what we mean by associations between variables, and with distinguishing between statistical and causal associations. In a brief return to descriptive statistics, you'll then learn how to describe bi-variate associations. At the end, I'll bring you back to the theory (and practice) of statistical inference, specifically to hypotheses and hypotheses testing, as this is again what allows us to move from sample descriptions to generalizable conclusions about the population of interest. Finally, I provide a brief discussion of the inevitability of uncertainty through introducing you to the two types of *errors of inference*.



---

## 7.1 Types of Bivariate Associations

To start with, what does it mean for two variables to be *associated*? Even without prior knowledge of statistics and statistics terminology, you likely have considered or at least noticed variables associations both during your studies and in general. For example, you probably know that fertility rates are higher in some countries and lower in others, and you might also know that the level of socioeconomic development also tends to differ between the two groups. You might also have noticed that, say, early childhood educators and hospital nurses tend to be women, while auto-mechanics or refrigerator repair technicians tend to be men. You certainly know that (for now) prime-ministers in Canada and presidents of the USA have tended to be white (and male, and Christian).

These of course are all examples of associations between variables. **Every time it can be noted that specific attributes of *one* variable tend to go or appear more often with certain attributes of *another* variable, you're looking at an association. That is, we're looking for a *pattern* between the sets of attributes of two variables; a pattern where some attribute combinations are seen more frequently while other attribute combinations are observed less often.**

Recall that we defined variables as characteristics that

vary across cases. Variables can vary *independently* of one another, or they can vary together — *in tandem*, as it were — in such a way that when some attributes of one variable are present, you'd expect to see some specific attributes of the other variable present too. Like so: Countries defined as *developed* tend to have lower fertility rates than countries defined as *developing*, so we have the variables *level of socioeconomic development* on the one hand, and *fertility rate* on the other. The association pits high levels of the former variable with low levels of the latter variable and vice versa — low levels of the former variable with high levels of the latter. These two combinations (high development/low fertility and low development/high fertility) are more likely to be observed than a no-pattern situation, where all sorts of combinations of development and fertility levels would be equally likely.

Similarly, research has repeatedly shown that some occupations tend to be male-dominated while others female-dominated. If there were no association (i.e., no pattern between the two sets of attributes), we would expect to observe approximately equal numbers of women and men in all occupations — but from what we have seen, that's not the case. That is, it seems there is an association between the variables *gender* and (choice of) *occupation*. Furthermore, participation in Canadian and US politics (and voters' preferences), especially at the highest levels of power, appears also to be gendered — as well as associated with other variables like *race/ethnicity* and *religious affiliation*.



*Do It! 7.1 Bivariate Associations*

Try to think of some other bivariate associations on your own. Start with something simple, like asking yourself if you commonly encounter some characteristic alongside a specific other characteristic; e.g., are dark-haired people more likely to have brown eyes while at the same time are blonde people more likely to have blue eyes? (Or, are the combinations dark hair/brown eyes and blond hair/blue eyes more common than dark hair/blue eyes and blond hair/brown eyes? Is hair colour related to — associated with — eye colour?) Etc.

Now that you are more familiar with the associations vocabulary, let's clarify the typology of variable associations. **There are two substantively different types of variable associations: *statistical* associations and *causal* associations.** Claiming a causal association between variables is stronger than the claim for statistical association. Further, **having a statistical association between two variables is a prerequisite for claiming a causal association between them — a prerequisite that is a *necessary but not sufficient* condition,** at that.

Statistical inference provides tests for establishing statistical association, to some basics of which I'll introduce you in the remaining chapters. Establishing causality, however, takes statistical associations as only but a starting point, as you will see in later on. **Statistical**

**associations are for the most part a *technical* matter — causality, on the other hand, is based on *logic*.** It involves one's ability to consider (and account for) multiple variables' associations at the same time.

When two variables vary together, we simply can say they are *associated*; however, when we claim causality, we call one variable *the cause* (or *predictor*) and the other *the effect* (or *outcome*).

In summary, finding if two variables are statistically associated (i.e., that some attributes of one of the variables tends to go with specific attributes of the other) is relatively easy. Claiming that one variable *affects* another (i.e., that changes in one variable produce/cause changes in the other variable), on the other hand, is not easy at all — rather, in the social world, it is quite difficult. But we'll get to that later.

For now, let's start with statistical associations and how to “find” them. To get there, first we need to take a brief trip to the (almost everyone's favourite) land of descriptive statistics in order to learn to even recognize potential statistical associations. We do that through bivariate description, i.e., by describing two variables together, considering them and their potential association at the same time.

---

## 7.2 Describing and Examining Bivariate Associations

Before we can get to establishing statistical associations between two variables, we need to know what we are looking for (or at), as it were. Social research, especially deductive reasoning, usually starts with an idea — a research question if you will — which is frequently grounded in an empirical observation of two variables' possible association (e.g., “Hey, it seems like all vegetarians/vegans I know tend to be well off. I wonder if income and vegetarianism/veganism are related...”) Then, if one is quantitatively inclined, a random sample can be used to “check” for such an association.

Most people conceive of that “check” as a one-step process but it actually involves two steps frequently undertaken in quick succession, so much so that to appear singular. As this is your introduction to the topic, we will take the steps slowly, one after the other.

**The first step is the descriptive part: given our sample data, does it *look like* there is an association between the two variables of interest?** This step concerns the data obtained through our sample, i.e., it describes *our sample*, and *only* our sample.

**The second step is the inferential part: assuming that it looks like there is an association between the two**

**variables of interest in the sample, is this association generalizable to the population?** That is, is this a “real” association reflecting the population or is it something we have observed in our sample due to the vagaries of random chance? This is the part where we formulate and test hypotheses in order to be able to make generalizable conclusions. We will focus on that starting with this chapter until the end of the book.

**We hereby start with the first step, describing bivariate associations** (again, based on sample data). What you need for this step is a recollection of the types of variables, and of the fact that we generally use both visual (graphical) and numerical descriptions.

From Chapter 3 (a *long* while back), recall that we *univariately* described a variable by 1) graphing its distribution (we used pie charts, bar graphs, and histograms, depending on level of measurement), and 2) providing numerical measures of central tendency and dispersion where applicable; this is how we used to “get a sense” of the variable and what it looked like. Similarly, we can also use graphical and numerical bivariate descriptives, this time depending on the combination of continuous-or-discrete variable type, to “get a sense” of the potential association between two variables and what it might look like.

Recall as well (from Section 1.5 (<https://pressbooks.bccampus.ca/simplestats/chapter/1-5-discrete-and-continuous-variables/>)), that we can classify variables as discrete and continuous<sup>1</sup> (I know,

1. Briefly, nominal and ordinal variables tend to be (but, especially the latter, are not always) *treated* as discrete, and interval/ratio variables tend to be (but

I know – it too has been awhile, but I did warn you eventually we'd get back to that).

**From this chapter on, we'll proceed by considering all three possible bivariate combinations of these: 1) associations between a *discrete* and a *continuous* variable<sup>2</sup>, 2) associations between *two discrete* variables<sup>3</sup>, and, finally, 3) associations between *two continuous* variables<sup>4</sup>.**

I discuss describing each of the three types of associations in the following subsections.

are not always) *treated* as continuous. Note, again, that social science data tends to be discrete -- we just treat some variables (with relatively large number of categories/values) as continuous. For the remainder of the text I will be referring to variables as discrete and continuous and you should take this to mean that that's how they are *treated* (and not as an indication of their "true nature").

2. We will soon learn to test this type of associations in one of the following sections and in Chapter 8.
3. We will learn to test this type of associations in Chapter 9.
4. We will learn to test this type of associations in Chapter 10



---

## 7.2.1 Between A Discrete and A Continuous Variable

**We can “get a sense” if a discrete and a continuous variable *seem* associated visually through a chart called a *boxplot* (discussed further below) and numerically through examining the *difference of means* (or medians, if one so prefers).**

What type of an association do we get when we consider a discrete and a continuous variable? The easiest way to represent this type of association is when we consider a binary (two-category) discrete variable and check if a continuous variable’s statistics (like the mean, or the median) vary between the discrete variable’s categories. This sounds far more complicated than it is. A couple of examples will show you that you have probably considered questions about “comparisons of means” even in your everyday life. The first one will explain it conceptually, the second with actual data.

*Example 7.1 Sex Differences in Upper Body Strength, American College Students*

Research has shown that, despite similar lower body

strength, women have less upper body strength than men, on average. [LIST CITATIONS FROM HERE <https://health.howstuffworks.com/wellness/diet-fitness/personal-training/men-vs-women-upper-body-strength.htm> AND THE FOLLOWING] One such study examined differences in upper body strength in a sample of Caucasian and East-Asian college students engaged in weight-lifting classes in American colleges (Chen, Liu and Yu, 2012) [<https://content.sciendo.com/view/journals/ssr/21/3-4/article-p153.xml>, pdf here <https://www.degruyter.com/downloadpdf/j/ssr.2012.xxi.issue-3-4/v10237-012-0015-5/v10237-012-0015-5.pdf>].

While the study examined numerous aspects of the difference in strength, I will take only one of the researchers' findings to illustrate my point: triceps strength in arm extension. The reported means were 46.2 pounds for women versus 87.4 pounds for men in the Caucasian sample, and 39.6 pounds for women versus 82.1 pounds for men in the East-Asian sample (Chen, Liu and Yu, 2012, p.156).

Consider what we are discussing here: We have two variables of interest<sup>1</sup>, *gender* and *upper-body strength*. *Gender* is a nominal discrete (and, in this study, binary) variable while *upper-body strength* (through various measurements in pounds) is a ratio continuous variable. The hypothesized association between the two posits that some categories of the discrete variable (e.g., men) tend to go with specific values of the continuous variable (e.g., higher values on upper body-strength). That is, if both men and

1. You could argue that *race/ethnicity* is also there. As reported in the study, however, *race/ethnicity* was a secondary variable bringing more detail to the study, through which the authors were able to demonstrate that upper-body strength differences based on sex exist in both race/ethnic groups considered.



women had the same means for, in this case, triceps strength in arm extension, *gender* and *upper-body strength* would be unrelated, as one's sex wouldn't be predictive of one's upper-body strength at all.

In effect, we are comparing the mean values (of a continuous variable) across groups (i.e., the categories of a discrete variable). Now, as far as a numerical description of that comparison goes, we have the two means (of men and of women) and we can thus calculate the difference of means:

$$\bar{x}_{men} - \bar{x}_{women} = 87.4 - 46.2 = 41.2$$

(Caucasian sub-sample)

$$\bar{x}_{men} - \bar{x}_{women} = 82.1 - 39.6 = 42.5$$

(East-Asian sub-sample)

Thus, what we observe *in this sample* is a 41.2 pounds difference in the upper-body strength (as measured by triceps strength in arm extension) between Caucasian men and women and a difference in upper-body strength of 42.5 pounds between East-Asian men and women. Again, note that **the fact that we see these differences in the sample does not mean they exist in the population — they may, or they might not. We wouldn't know this unless we test if the differences are generalizable to the population<sup>2</sup>.** We will get to testing later, for now we are only interested in the differences *descriptively*, i.e., that they exist *in the sample*.

2. If you are interested, the authors of the study did test these differences (with a *t*-test, discussed later) and found them generalizable to the population indeed (Chen, Liu and Yo, 2012).

Example 7.1 above shows that every time we compare averages of two (or indeed, more than two) groups and calculate the differences in the means, we are effectively describing associations between variables. I could have easily presented other examples like gender or race/ethnic differences in annual income<sup>3</sup>, years of education, occupational prestige, test scores<sup>3</sup>, etc., etc. The reason I chose an example about a sex-based rather than gender-based difference (that is, a kinesiological rather than a sociological study) was so that I can warn you in passing about a common mistake, called the *ecological fallacy*.

**Watch Out!! #12 . . . for The Ecological Fallacy**

Consider the findings from the study in Example 7.1 above: men's average upper-body strength is higher than women's. Assuming we can generalize the findings to the general population<sup>4</sup>, the evidence suggests that when it comes to upper-body strength men are stronger than women *on average*. Many people take this to mean that a randomly selected man would be *always* stronger than a randomly

3. For an example of a brief study on the association between *race/ethnicity* (a five-category discrete variable) and SAT scores of Harvard University applicants, see here: <https://www.thecrimson.com/article/2018/10/22/asian-american-admit-sat-scores/>.
4. As mentioned above, many studies support this as a real, physiological sex difference; the reason I chose this example instead of more controversial/debated issues like gender differences in IQ, or the gender pay gap, etc.

selected woman . . . which does not follow at all from the difference in mean strength.

Statistically speaking, it is a matter of the dispersion around the means of the two groups, and of how big the difference in means is. It is quite possible for a lot of women to have more upper-body strength than the men's average, as well as that a lot of men to have less upper-body strength than the women's average.

Ultimately, the takeaway from this caveat is to not over-interpret differences in averages to mean more than what they actually are: differences in *averaged* values, not of the specific values of *individuals* belonging to the different groups that are compared. (You can find an excellent account of how common this ecological-fallacy mistake is here: <https://www.americanscientist.org/article/what-everyone-should-know-about-statistical-correlation>.)

With that warning out of the way, let's take another (this time, sociologically motivated) example for examining differences of means, along with a proper visual description — boxplots.

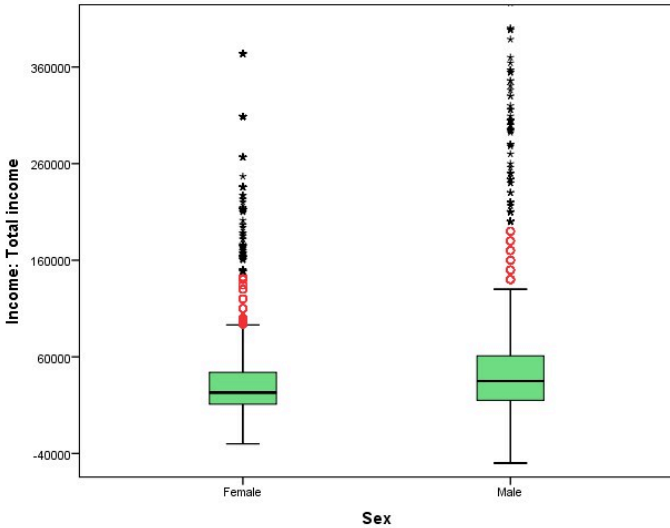
*Example 7.2 Gender Differences in Total Income, NHS 2011*

Statistics Canada's *National Household Survey 2011* (NHS 2011) was designed to replace the until-that-time mandatory long form of the Census<sup>5</sup>. For this example, I use a random sample of about 3 percent of the *NHS 2011* individual data (aka a *Public Use Microdata File*, or PUMF), resulting in  $N=22,123$ . I am interested in whether men and women's income for the year preceding the survey differed, i.e., whether the variables *gender* (called *sex* in the dataset) and *total income* (i.e., income from all possible sources) appear associated.

With the help of SPSS, I plot the data. The resulting boxplots graph is given in Figure 7.1 below.

*Figure 7.1 Gender Differences in Total Income, NHS 2011*

5. For the problematic nature of the (Harper) Government's decision in 2010 to make the survey voluntary and its related implications, see for example here: <https://ocul.on.ca/node/3400>. The mandatory long-form census was restored in 2016 by the Liberal Government. My usage of the data here is strictly for demonstration purposes and as such shouldn't be taken as an endorsement of the NHS 2011.



Boxplots are charts visually incorporating a lot of statistical information in one neat little package; I encourage you to make use of them when exploring your data as they can be quite useful. What do we see in Figure 7.1 in our case? Obviously, we have two groups to compare (as per the two categories in the nominal variable gender), *women* and *men*, and therefore the graph presents two boxplots. (Had we multiple categories in our discrete variable, we'd have had multiple boxplots.)

**How to read a boxplot.** Each boxplot consists of the eponymous “box” and two so-called “whiskers” protruding from it. The “box” (in green above) represents the middle 50 percent of the data (i.e., the two middle quartiles, or

the IQR); the lower “whisker” represents the first/bottom quartile of the data, and the upper “whisker” represents the last/top quartile of the data. The dark line bisecting the box indicates the median. The two ends of the “whiskers” are the lowest and the highest values. Note, however, that the quartiles (as represented by the “whiskers”) exclude outliers as to not visually distort the “regular” spread of the data. As such, the chart plots run-of-the-mill outlier cases as small circles (above they are in red) outside of the “whiskers”; extreme outliers are indicated by stars (in black above)<sup>6</sup>.

Now that you know how to read them, compare the two boxplots above. First, we see that the median for men is higher than the median for women (again, these are the dark lines within the boxes); as well, total income appears to be more spread out for men than for women (the “whiskers” in the men’s boxplot reach further, indicating larger range and IQR. Further, while both men and women appear to have outliers, the men’s group seems to include more extreme outliers and at higher values than those observed in the women’s group<sup>7</sup>.

6. Also note that to make the boxplot readable in an appropriate size, in Figure 7.1 I cut some *extremely* extreme outliers off at the top of the men’s boxplot.
7. You might have noticed that the first quartile (i.e., the bottom “whisker”) includes negative values. Statistics Canada uses several income variables for which this is the case. Negative income exists as an accounting possibility: when one’s annual expenses end up larger than one’s annual income (e.g., like for a self-employed individual whose business hasn’t been as successful, etc.). In many real-life research, negative income values are frequently dropped/removed if that course of action is justified by the study’s design, research question, and purposes. In the case of this example

All this points to the conclusion that men in the sample had higher (median, and quite likely average) total income for 2010 than women did, despite that the individuals with the lowest incomes also appear to be men.

As useful the general information we gleaned from the boxplots, we should look at the precise numbers too. SPSS calculates the mean total income as \$32,465 for women and \$48,866 for men — that is, there is \$16,401 difference in mean total income in favour of men. In this sample of 22,123 people, men's average total income is \$16,401 more than women's.

We could also compare the medians (especially useful when dealing with income variables): SPSS gives the median total income of women in the sample as \$23,000, while the median total income for men is \$35,000 — a difference of medians of \$12,000, again in favour of men.

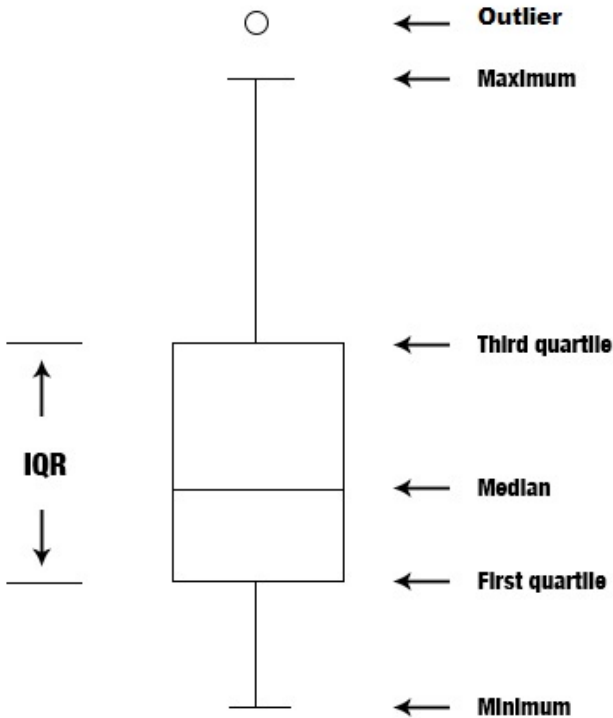
To summarize, **you can explore a potential association between a discrete and a continuous variables of interest in two ways: 1) visually — by plotting and comparing boxplots, and 2) numerically, by inspecting the means (or medians) for the groups (i.e., the categories in the discrete variable being compared) and reporting their difference.**

I have no reason to do that, hence I left the negative income values in the data.

Keep in mind that we are not estimating anything at this point and are not claiming anything about the population: we are simply describing data based on a specific, actual sample.

Figure 7.2 below shows a quick reference for interpreting boxplots.

*Figure 7.2 How to Interpret a Boxplot*



[Source: [https://commons.wikimedia.org/wiki/File:Box\\_plot\\_description.jpg](https://commons.wikimedia.org/wiki/File:Box_plot_description.jpg)]



*SPSS Tip 7.1 Bivariate Descriptions of Discrete and Continuous Variables: Boxplots and Comparisons of Means*

This is **how you can get boxplots** like the ones in Figure 7.1 above:

- From the *Main Menu*, select *Graphs*, then from the pull-down menu *Legacy Dialogues*, and finally *Boxplot*;
- In the resulting *Boxplot* window select *Simple* and, keeping *Summaries of groups of cases* checked, click *Define*;
- Select your continuous variable of interest from the list on the left and, using the appropriate arrow, move it into the *Variable* empty space on the right (at the top);
- Select your discrete variable of interest from the list on the left and, using the appropriate arrow, move it into the *Category Axis* empty space on the right (below the *Variable*), then click *OK*;
- Your boxplots will appear in the *Output* window. (Note that the graph will appear in its default SPSS colours and specifications. Double-clicking the chart will make a *Chart Editor* window appear. In the *Chart Editor* you can change, edit, and modify the appearance of your boxplots to your heart's content.)

This is **how you can get means, medians (or any descriptive statistic really) for different groups**:

- From the *Main Menu*, select *Data* and then

from the pull-down menu, select *Split File*;

- In the new window, select *Compare groups*, then find your discrete variable of interest from the left-hand side, and using the arrow, move it into the *Groups Based on* empty space; click *OK*.
- You would have just placed a filter on your data. From this point on (until you switch the filter off), everything you do in SPSS will be done for each separate group (this is indicated by a message “SORT CASES BY [your discrete variable name]. SPLIT FILE LAYERED BY [your discrete variable name].” appearing in the *Output* window.
- Then, from the *Main Menu*, select *Analyze*, and then *Frequencies*, etc. to request any descriptive statistics you may like, e.g., the mean, the median, the standard deviation, etc. as discussed in the SPSS Tips in Chapter 3.
- Your output in the *Output* window will list the requested descriptives by the different groups (categories of the discrete variable).
- Once you are done with the comparisons, do not forget to switch the filter off (or your data file will remain split by groups): go again to *Data* in the *Main Menu*, select *Split File* and click *Analyze all cases, do not create groups* on the right-hand side; click *OK*.
- Your *Output* window will give a message of “SPLIT FILE OFF.” to indicate that the data is no longer split by group and it’s in its original condition.

Now let's see how to “spot” and describe potential associations between two discrete variables.



---

## 7.2.2 Between Two Discrete Variables

Examining a potential statistical association between two discrete variables amounts to comparing groups (as per the categories of one of the variables) on the number (and proportion) of their respective members that fall in the categories of the other variable<sup>1</sup>. Again, this sounds far worse than it actually is, as you will see in the examples that follow.

The potential association between discrete variables can be examined both visually and numerically via a special table called *cross-tabulation* table (“cross-table” or “crosstab” for short) or *contingency* table. While a contingency table can have any number of rows and columns, *too* large a number of either/or both can easily make the table unreadable as it would contain too much data to contemplate at once. (This is also the reason why we chose to treat some variables as continuous — when they have too many categories — as then we can use another tool to visualize and examine them, as we will see later.) Thus, below I introduce the simplest form of a contingency table, a 2×2 crosstab (i.e., 2 rows and 2 columns).

In the general sense a  $K \times J$  cross-table would be a table

1. Since both variables are discrete, for clarity's sake I refer to the attributes of one variable as *groups* and to the attributes of the other variable as *categories*. (I have used them interchangeably until now but here it helps to distinguish the two variables by using two different words for their attributes.)

containing  $K$  rows and  $J$  columns, where the categories of one variable go into the rows (a  $K$  number of them) and the categories of the second variable (a  $J$  number of them) go into the columns of the table (therefore *crossing* in the interior cells of the table).

Thus, a  $2 \times 2$  contingency table would mean we have two binary variables, each with two categories. Before I show you an actual data exploration, Table 7.1 presents an “empty shell” of one such table which I use to introduce some needed vocabulary.

*Table 7.1 A Generic Cross-tabulation Table*

	Variable 1 Group 1	Variable 1 Group 2	Total
Variable 2 Category 1	Number A	Number B	Category 1 Total (A+B)
Variable 2 Category 2	Number C	Number D	Category 2 Total (C+D)
Total	Group 1 Total (A+C)	Group 2 Total (B+D)	Total All (A+B+C+D)

The first thing you should notice is that the  $K \times J$ , or the  $2 \times 2$  in our case, refers to the groups/categories of the variables in questions, not to the actual number of rows and columns in the contingency table. Technically speaking, Table 7.1 contains four rows and four columns — but the ones that count are only the ones in green: two “green” rows and two “green” columns, indicating the number of groups and categories of the variables. The last row and

the last column (in blue above) are called *margins* and are reserved for reporting totals<sup>2</sup>. The first column and the first row (in bold above) are simply titles.

The central cells of the table are the most important ones. In the example above, *Number A* indicates the number of cases (observations/individuals/etc.) that belong simultaneously to Group 1 (of the first variable) and Category 1 (of the second variable). By analogy, *Number B* indicates the number of cases that belong simultaneously to Group 2 and Category 1; *Number C* stands for the number of cases that belong to both Group 1 and Category 2; and finally, *Number D* is the number of cases that belong to both Group 2 and Category 2.

The margins contain the totals by row and by column, and the last cell (last row, last column) is reserved for the total *N*.

*So what is so special about this table? I've seen such tables all my life!* you might be saying right about now. Bear with me, we'll eventually get to the special — and somewhat complicated — part (and likely you'll be sorry for it). First though, let's look at a contingency table with some actual numbers.

### Example 7.3 Do You Like The Campus Cafeteria?

2. The more observant of you may notice that the *horizontal margin* (the last row) shows the frequency distribution of Variable 1 (i.e., the number of cases per group), while the *vertical margin* (the last column) shows the frequency distribution of Variable 2 (the number of cases per category).

Imagine you are frustrated within the food options available in your campus cafeteria and you wonder if others share your thoughts on the matter (perhaps in order to gauge support for changes you'd like to see enacted, or similar type of activism). Before you devote time to do an actual random sample study (now that you know), you do a quick exploratory poll of your classmates in one of your classes. You ask 35 people whether they like the campus cafeteria, and in the process, you get the inkling that second-year students seem to have different opinion about the food options than the first-year students in the class. You plot your results:

*Table 7.1(A) Do You Like The Campus Cafeteria?*

	First Year Students	Second Year Students	Total
<b>YES</b>	7	5	12
<b>NO</b>	8	15	23
<b>Total</b>	15	20	35

I am certain you know how to read this: 7 first-year and 5 second-year students like the cafeteria, while 8 first-year and 15 second-year students do not. There is a total of 12<sup>3</sup> students who like the cafeteria and 23<sup>4</sup> students who do not.

3. As  $7+5=12$ .

4. As  $8+15=23$ .



You talked to 15<sup>5</sup> first-year and 20<sup>6</sup> second-year students, a total of 35 students.

Can you compare the relevant numbers as they are presented in the table? And, for that matter, what *are* the relevant numbers?

Let's answer both questions in turn.

Recall from Chapter 2: No, you cannot compare the numbers as stated since the two groups you have to compare are of different size. The relevant comparison is between the different year students who like the cafeteria — first-years vs. second-years — as this is what you want to know.

It's true that 2 more first-years like the food in the cafeteria than the second year students ( $7 > 5$ ) but at the same time you had 5 more second-year students in your sample ( $20 > 15$ ). To take into account the differing group size, you need to compare proportions (or percentages): the proportion of first-year students who like the cafeteria against the proportion of second-year students who like the cafeteria. You therefore calculate the respective proportions, turning them into percentages at the end:

- $\frac{7}{15} = 0.467$ , or 46.7% of first-years like the cafeteria
- $\frac{5}{20} = 0.250$ , or 25% of second-years like the

5. As  $7+8=15$ .

6. As  $5+15=20$ .

cafeteria

- $\frac{8}{15} = 0.533$ , or 53.3% of first years do NOT like the cafeteria
- $\frac{15}{20} = 0.750$ , or 75% of the second-years do NOT like the cafeteria
- $\frac{12}{35} = 0.342$ , or 34.2% of ALL students like the cafeteria
- $\frac{23}{35} = 0.718$ , or 71.8% of ALL students do NOT like the cafeteria

To summarize the information neatly, we modify our table to this:

*Table 7.1(B) Do You Like The Campus Cafeteria?  
(Column Percentages)*

	First Year Students	Second Year Students	Total
YES	46.7%	25%	34.3%
NO	53.3%	75%	71.8%
Total	100%	100%	100%

So far so good? From Table 7.2(B) now we clearly see that your initial hunch was right: there does seem to be a difference in the opinions of your classmates based on which year they are in their studies. That is, while you do

have support for anti-cafeteria activism (only 34.3% of your classmates like the campus cafeteria, while 71.8% dislike it) first-year students seem to like the cafeteria a lot (almost twice) more than second-year students do: 46.7% of first-years like the food options in the cafeteria compared to only 25% of the second-years, a difference of 21.7 percentage points.

The example above shows **what you need to examine the possible association between two discrete variables: a *cross-tabulation* (listing percentages, not absolute numbers!), visually, and a *difference in proportions* (or percentages), numerically.**

Again, a reminder that this is sample-only exploration. We make no predictions or inferences about a population, we just explore what the data we have at hand shows.

So far, I purposefully show you how the logic of the descriptive analysis of contingency table goes, *the right way*. Here comes the complication, however: why did I calculate the proportions in the example the way I did? Consider the alternative:

- $\frac{7}{12} = 0.583$ , or 58.3% of the students who like the cafeteria are first-years
- $\frac{5}{12} = 0.417$ , or 41.7% of the students who like the cafeteria are second-years
- $\frac{8}{23} = 0.348$ , or 34.8% of students who do NOT

like the cafeteria are first-years

- $\frac{15}{23} = 0.652$ , or 65.2% of students who do NOT like the cafeteria are second-years
- $\frac{15}{35} = 0.429$ , or 42.9% of ALL students are first-years
- $\frac{20}{35} = 0.571$ , or 57.1% of ALL students are second-years

Table 7.1(C) below demonstrates this alternative.

*Table 7.1(C) Do You Like The Campus Cafeteria? (Row Percentages)*

	First Year Students	Second Year Students	Total
YES	58.3%	41.7%	100%
NO	34.8%	65.2%	100%
Total	42.9%	57.1%	100%

Table 7.1(B) and Table 7.1(C) contain two different sets of percentages. The percentages in Table 7.1(B) are called *column* percentages, while the percentages in Table 7.1(C) are called *row* percentages. **Column percentages are calculated “down the columns” (i.e., the proportions are based on the numbers on the horizontal margin/ last row, which in turns lists “100%” in each column). Row percentages are calculated “right/across the rows” (i.e., proportions are based on the vertical**

**margin/last column, which in turn lists “100%”).**

Why didn’t we use Table 7.1(C) in the example above?

The answer is in the warning box below.

**Watch Out!! #13.** . . . *for Choosing The Wrong Percentages in Contingency Tables*

The complication regarding choosing the “right” percentage arises due to the fact that what is considered the “right” or the “wrong” percentage depends on what you actually want to know, as in, what your research question/question of interest is. The percentages in Table 7.1(C) are “wrong” only because they are not helpful to answer the question whether there is a difference in the two groups of students we compare, first-years and second-years. Had we been comparing the YES group and the NO group on how many first-year students they each contained, we’d have used Table 7.1(C). However, this doesn’t seem like the most relevant question we could ask in *this* hypothetical study.

Unfortunately, that’s not all. If you thought *OK, then, I’ll just always use column percentages and be done with it*, you’d have been too hasty. You see, **the “correctness” of the percentages you need depends on where your compared-groups variable is placed.** In Table 7.1(B) I placed the groups-to-be-compared (first-years vs. second-years) in the columns, and therefore I calculated the column percentages. If I had put the groups-to-be-compared in the

rows, I would have calculated the row percentages (which would have resulted in a transposed Table 7.1(B))<sup>7</sup>

	YES	NO	Total
First Year Students	46.7%	53.3	100%
Second Year Students	25%	75%	100%
Total	34.3%	71.8%	100%

Many students faced with contingency tables have trouble deciding whether they need column or row percentages. My advice is (which you can take as a rule of thumb) to be clear what groups you compare based on your question: **if you compare the groups in the columns, you need column percentages; if you compare the groups in the rows, you need row percentages.** (This is also the reason why I labeled Variable 2's attributes as *categories* early in this section, not to confuse them with the Variable 1's *groups*.)

Another rule of thumb you might find useful: try to always put your groups-to-be-compared in the columns (as most people find comparing a left column to a right column, horizontally, easier), then you'll always need column percentages. That said, do not assume that everyone follows this last advice: sometimes you might find a table where the relevant comparison is top row to bottom row, vertically.

7. Table 7.1(D) Do You Like The Campus Cafeteria? (Transposed -- and still correct)

To orient yourself in the organization of the table, look for which margin contains the “100%”s — if it’s the horizontal margin (bottom row), you’re dealing with column percentages, if it’s the vertical margin (last column), you’re dealing with row percentages.

Finally, **never try to “compare” the percentages that add to 100%** (be they in the rows or in the columns) as this would not constitute a comparison at all — instead, it would be a breakdown of the groups in terms of composition (that’s why they’d add up to 100%, like the 25% of second-years who liked the cafeteria and the 75% who did not in Table 7.1(B) above). Again, **what you need to compare is always the fraction of cases from one group falling in a category of interest to the fraction of cases from the other group in the same category of interest.**

All of this is arguably complicated at first blush. The light at the end of the tunnel is that the more you work with contingency tables, the easier you will find constructing them and/or interpreting them correctly.

To that effect, let’s take an example with real existing data.

*Example 7.4 Gender Differences in the Speaking Aboriginal Language Ability among Indigenous Canadians , APS 2012*

Statistics Canada’s *Aboriginal People Survey (APS) 2012* is a nationally representative survey of First Nations peoples (living off reserve), Métis and Inuit, 6 years of age and older (Statistics Canada, 2019)<sup>8</sup>. Language is a key element in retaining, preserving, and transmitting culture; as such, the ability of Indigenous peoples to speak their ancestral languages is of special interest given the recommendations of the Truth and Reconciliation Commission’s (TRC) final report (2015) [REFERENCE].

For the purposes of this example, I am interested in if there are gender differences in the ability to speak an Aboriginal Language among the collected sample. Table 7.2 shows the cross-tabulation of *gender* (called *sex* in the *APS 2012*) and *speaking Aboriginal language* variables. (Both variables are binary in the survey.)

Table 7.2(A) *Speaking Aboriginal Language Ability by Gender, APS 2012*

Lang. - Speaking Aboriginal language * Sex of respondent Crosstabulation				
Count		Sex of respondent		Total
		MALE	FEMALE	
Lang. - Speaking Aboriginal language	Yes	4877	5672	10549
	No	6898	6933	13831
Total		11775	12605	24380

8. One could perhaps see the *APS 2012* as an effort by Statistics Canada to address some of the voluntary *NHS 2011*’s issues with coverage/non-response of the listed population groups.



As you can see, working with real, large  $N$  data makes proportions even more indispensable for making sense of the table. We need to compare the fraction of women who speak an Aboriginal language (or languages) to the fraction of men who are able to do that.

To make things easier, I followed the rules of thumb I listed in the *Watch Out!!* 13 above: the groups-to-be-compared are in the columns, and we need to compare them horizontally<sup>9</sup>. Therefore, I need column percentages. Table 7.2(B) does just that.

*Table 7.2(B) Speaking Aboriginal Language Ability by Gender, APS 2012 (Column Percentages)*

Lang. - Speaking Aboriginal language * Sex of respondent Crosstabulation					
			Sex of respondent		
			MALE	FEMALE	Total
Lang. - Speaking Aboriginal language	Yes	Count	4877	5672	10549
		% within Sex of respondent	41.4%	45.0%	43.3%
	No	Count	6898	6933	13831
		% within Sex of respondent	58.6%	55.0%	56.7%
Total	Count		11775	12605	24380
	% within Sex of respondent		100.0%	100.0%	100.0%

9. A point to be made here is that when working with binary data, it's enough to focus on one of the categories on which you compare the groups, as the other category would be a complement of the first as we are working with proportions. That is, here we need consider only the YES category (as the NO category is it's exact opposite, i.e., "1- YES") due to the fact that we're interested in those who can speak the language, not those who don't.

SPSS provides both the original cell count (i.e., frequency) and the respective percentage below it<sup>10</sup>.

We can thus easily see that while only 41.4% of men in the sample can speak an Aboriginal language, 45% of women in the sample can do that; i.e., there is a gender difference of 3.6 percentage points in favour of women<sup>11</sup>.

So far, we discussed only  $2 \times 2$  contingency tables, i.e., binary variables. Of course, discrete variables can have more than two categories each. In the case of a  $2 \times 3$  table (and assuming our groups to-be-compared are in the columns), we'd simply have three groups/proportions to compare. In the case of  $2 \times J$ , where  $J > 3$ , we'd have  $J$  groups/proportions to compare. The proportions can be compared in two ways: one against the remaining ones together (through one difference of proportions), or each compared to each of the remaining ones (through several difference of proportions).

Matters become more complicated when we let go of binary variables altogether and have  $K \times J$  table where both  $K > 2$  and  $J > 2$  instead. This type of table can be visually complicated, the larger the  $K$  and  $J$ . However, the

10. Yet another useful rule of thumb: make sure that SPSS lists "% within [groups-to-be-compared]" as this indicates that the correct percentages appear in the table. In this case, SPSS tells us that it has listed "% within Sex of respondent", i.e., the ones we need in order to compare the two gender groups.
11. You might think it a small difference, but the magnitude of the difference is not the most important thing when establishing statistical associations. More on the topic in below.

comparison can still be done between groups on a category of interest in the manner described above. For a brief illustration, see Table 7.3 below.

*Table 7.3 Marital Status Differences in Perceived Health, CCHS 2016*

Perceived health * Marital status Crosstabulation			Marital status				Total
			Married	Common-law	Widowed/Divorced/ Separated	Single	
Perceived health	Excellent	Count	9967	2549	3478	7952	23946
		% within Marital status	22.2%	23.8%	15.9%	25.1%	21.9%
	Very good	Count	16913	4083	6861	11738	39595
		% within Marital status	37.7%	38.2%	31.4%	37.0%	36.3%
	Good	Count	12606	2971	6780	8648	31005
		% within Marital status	28.1%	27.8%	31.0%	27.3%	28.4%
	Fair	Count	3903	843	3316	2559	10621
		% within Marital status	8.7%	7.9%	15.2%	8.1%	9.7%
	Poor	Count	1485	242	1402	830	3959
		% within Marital status	3.3%	2.3%	6.4%	2.6%	3.6%
	Total	Count	44874	10688	21837	31727	109126
		% within Marital status	100.0%	100.0%	100.0%	100.0%	100.0%

Table 7.3 is a 5×4 table and it presents data from Statistics Canada's *Canadian Community Health Survey (CCHS) 2015-2016*, crosstabulating *marital status* (in 4 groups) and *perceived health* (in 5 categories). Considering that the latter is an ordinal variable, a way to mentally simplify the presented information is to focus on the

extremes — the proportions of people in the different marital status groups who reported excellent or poor health.

A quick examination of the relevant percentages reveals that fewer widowed/divorced/separated respondents appear to report their health as excellent than any of the other groups (15.9% vs. 22.2%, 23.8%, and 25.1% of married, common-law, and single respondents, respectively) — a difference of 6.3 percentage points at the minimum in favour of the other groups. Correspondingly, widowed/separated/divorced respondents also report their health as poor more often than the other marital status groups (6.4% vs. 3.3%, 2.3%, and 2.6% for married, common-law, and single individuals, respectively) — a difference of 3.1 percentage points at the minimum in favour (or rather, *disfavour*) of the widowed/married/separated group.

As such, it appears that while the other groups do not seem to differ much on their self-reported health, the widowed/separated/divorced group stands out by reporting lower levels of health, an observation consistent through all five health categories, an indication that the variables *marital status* and *perceived health* could be associated.

This concludes my presentation on how to analyze contingency tables data for possible discrete variable associations; the only thing left is to tell you how to produce a table with SPSS.

- From the *Main Menu*, select *Analyze*, and then from the pull-down menu, *Descriptive Statistics* and then *Crosstabs*;
- Select your pair of discrete variables of interest from the list on the left-hand side, and, using the appropriate arrows, move each to their respective slot on the right: *Row(s)* or *Column(s)*;
- Click on the *Cells* button in the top right corner<sup>12</sup>; in the resultant window select *Observed* in *Counts*, and either *Row* **or** *Column* in *Percentages*<sup>13</sup>, depending on where you put your groups-to-be-compared, and click *Continue*;
- Once back at the original window, click *OK*.
- The *Output* window will show the contingency table of the variables you selected.

So far, we have seen how we examine potential bivariate associations between a discrete and a continuous variable (previous Section 7.2.1) and between two discrete variable

12. This is important as if you fail to click on *Cells* and just click *OK* at the bottom, SPSS will produce a table with only the observed count (i.e., number of elements in each cell) which will make comparison between the groups impossible. Clicking *Cells* allows you to choose which percentages you want calculated and included in the table.
13. Avoid selecting both, and even more so, avoid selecting all three options (*Row*, *Column*, and *Total*). I guarantee you wouldn't want to interpret the resulting table should you choose more than one set of percentages. Again, be careful to request the percentages for the place, rows or columns, where you put your groups-to-be-compared. If they are in the rows, select *Row* in *Percentages*; if they are in the columns, select *Column* in *Percentages*.

(presently). We now turn to the last bivariate combination, between two continuous variables, next.

---

## 7.2.3 Between Two Continuous Variables

The distinctive feature of continuous variables is their large number of values. As discussed previously, typically we treat most interval/ratio variables as continuous. However, sometimes ordinal variables too can have a number of categories, large enough to justify their treatment as continuous for the purposes of statistical analysis. (Think back to the previous Section 7.2.2 and imagine crosstabulating a variable with, say, 10+ categories on another; the resulting table will be too unwieldy for meaningful examination.)

As well, continuous variables have values of different magnitudes, which can be ordered from low to high. Thus, what we will be looking for when examining two such variables for a possible association is whether a pattern exists between their values, or, alternatively, if their values do not exhibit any predictable combination. While many types of patterns can exist, for the purposes of this introductory text we'll focus on the two simplest ones: a *positive linear* association and a *negative linear* association. The way we describe and examine such associations is visually through a graph called a *scatterplot* and numerically through a special indicator called *Pearson's correlation coefficient  $r$*  (or *Pearson's  $r$* , or just  *$r$* ). I explain both below.

**A positive linear association is a pattern in which low values of one variable go with low values of the other variable alongside with high values of the former going with high values of the latter.** That is, in a positive linear association when the values of Variable 1 increase or decrease, so do the values of Variable 2. As its name suggests, **a negative linear association is the exact opposite: low values of one variable go with high values of the other variable and vice versa.** Then, as the values of Variable 1 *increase*, the values of Variable 2 will tend to *decrease*, or vice versa.

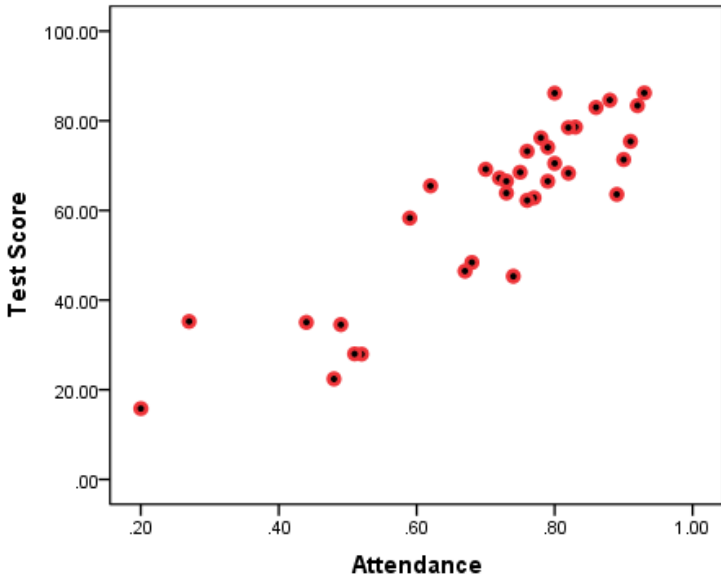
Both the positive and the negative version of this pattern are called *linear* because plotting the values of the two variables on a coordinate system shows the data points “congregating” in an approximately “straight” fashion, as if along an imaginary straight line with an upward (i.e., positive) or downward (i.e., negative) slope<sup>1</sup>.

Consider the following example two figures.

*Figure 7.3(A) Positive Association: Test Scores by Class Attendance (Simulated Data<sup>2</sup>)*

1. Other than linear associations exists, e.g., *curvilinear* (imagine U-shaped or inverted U-shaped *curves* in the data, instead of a straight line). Analyzing these is more complicated and beyond the scope of this book. The discussion hereafter will consider only bivariate linear associations, regardless if I mention it explicitly or not.
2. The simulated data used here for illustration purposes only is provided by DataBake ([www.databake.io](http://www.databake.io)). [see terms of use 3.6, 3.7: (free) datasets can be copied, modified, stored or otherwise used for your own personal, academic, or internal business purposes"]





In the **scatterplot** in Figure 7.3(A) above, I have plotted data from 35 imaginary students on their class attendance and subsequent final test scores<sup>3</sup>. Both *class attendance* and *test scores* are continuous variables. (Attendance is a ratio variable measuring proportion of the class time attended while test scores is an interval variable measured in percentages.) Each point of the data represents *simultaneously* a student's attendance (on the horizontal axis) and a student's test score (on the vertical axis); e.g., the lowest/left-most data point stands for a student who attended about 20% of class time and scored less than 20% on the final exam. The data points look “scattered” all over the graph, hence the name *scatterplot*.

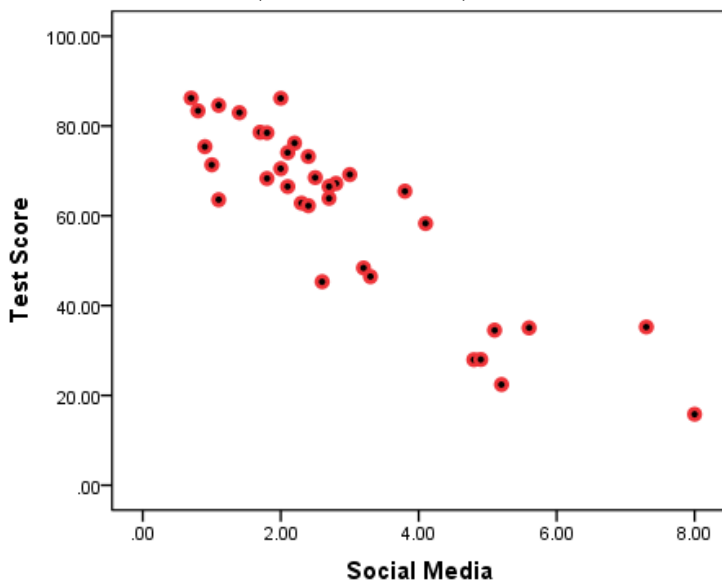
You can easily see the pattern in the data in Figure

3. The data is called *simulated* as it's computer-generated for the purposes of the exercise.

7.3(A): lower attendance seems to go with lower test cores, and higher attendance with higher scores. The bottom right side (high attendance/low scores) and the top left side (low attendance/high scores) of the graph are empty: there seem to be no students who attended classes a lot but scored low on the test, nor students who didn't attend much but scored high on the test. Had there been no pattern, the data points would spread all over the graph, identifying no clear "congregation" of values based on their magnitude.

Since class attendance and test scores seem to go *concordantly* "together" (i.e., low/low and high/high), we have indication of a *positive* association.

*Figure 7.4(A) Negative Association: Test Scores by Time Spent On Social Media (Simulated Data)*



Again, both *time on social media* and *test scores* are

continuous variables, with time on social media measured in average hours per day.

The pattern in Figure 7.3(A) is the opposite of the one we had before: lower number of hours spent on social media seem to go with higher test cores, and higher social media usage with lower scores. This time, the bottom left side (low on social media/low scores) and the top right side (high on social media/high scores) of the graph are empty: there seem to be no students who spent very little time on social media but scored low on the test nor students who had high usage of social media but scored high on the test.

Since social media usage and test scores seem to go *discordantly* “together” (i.e., low/high and high/low), here we have an indication of a *negative* association.

Figure 7.3(B) and Figure 7.4(B) below make the point about linearity clearer by adding something called a *line of best fit* to the original graphs<sup>4</sup>. **The slope of the line indicates the nature of the supposed association: upward/positive or downward/negative.**

Figure 7.3(B) *Positive Association: Test Scores by Class Attendance With Line of Best Fit*

4. We discuss the line of best fit (aka regression line) in Chapter 10 later.

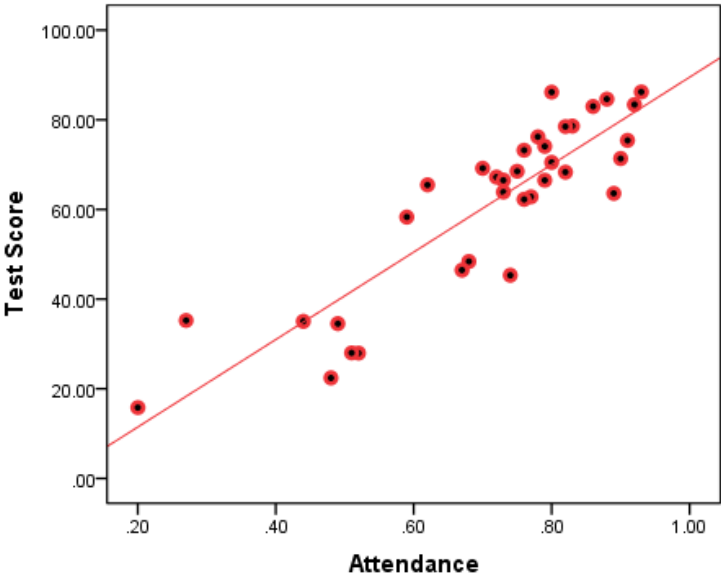
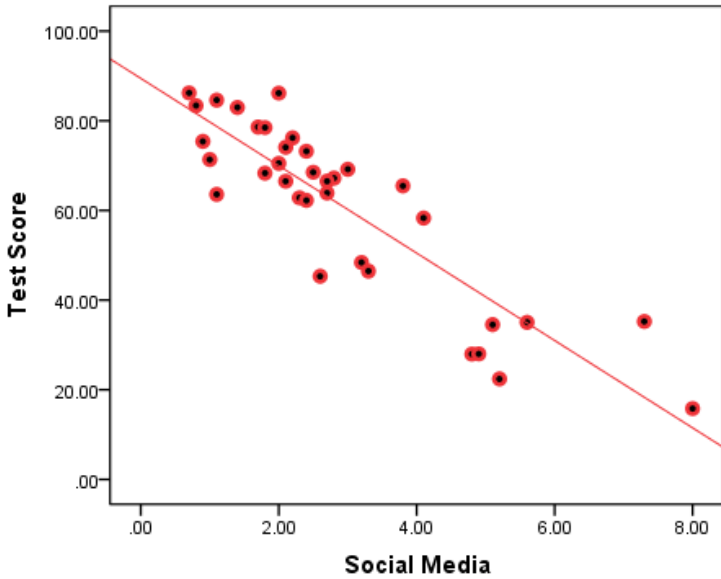
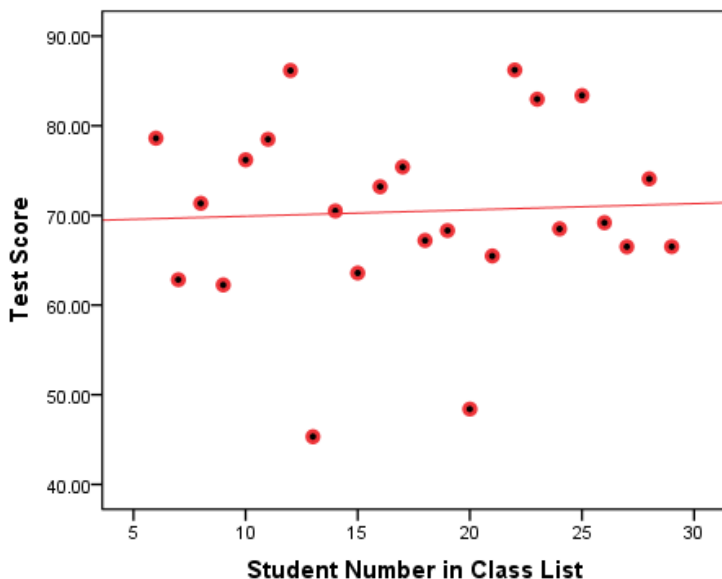


Figure 7.4(B) *Negative Association: Test Scores by Time Spent On Social Media With Line of Best Fit*



Compare the slopes of the lines in the figures above to the one in Figure 7.5 below.

*Figure 7.5 No Association: Test Scores By Student Number in Class (Selected Scores)*



The graph in Figure 7.5 above plots the non-existent association between a student’s number in in the class and their final test score. Of course, this is a bogus “association” which I’m showing here only as an example of a *flat line of best fit*, an indication that the two variable have nothing to do with each other. The line in Figure 7.5 is not perfectly flat, however, so it helps to have a numerical indication of association in addition to the visual ones the scatterplots give us.

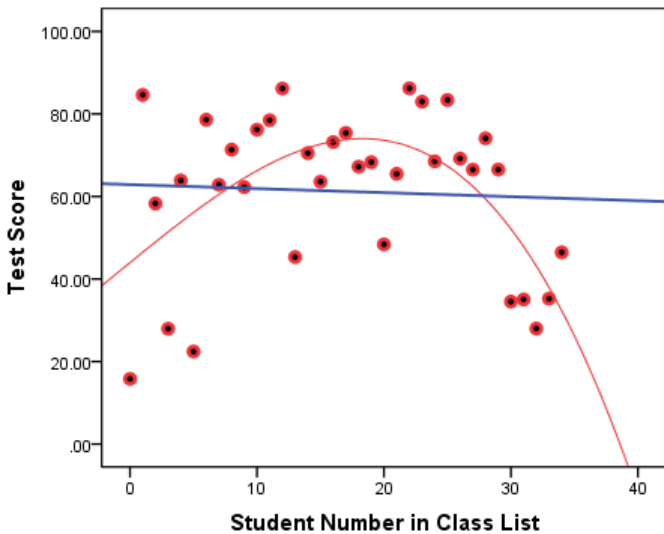
Before we get to that, a word of warning. The presumption of linearity for this type of analysis is very important and **you should make sure to not impose linearity where it doesn’t exist**. The caveat below explains.

**Watch Out!! #14 . . . For Non-Linear Associations**

Data points without a pattern produce a flat (i.e., with no linear slope) line of best fit, as shown in Figure 7.5 above. However, **data points in a non-linear pattern will also result in a flat (i.e., with no linear slope) line of best fit**, if we insist on seeing the variables as linearly associated. This can lead to dismissing a potential association only because it's non-linear, which would be a mistake. While this textbook doesn't go into non-linear associations, this doesn't mean they do not exist or they are not important: on the contrary, but they do require you to use different methods to investigate them.

My warning here is simple: **When working with given data, keep an eye on potential non-linearity. Otherwise you may incorrectly assume no association when in fact a non-linear association exists.** Figure 7.6 below illustrates.


*Figure 7.6 Curvilinear Association: Test Scores By Student Number in Class (All Scores)*



Surprisingly enough, Figure 7.6 shows that students at the beginning and at the end of the class list scored lower on their final test than their peers for whatever reason, or simply by chance (my bet would be on the latter).

Regardless of the reason or lack thereof, my goal here is to show you that imposing linearity by drawing a *linear* line of best fit will end up as a flat line, which one hastily may take as an indication of no association (see the straight blue line on the graph). A closer and more careful look, however, reveals the inverted-U shape pattern of the data points in the scatterplot: As the student numbers increase initially, so do the test scores. Then, as the student numbers continue to increase, the test scores start decreasing (see the curved red line following the data points much more closely than the blue flat one). This is clearly a pattern that should not be ignored in any serious, real-life study.





A visual summary of the data and any potential bivariate associations like the scatterplot is thus very useful. Scatterplots are in fact rather indispensable if one is to base their analysis on the assumption of a linear association between two continuous variables. Still, like in the previous two cases of two discrete variables and a discrete and a continuous variable, a numerical summary of the potential association can be of great help.

For discrete variables we could examine and report differences of proportions, while for a discrete and continuous variables we use differences of means (or medians). In both cases we could compare groups (on proportions, or means). In the case of continuous variables, we have neither groups, nor a convenient number to compare them on. Instead, here we have a correlation coefficient, *Pearson's  $r$* . The correlation coefficient takes all data points simultaneously and summarizes to what extent certain values of one of the variables go with certain values of the other variable (i.e., if they form a pattern or they vary independently of each other).

While we will examine the exact definition and calculation of the Pearson's  $r$  in Chapter 10 later, for now we'll focus on its interpretation.

**The correlation coefficient  $r$  is a number between -1 and +1, indicating the strength of any possible (linear) association between two continuous variables.** However,

there is a catch: **the strength of the association is calculated in absolute terms while the  $\pm$  sign is there to indicate whether the association is positive or negative.** Thus, both  $r=-1$  and  $r=1$  stand for the strongest possible (i.e., perfect) correlation, the former perfect *negative association*, the latter perfect *positive association*. Between them is  $r=0$ , or no association.

While perfect correlations ( $r=\pm 1$ ) are very rare (if not non-existent)<sup>5</sup>, most variables's associations are somewhere between 0 and  $\pm 1$ . **The closer a correlation is to 0, the weaker it is; the closer the correlation is to -1 or +1, the stronger it is.** Typically, in the social sciences a correlation of about  $r=\pm 0.7$  would be considered strong, a correlation of about  $r=\pm 0.5$  would be considered moderate, and a correlation about  $r=\pm 0.3$  would be considered weak. Correlations around  $\pm 0.8$  or  $\pm 0.9$  would therefore be very strong, while associations around  $\pm 0.2$  and  $\pm 0.1$  would be quite weak.

Now that you are well-equipped with knowledge about interpreting correlations, let's see what the correlations of the associations discussed above were.

First we looked at class attendance and test scores (Figures 7.3(A) and 7.3(B)); the correlation between the two variables was a very strong  $r=0.881$ . Then, we looked at social media usage and test scores (Figures 7.4(A) and 7.4(B)), where the correlation was equally strong  $r=-0.882$ <sup>6</sup>. Finally, we discussed the practically non-

5. The obvious exception here is the correlation of a variable on itself, which will produce  $r=1$ .
6. If you're wondering why the correlations appear to be of the same strength, the reason lies in the way I created the synthetic variable *social media*

existent linear association between student number and test scores (of selected students, Figure 7.5) whose  $r=0.049$ , while the improperly imposed linearity in Figure 7.6 from the caveat had a similar so-weak-almost-zero linear correlation of  $r=-0.051$ .

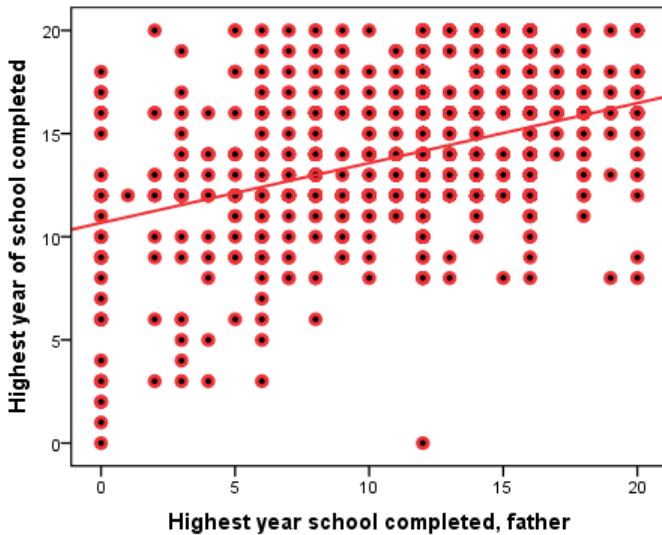
Tired of fake data? Ready to return to the real world of sociological research? Then let's take a real example with existing data and see how it all works out.

*Example 7.5 Intergenerational Reproduction of Privilege in Education in the USA (GSS 2018)*

For this example I used data from the National Opinion Research Center's (NORC) at the University of Chicago *General Social Survey (GSS) 2018*. I'm interested in exploring whether *father's education* and the *education* of the respondent are potentially correlated. Both father's education and education of the respondent are measured in years of schooling, ranging from 0 (no education) to 20 years. As such they are discrete ratio variables which we can treat as continuous due to their number of values being quite large (twenty-one to be precise). Figure 7.7 shows the relevant scatterplot.

*usage* -- as an inversion of the simulated variable *class attendance*. I did warn you the data is made up as a heuristic. (Do not take this to mean that such associations -- between attendance and class performance and social media usage and test scores -- do not exist in real life.)

*Figure 7.7 Respondent's Years of Schooling by Father's Years of Schooling (GSS, 2018)*



There are several things to note in the graph above. One is that the data points look less “scattered” and more orderly arranged in neat rows and columns than would be the case, had we variables with much larger number of values. Furthermore, while  $N=1,687$ , there are much fewer data points on the scatterplot: the reason, of course, is that there are many observations “on top” of each other, i.e., most data points represent more than one person’s combination of their own years of education and their respective father’s years of education. (After all, most such combinations are unlikely to be unique; we can arguably expect there to be more than one respondent and their father both having, say, 12 years of education in the dataset.)

Substantively, however, what do we see in the scatterplot above? To the extent that there are respondents with low levels of education, they seem to have fathers with low levels of education too. As well, while respondents with higher levels of education seem to have fathers with all levels of education, those with higher parental education appear to be more than those with lower parental education. (That is, both the left and the right side of the upper half of the scatterplot have many observations, but the top right area do seem to contain more observations than the top left area). Finally, and most importantly, there seem to be almost no respondents with low levels of education whose fathers had high levels of education (note the empty bottom right area of the graph).

All in all, it seems like more years of father's education "go" with more years of respondent's education, and fewer years of father's education "go" with fewer years of respondent's education — though not completely so, or the top left area of the graph (the less educated fathers with more educated offspring) would be empty too. This is reflected in the line of best fit whose slope, while positive, is not very steep.

**Ultimately, the scatterplot indicates that *father's education and respondent's education* seem positively associated in the dataset but also that this association is not very strong.** That is, there appears to be intergenerational reproduction of privilege in education, however, fortunately, one's father's lower levels of education don't seem to completely preclude one's own educational attainment.

The correlation coefficient provides a numerical summary of the potential association described above.

Table 7.4 Correlation between Father’s Years of Schooling and Respondent’s Years of Schooling (GSS 2018)

Correlations			
		Highest year of school completed	Highest year school completed, father
Highest year of school completed	Pearson Correlation	1	.413**
	Sig. (2-tailed)		.000
	N	2345	1687
Highest year school completed, father	Pearson Correlation	.413**	1
	Sig. (2-tailed)	.000	
	N	1687	1687

\*\* . Correlation is significant at the 0.01 level (2-tailed).

SPSS’s output provides *r* as “Pearson Correlation”, and here *r*=0.413. As suspected, this reflects positive a moderate/moderately-weak association.<sup>7</sup>

To summarize, you can describe and examine potential

7. Note that SPSS's bivariate correlation tables are 2x2 tables, with the information repeated twice. Thus, while four coefficients are provided in the central cells of the table, they are actually two pairs of the same two correlations. (That is, correlations are symmetric: correlating *Variable 1* on *Variable 2* is the same as correlating *Variable 2* on *Variable 1*.) As well, one of these two pairs is always equal to 1, as a variable correlated on itself is a perfect correlation. This is shown in the table as *corr*(Highest year school completed, Highest years school completed, father)=0.413=(Highest year school completed, father, Highest years school completed) and *corr*(Highest year school completed, Highest year school completed)=1=(Highest year school completed, father, Highest year school completed, father).

associations between continuous variables through scatterplots with lines of best fit (looking for a concordant or discordant pattern in the data points) and the coefficient of correlation  $r$  (ranging from 0 to  $\pm 1$  in strength, with 0 standing for no correlation and  $\pm 1$  constituting a perfect negative or a perfect positive correlation).

Before we move on, the tip below shows how to get the visual and the numerical summary of continuous bivariate associations in SPSS.

#### *SPSS Tip 7.3 Scatterplot and Correlation Coefficient*

##### **For Scatterplots:**

- From the *Main Menu* select *Graphs* and, from the pull-down menu, *Legacy Dialogues*; click on *Scatter/Dot*;
- Keep the pre-selected *Simple Scatter* option and click *Define*;
- In the new window, select one by one your variables of interest from the list on the left and, using the arrow buttons, move them to the *X-Axis* and *Y-Axis*<sup>8</sup> empty spaces on the right; click *OK*.
- The *Output* window will show the resulting scatterplot; double-clicking on it will open a *Chart Editor* window from where you can change the text, colours, size, etc. of the graph to suit your needs.

##### **For the correlation coefficient (Pearson's $r$ ):**

8. At this point, it doesn't really matter which one you put in the X- or Y-Axis though I would suggest placing the variable that precedes the other in time (like father's education generally precedes offspring's education) in the X-Axis. The reasons for this will be explained in Chapter 10.

- From the *Main Menu*, select *Analyze*;
- From the pull-down menu, select *Correlate* and then *Bivariate*;
- In the resulting window, select one at a time your two variables of interest from the list on the left and, using the arrow button, move them to the *Variables* space on the right (the order is not important); click *OK*.
- The *Output* window will display a symmetric 2×2 table with your requested correlation coefficient.



---

## 7.3 Summary [EMPTY]



---

## 8 Hypotheses Testing

In Chapter 7 we learned how to look for associations between two variables in random sample data. Just because two variables' observations exhibit a pattern that we can see in the sample doesn't mean that the variables are necessarily *truly* related in the population. Recall the purpose of sampling from Chapter 6: to infer something about a population based on a sample, i.e., to use sample statistics to estimate population parameters.

Given this, the questions you should be asking at this point are: Is an association we observe in the sample data something that exists in the population of interest? That is, do we observe this association because it really exists in the population and is reflected in the sample? Or is our sample unusual enough so that the association is an artifact of random chance, present only in this one sample? How certain can we be in our conclusion either way?

To answer these questions, you need to learn how to *test potential associations for statistical significance*. The last section of this chapter and the next two chapters are devoted to just that. First, however, there is some preliminary work to do. To that effect, in this chapter I introduce you to the concept of a *hypothesis* in social science research and the logic of hypotheses testing, both as a theory and in practical terms.

Before we delve into this (rather extensive) topic, I still have to address the elephant in the room when it comes to statistical associations: *causality*, next.

---

## 8.1 Causality

From the start, I need to make one thing clear: regardless if observed only in sample data or generalizable to populations, so far we have only discussed *statistical* associations.

*Well, what kind of other associations could we discuss, I can imagine you grumbling, it's a statistics textbook — of course the associations will be statistical!*

You are correct, of course, but (you knew there will be a “but”) — “statistical” here has a *very* narrow meaning, something most people unfamiliar with statistics seem unaware of and thus interpreting to mean a lot more than it actually does.

You see, *statistical association* refers only to whether there is a pattern in the data or not; whether certain attributes of one variable tend to go with specific attributes of another variable. In no way does this imply that one variable is what it is *because* of another, or that a change in one *causes* another variable to change, or that a variable is dependent on another.

If we can state any of these, we make a much stronger claim — one of *causality* — and the associations are then

called *causal*<sup>1</sup>. When we have a causal association, we call one variable *independent* and the other *dependent*<sup>2</sup>.

See if you can differentiate statistical and causal associations. Smoking is associated with lung cancer: people who smoke (or smoke more) have lung cancer at higher rates than those who don't. Smoking *causes* lung cancer: smokers are *more likely* to get lung cancer *because* of the fact that they smoke. Class attendance and test scores are associated: students who attend more classes have higher test scores. Test scores are *dependent* on class attendance: coming to class more often is partly *responsible* for higher test scores. Parental education and offspring education are positively correlated: higher levels of parental schooling are associated with higher levels of schooling for the offspring. Individuals with higher levels of schooling have more education *because* their parents were better educated themselves.

The first sentence in any of the examples in the previous paragraph was a statement of statistical association, the second statement was one of causality. If they generally sound the same to you, you should start paying more explicit attention to phrasing, specifically how the claims of association are put into words. As the one of most often-quoted sayings in statistics goes, **correlation is not**

1. Please make sure you don't confuse causal ['KO-zal] and causality [ko-'ZA-liti] with casual ['KEH-jwal] and causality (which doesn't exist).
2. You can think of the independent variable (i.e., the cause) as free to vary on its own; with or without the dependent variable, the independent is what it is. On the other hand, the dependent variable (the effect) varies *because* of the independent one, that's why it's called *dependent*. (Note that it's *dependent* variable and not *dependant*. The latter applies to people who are economically supported by others, like children are dependants of their parents.)

**causation.** Apart from urging caution in interpreting results, it also brings attention to how careful researchers must be when reporting results and conclusions in order to not overstate their claims.

What is the main difference between statistical association<sup>3</sup>

Establishing a statistical association between two variables is relatively straightforward and easy: there are tests for that (as we shall shortly see)<sup>4</sup>. Establishing a causal association between two variables (especially in the social sciences), on the other hand, is notoriously hard.

**Criteria for establishing causality.** There are three basic requirements for establishing causal associations, and an additional, overarching one related to the logic of research as a whole.

1. **Does the variable we claim is the *cause* come before the variable we claim as an *effect* in time?**
3. While many times the words *association* and *correlation* are used interchangeably, I prefer to use *correlation* only in relation to continuous variables in the context of the correlation coefficient. Referring to any statistical association as *correlation*, however, is technically not wrong; the usage is simply a matter of preference. and causation? Briefly, the method of establishing either; what is necessary for us to be able to claim one or the other.
4. Of course, it's not as easy as I present it further in this text. As an introduction to the topic, however, it will suffice. My point is that relative to establishing causality, it is easier.

This requirement is also known as *temporal precedence* — that is, **whether the potential cause happens before the potential outcome**. It is squarely based on logic: after all, an outcome cannot logically precede its cause. You can't take a test on the first day of class, and claim that your test score was due to your attending class or being absent later in the semester: that's not how time works. Similarly, you cannot claim that the bachelor's degree you will get in the near future is somehow responsible for your parents college degrees from twenty or so years ago.

While in these examples the temporal precedence is crystal clear, keep in mind that this is not always the case. There are plenty of situations in social research when it's difficult to adjudicate which one of a pair of variables came first, as well as cases of mutual causality and reverse causality. Without getting into too much detail take, for example, the popular finding [citation: Waite] that married people tend to be happier, on average. One can easily conclude that marriage promotes happiness. But what if happier people tend to have more successful relationships leading to marriage and a related propensity to stay married? Which one, marriage or happiness, is the cause of the association and which one the outcome? Further analysis and investigation of the variables' association is necessary in such a case (and even that might not lead to definite conclusion).

## 2. Are the two variables statistically associated?

This provides further evidence that statistical association is different from causation by listing the presence of a statistical association as a necessary requirement for establishing causality, among others. In short, **the presence**



**of a statistical association between two variables is a *necessary but not sufficient* condition for claiming causality.**

Why it's necessary should be obvious: we cannot claim that we have a variable we think is a cause to a potential outcome variable, if we have no evidence whatsoever that they are statistically associated in the first place. Otherwise, if there is no observable pattern between the values/categories of the two variables, how can we claim that changes in one variable *cause* changes in the other? Again, logically, the cause and the effect must be related in some way for which association we have enough evidence with a specific desired level of certainty. (The remaining chapters are devoted to finding just that type of evidence.)

### **3. Are there no alternative explanations of the variables's statistical association?**

This condition is the most complicated one of the three, as it requires the examination of other variables and not just the two of initial interest. Again briefly, there are concerns about causality due to the social world being vastly complex and to the social science variables' complicated interplay in real life. Basically, in the social world there rarely is a single cause of anything.

For example, is the statistical association in question observed because the potential cause variable *indeed* affects the potential outcome variable — or because both variables are in fact effects of a *third* variable (sometimes without any association between the original two variables)? Can we differentiate between a genuine relationship and a so-called *spurious* (i.e., fake, bogus) one, like the one described? As well, perhaps we only

observe a statistical association between two variables and claim one as the cause because we haven't considered different potential causes. How can we be certain that it is (solely) the "cause" we have identified, or that if we considered alternative causes, the original so-called "cause" will remain as one?

Regarding the latter, consider again the association between *class attendance* and *test scores*. Would you believe me if I told you that your statistics test scores depended *only* on your class attendance? What about hours of studying, potential after-class tutoring, doing exercises, pre-existing math knowledge, searching for/reading additional sources online or in the library, asking relevant questions in class and/or office hours, etc., etc.?

There are numerous reasons why anyone would score higher or lower on a test, and I just listed a few of the study-related ones. We don't need to limit ourselves to these though. How about general health on the date of the exam (maybe you have come to the test sick)? Or romantic relationship or family problems one might be going through? A sick relative at home? Episodes of anxiety and/or depression? Being overworked, working a night shift before the test, and/or not getting enough sleep for another reason?

You certainly can add even more reasons for why a particular test score ends up what it is, and that class attendance is merely *one* such potential cause. (Are we even certain that, if we somehow accounted for all the other potential causes, we would still observe an association between attendance and scores?)

As to spurious associations, consider that it's possible for two variables to *seem* associated (i.e., there is a pattern between their values/categories; changes in one are accompanied by changes in the other) only because a third variable is causing the changes in both. Then if, instead of focusing on the two genuine associations, we ignore the third variable and focus on its two outcomes which just happen to change at the same time, we would make a wrong conclusion in attributing causality to an association that essentially doesn't exist.

Take for example *life expectancy* and *internet*: Since 1990s, as internet was becoming more and more widespread in Canada, the Canadian life expectancy at birth was also increasing. We can therefore conclude that internet prolongs life. But there is a reason why you've never before heard about this particular beneficial effect of internet on one's health and life — it's extremely doubtful it exists. After all, wouldn't it make more sense to attribute both to general technological progress (not only in communications, IT, and infrastructure but also in healthcare and medicine)?

Finally, this is where the additional, overarching general condition for causality comes into play. Assuming the three conditions listed above are met, **claiming causality essentially implies providing a *logical explanation of the observed association***. In and of itself, causality is about having a theory — an idea, if you will, *why* there is such an association. Without such an idea, we are left simply with two variables which may be or may not be *statistically* — *but definitely not causally* — associated, and the statistical

association doesn't mean much, on its own<sup>5</sup>. And given that the potential statistical association you may think exists might not even be there once other alternative causes are considered, you should realize by now that making a causal claim is indeed not a walk in the park.

What is to be done then? Obviously, such a brief presentation on the topic leaves a lot to be desired and is not going to be enough to fully prepare you for the task of comprehensively establishing causality in real-life research. What you should be able to do even now, however, is appreciate causality's complexity, keep in mind the necessary conditions for claiming causality (and apply these when reading about research findings and questioning conclusions), and always, always keep an eye on alternative explanations in particular (by asking yourself "what else could be causing this?"). These should provide enough basis for you not to take statements about statistical association between variables as more than they are, and to not confuse them with claims about causality.

As well, I hope you would be careful in phrasing your own conclusions when communicating statistical research to others by not overstating the findings of any analyses

5. You most certainly need to check these associations

out: <http://www.tylervigen.com/spurious-correlations>. (At this point you need any distraction you can get, and this time you can even say it's for a good, pedagogically meaningful cause. Or so I can tell myself.) Among them, you'll learn that the number of doctorates in Sociology awarded in the USA is very strongly correlated over time with worldwide non-commercial space launches, not to mention that the number of drownings by people falling into a pool correlates moderately strongly with the number of movies in which Nicholas Cage appeared for the ten years between 1999 and 2009 (CITATION Spurious Media LLC/Tyler Vigen <http://www.tylervigen.com/spurious-correlations>

you might end up doing, especially if they involve only two variables, as per our discussion. By now it should be clear that real-life research considers many variables at the same time. Such *multivariate* analysis lies beyond the scope of this book so you should take any bivariate associations we discuss to be of solely *indicative* (or exploratory) nature — something that additional, multivariate analysis may establish at a later point, but definitely not a finished product. After all, you didn't expect that you can establish causality by considering only two variables, did you?

With this in mind, we proceed with the question of how to establish *statistical* associations — and not just the ones observable in sample data, but the associations in which we are truly interested, i.e., those generalizable to populations. You may not be able to make claims about causality at this point but you can certainly learn how to test for evidence of statistical associations between two variables. To that purpose, the next section introduces the logic of using hypotheses in research and how hypotheses get tested.



---

## 8.2 Hypotheses

Now that we have come to terms with the fact that we will not be making causal statements at this point, let's turn our attention to establishing statistical associations. As I mentioned in the previous section, this is done through testing. In order to test variables' associations we need to know how hypotheses are scientifically tested.

To have a hypothesis about something means to have an idea about how to explain it. This idea, or proposed explanation, might be based on any combination of logic, previous related observations, experience, etc. In science, hypotheses are formulated as relatively concise, *testable* statements. If a statement cannot be tested, it doesn't qualify as a scientific hypothesis.

Most students unfamiliar with the scientific method of testing hypothesis are surprised to learn that the testing is done in a roundabout, method-of-exclusion kind of way: **we don't set out to confirm our hypothesis but rather to reject the opposite of what we claim.** To baffle you further, if we reject the opposite, we have found evidence in support of our hypothesis but we have not *proven* that it's true. Similarly, if we do *not* reject the opposite, it doesn't mean that we've *proven* it as true *or* that we have *proven* our hypothesis wrong. (Nothing is ever proven in science as that would require 100 percent certainty and we already established that is impossible.) Thus, interpreting

a hypothesis test requires careful, qualified language as to not overstate findings.

Confused? Not to worry. I am getting ahead of myself here to give you a quick sketch of where we are headed in this section, but of course I will go over and explain the parts of the paragraph above in greater detail below. Also a heads-up: after the brief respite, things are about to get technical again (in the next section). But first things first.

**To test a hypothesis of interest, we make two contradictory statements: one about what we hypothesize and another stating the *exact* opposite<sup>1</sup>.** The “opposite” hypothesis is called a *null hypothesis* (frequently designated as  $H_0$ ) and is usually stated first; the original hypothesis of interest is called an *alternative hypothesis* (usually designated as  $H_a$ ) and is stated second<sup>2</sup>.

When we apply all this to testing variables’ associations, we end up with null hypotheses such as “the two variables are not associated”, “there is no association between the

1. Why? Beyond what I already explained about proofs, also because scientists need to be impartial about what a test will reveal. As a scientist, you want to test a hypothesis with an open mind and to be equally prepared to accept the result either way it goes -- so you cannot set out from the start to find your hypothesis supported.
2. Do not get alarmed if you see different notation in published research. When researchers test many hypotheses in the same study, they may designate them as  $H_1$ ,  $H_2$ ,  $H_3$ , etc. Even more importantly, experienced researchers don't explicitly state the null hypotheses in their studies -- they are self-understood as the opposite of whatever each alternative hypothesis states. Further, some researchers never explicitly designate a hypothesis as it is taken as evident that this is what they do. Beginner researchers like you, however, should practice stating -- and clearly designating -- both null and alternative hypotheses.



two variables”, or “Variable 1 does not affect Variable 2”, or “the two variables are independent of each other”, etc. The alternative hypotheses then would be something like “the two variables are associated”, “there is an association between the two variables”, or “Variable 1 does affect Variable 2”, etc. (However, recall that when interpreting and reporting results it is always better to state the findings not only in terms of variables but also in terms of people.) See some examples in the box below.

### *Example 8.1 Stating Hypotheses*

#### *Hair colour and eye colour:*

- $H_0$ : Hair colour and eye colour are not associated; e.g., dark-haired individuals are equally likely to have blue eyes as blond individuals are.
- $H_a$ : Hair colour and eye colour are associated; e.g., dark-haired individuals’ and blond individuals’ likelihood of having blue eyes is different.

#### *Smoking and lung disease:*

- $H_0$ : Smoking and lung disease are not associated; e.g., smokers and non-smokers have the same odds of developing lung disease.
- $H_a$ : Smoking and lung disease are associated; e.g., smokers and non-smokers have different odds of developing lung disease.

*Gender and income:*

- $H_0$ : Income is independent of gender; e.g., men and women have the same average income.
- $H_a$ : Income is dependent on gender; e.g., women and men have different income on average.

*Parental education and offspring education:*

- $H_0$ : Parental education is unrelated to the education of their offspring; e.g., the level of parental education has no effect on children's level of education.
- $H_a$ : Parental education and their offspring's education are related; e.g., the level of parental education is associated with the children's level of education.

There are three things that you can learn from the examples presented above. **First, the hypotheses are formulated as short statements that can be evaluated in a simple yes-or-no kind of way:** “Average income is independent of gender”: YES, or “Average income is independent of gender”: NO. Thus you really need only one statement per hypothesis; if your proposed explanation is complicated and involves more than two variables, this means you are dealing with multiple hypotheses, each of which needs to be tested separately.

**Second,** while there are many ways you can state

essentially the same hypothesis, **try to keep the null hypothesis as the same statement the alternative hypothesis has but in opposition**, such as “...are not related/associated” and “...are related/associated”, or “...are independent” and “...are not independent”, etc.

**Third**, you may have noticed the slightly awkward way in which some of the alternative hypotheses are listed above. Couldn't I have stated “women have lower income than men on average”? Or, “blond individuals are more likely to have blue eyes than dark-haired individuals”? I could but then these would have been different alternative hypotheses. The reason I did not imply who is more or less likely to have blue eyes, or who has a higher income on average but **kept the statements as a generic “different likelihood” and “different income” is because it affects the kind of test that needs to be used**. Briefly, there is a general test for association/difference (aka *two-tailed test*), and a more specific version (aka *one-tailed test*) which implies “direction”; the former is more “open-minded” as it doesn't rely on or imply prior knowledge and is therefore more conservative. The latter indicates not only a difference (i.e., association) but of what specific type so its usage needs to be justified. More on that in the next section but for now keep in mind that as beginner researchers, it's recommended you use the general, two-tailed, version of the test.

Before we move to some actual hypothesis testing, see if you can formulate some hypotheses on your own.

*Do It! 8.1 Stating Hypotheses*

Formally state the null and alternative hypotheses about each of the following pairs of variables: class attendance and test scores, time spent on social media prior to a test and test scores, race/ethnicity and years of schooling, gender and belief in climate change, political affiliation and attitudes toward gun control. In fact, just go ahead and practice formulating hypotheses about anything you like.

---

## 8.3 Hypothesis Testing

The first thing you should know about testing hypotheses is their relationship to statistical inference: **We formulate hypotheses about the population of interest, and *only* about the population of interest. We test them through sample data.**

Like so: imagine I have read enough on the topic of the gender gap that I hypothesize that women and men receive different income on average. I explore my sample data and I do find that in *the sample with which I'm working* men have a higher average income than women. *It seems* like there is an association between gender and income; however, *I do not know* if there is an association between gender and income *in the population* in general. To that effect, I want to estimate (with a given level of certainty) whether such an association exists *in the population*. My hypothesis is about *the gender/income association in the population*. (After all, I can *see* the different average income levels in the sample; there is no need to *hypothesize* about the sample.)

You may be getting tired of my italicizing “the population” but it really *is* that important: hypotheses are stated about *the population*. This is key for testing, so keep it in mind.

If the test provides us with evidence in support of our

alternative hypothesis, we call the association being tested *statistically significant*<sup>1</sup>.

Before we get to the nitty-gritty details of hypotheses testing, here's **an overview to show you the underlying logic of how it all works:**

1. State the null and the alternative hypotheses;
2. Assuming the null hypothesis as “true”, calculate the related score (e.g., *z*-value, *t*-value, etc.);
3. Find the probability associated with that score (essentially the probability that the null hypothesis is indeed “true”, called *p*-value).
4. If that probability is *low enough* (i.e., below the *level of significance*, explained below), reject the null hypothesis; if the probability is *too high* (above the level of significance), fail to reject the null hypothesis.
5. **If the null hypothesis has been rejected, you have found support for the alternative hypothesis: your bivariate association is statistically significant, and therefore generalizable to the population.**
6. **If the null hypothesis has not been rejected, you have found no support for the alternative hypothesis: your bivariate association is not statistically significant, and is *perhaps* due to**

1. Statistical significance has a very narrow, very specific meaning as you will learn further in this section. On the difference between statistical significance and significance in general, see warning in the Watch Out!! #15 box in the next section.

**expected sampling variability (i.e., to random error) appearing in this one particular sample.**

Example 8.2 below illustrates the whole process in detail. As with applying the Central Limit Theorem to confidence intervals, it is easier to start with an example where we assume we have the population parameters. Once you grasp the underlying logic, we can move on to properly testing bivariate associations. (The example below is for heuristic purposes only.)

*Example 8.2 (A) Employee Productivity (Finding Statistically Significant Results,  $N=100$ )*

Imagine a large company has created a productivity index to measure its employees' productivity. The (interval scale) index is constructed to be normally distributed, with a mean of 600 points and a standard deviation of 100 points.

Imagine further that a hundred of the company's employees were randomly selected to attend a new specialized training course, after which their average productivity score was measured as 650 points. (To simplify things, we'll also assume that their standard deviation is the same as the general group of employees.) Can we conclude that the training course had indeed increased productivity? Or is the gain of 50 points something due to regular

sampling variability? That is — is this 50-points gain statistically significant?

Here's what we have, formally stated:

$$\mu = 600$$

$$\sigma = 100$$

$$\bar{x} = 650$$

$$N = 100$$

What we want to know is the probability of a score of 650 if the training course didn't contribute to the gain, i.e., **the probability of a score of 650 under the condition of the null hypothesis.**

- $H_0$ : The training course did not affect productivity (the 650 score was due to random chance);  $\mu_{\bar{x}} = \mu$ . The true/population mean of the trained is the same as that of the untrained employees.
- $H_a$ : The training course affected productivity (the 650 score was a true gain);  $\mu_{\bar{x}} \neq \mu$ . The true/population mean of the trained is not the same as the population mean of the untrained employees.

Recall from Chapter 5 and Chapter 6 that to obtain the probability of a score we need to express it in terms of standard deviations (i.e., here in standard errors, as we are working with a sampling distribution).



The standard error is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{100}{\sqrt{100}} = \frac{100}{10} = 10$$

Then the z-value of 650 is:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{650 - 600}{10} = \frac{50}{10} = 5$$

That is, the trained group's mean of 650 is five standard errors above the 'general' (not-trained employees') mean of 600. Considering that we know that 99 percent of the time a sample mean will fall within 3 standard errors away from the population mean, the probability for the trained group's mean in the sample to be 5 standard errors above the mean of everyone else is extremely small (smaller than 0.5% to be exact, as explained below). Given the properties of the normal curve, we know that 68 percent of all means in infinite sampling will fall between  $\pm 1$  standard error (i.e., between 590 and 610), 95 percent will fall between  $\pm 1.96$  standard errors (i.e., approximately between 580 and 620), and 99 percent will fall between  $\pm 2.58$  standard errors (i.e., approximately between 570 and 630). The score of 650 which is 5 standard errors above the mean indeed would fall very, very far in the right tail.

In terms of probabilities, consider the following: if a sample mean has a 99 percent probability of being approximately between 570 and 630, and the remaining 1% is distributed equally in the two tails, the probability beyond 630 is 0.5%. *Assuming the null hypothesis were true* (i.e., training had no effect and we see the 650 by chance instead

of a real increase while the true/population mean of the trained is 600), our calculations show that the 650 score then appears with a probability of  $p < 0.005^2$  — a very small probability, so small that a score of 650 seems highly unusual.

**And this is where the crux of the logic of hypotheses testing lies: the chance of the 100 employees getting an average productivity score of 650 after a training course *if the course had no effect* (i.e., if their population mean is indistinguishable from the general/untrained mean) is so small, that it is *highly unlikely to be the case*. It is much likelier that the course had an effect, so that the trained employees' population mean is not the same as the untrained ones:  $\mu_{\bar{x}} \neq \mu$  (and in fact  $\mu_{\bar{x}} > \mu$ ). The null hypothesis is thus not supported.**

**We therefore reject the null hypothesis and conclude that the score of 650 does not appear to be just due to random variability (otherwise it would be within 3 standard errors away from the not-trained employees' mean — while it stands at 5, under the null hypothesis). Rather, it is statistically significantly different from 600.** In other words, our evidence suggests that the training course may have affected the productivity score of employees who took it. (Again, causality aside, note that we have not proven beyond a shadow of a doubt that it did, rather that *given our evidence at this point in time, we have a reason to believe it did.*)

2. The  $p$  here stands for "probability".

In the example above we ended up rejecting the null hypothesis. I will also show how it can turn out that we cannot reject the null hypothesis but first I will use the opportunity to 1) make a connection to a concept with which you are already familiar — confidence intervals, below; and 2) introduce two interrelated important theoretical concepts, the level of significance and the  $p$ -value, in the next section.

Believe it or not, hypothesis testing and confidence intervals are complementary as both testing a hypothesis and constructing a confidence interval allow us to arrive at the same conclusion. To see this, we just need to construct a, say, 95% confidence interval for  $\mu_{\bar{x}}$  from Example 8.2 (A) above:

- 95% CI:  $\bar{x} \pm 1.96 \times \sigma_{\bar{x}}$   
 $= 650 \pm 1.96 \times 10 = 650 \pm 19.6 = (630.4; 669.6)$

That is, we can be 95% certain that the average score for the population of employees who take the training course would be between approximately 630 points and 670 points. The average general score of 600 points is not part of the plausible values for  $\mu_{\bar{x}}$ , which is consistent with our decision to reject the null hypothesis.



---

## 8.4 Level of Significance and the p-Value

The concept *level of significance* is used to adjudicate whether the probability (of our results if the null hypothesis is true) is too high to dismiss the null hypothesis or low enough to allow us to reject the null hypothesis. In other words, the level of significance is what we use to proclaim results as statistically significant (when we reject the null hypothesis) or not statistically significant (when we fail to reject the null hypothesis).

Think about it this way: recall that with confidence intervals we had selected 95% certainty and 99% certainty as meaningful levels of confidence. What is left is 5% and 1% “uncertainty”, as it were, which we agree to tolerate. These 5% or 1% are distributed equally between the two tails of the normal distribution (2.5% on each side or 0.5% on each side, respectively). They also correspond to  $z=1.96$  and  $z=2.58$ . Following the logic of Example 8.2 (A) from the previous section, in order to reject a null hypothesis, we want the probability to be lower than these 5% or 1% (so that we can “feel confident enough”).

And this is exactly it: When we put it that way, saying that we want the probability (of the null hypothesis being true) — called a *p-value* — to be less than 5%, we have essentially set the level of significance at 0.05. If we want the probability to be less than 1%, we have set the level of significance at 0.01. We can go even further: we might

want to be extra cautious and to want a “confidence” of 99.99%, so that we want the probability to be less than 0.01% — then we have set the level of significance at 0.001.

These three numbers — 0.05, 0.01, and 0.001 — are the most commonly used levels of significance. The level of significance is denoted by the small-case Greek letter  $\alpha$ , i.e.,  $\alpha$ , thus we usually choose one of the following:

$$\alpha = 0.05$$

$$\alpha = 0.01$$

$$\alpha = 0.001$$

**You can think of the significance level as the acceptable probability of being wrong** — and what is acceptable is left to the discretion of the researcher, subject to the purposes of the particular study.

Following the logic presented in Example 8.2(A) then, **if the probability of the result under the null hypothesis — the  $p$ -value — is smaller than a pre-selected significance level  $\alpha$ , the null hypothesis is rejected and the result is considered statistically significant<sup>1</sup>**. This is denoted in one of the following ways:

$$p \leq 0.05$$

$$p \leq 0.01$$

1. Note the difference between  $\alpha$  and the  $p$ -value. While  $\alpha$  indicates what probability of being wrong we are willing to tolerate, the actual  $p$ -value we obtain is *not* the probability of being wrong. The  $p$ -value, again, is the probability of our result if the null hypothesis were true; in other words, if the null hypothesis is in fact true, and our  $p$ -value is, say, 0.03, we'd obtain our results 3% of the time simply due to random sampling error.

$$p \leq 0.001^2$$

**To summarize, when a hypothesis is tested, we end up with an associated  $p$ -value (again, the probability of the observed sample statistics if the null hypothesis is true). We compare the  $p$ -value to the pre-selected significance level  $\alpha$ : if  $p \leq \alpha$ , the results are statistically significant and therefore generalizable to the population.**

So far so good? Good. However, unfortunately this isn't all (sorry!). What I have presented above is the most conventional treatment of how to use and interpret  $p$ -values. It is attractively straightforward — but it's also arbitrary, and its *true* interpretation is subject of an ongoing debate. As an introduction to the topic, I will leave it at that but you should be aware that there's more to the  $p$ -value, and that its usage has been (rightfully) questioned and/or challenged in recent years.<sup>3</sup>  $p$ -value without context

2. In published research you will find results marked by one asterisk, two asterisks, and three asterisks. These correspond to their significance based on the level used:  $\alpha=0.05$ ,  $\alpha=0.01$ , and  $\alpha=0.001$ , respectively. The smaller the level of significance, the more strongly statistically significant the result is (i.e., most consider  $\alpha=0.001$  to indicate "highly statistically significant" results). (If you happen upon a dagger ( $\dagger$ ), it indicates significance at  $\alpha=0.1$  level, or 10% probability of being wrong, which most researchers consider too high, but some still use.
3. You can find plenty of information on the topic online; from journals banning the use of  $p$ -values and hypothesis testing in favour of effect size (the *Journal of Applied and Social Psychology*, see Trafimow & Marks, 2015 <https://www.tandfonline.com/doi/full/10.1080/01973533.2015.1012991>), to calls to abandon statistical significance (e.g., McShane, Gal, Gelman, Robert & Tackett, 2019 <https://www.tandfonline.com/doi/abs/10.1080/00031305.2018.1527253>), to others calling for its and  $p$ -values' defense (e.g., Kuffner & Walker, 2016 <https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1277161?src=recsys>; Greenland, 2019 <https://www.tandfonline.com/doi/full/10.1080/>

or other evidence provides limited information. For example, a  $p$ -value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large  $p$ -value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a  $p$ -value when other approaches are appropriate and feasible" (Wasserstein & Lazar, 2016<https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108?src=recsys>). Finally, if you really want to not to overstate what the  $p$ -value actually shows, see Greenland et al. (2016) for a of common misinterpretations and over-interpretations of the  $p$ -value, of confidence intervals, and tests significance (here: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877414/>). Because of its enormity, the topic is still conventionally taught as I presented it above (as it goes way beyond the scope of this book), at least at introductory level..

Going back to our example from the preivious section, let's see how the  $p$ -values can change due to particular features of the study, like the sample size. Example 8.2(B) illustrates.

00031305.2018.1529625?src=recsys). One thing is clear:  $p$ -values and levels of significance have become increasingly controversial. Still, the American Statistical Association's position is that although caution against over-reliance on a single indicator is necessary,  $p$ -values can still be used, *alongside with other appropriate methods*: "Researchers should recognize that a



*Example 8.2(B) Employee Productivity (Finding Statistically Non-significant Results,  $N=25$ )*

Imagine that we had the same information as in Example 8.2(A), however, 25 employees took the training course instead of 100 and their average score was 620. The we have:

$$\mu = 600$$

$$\sigma = 100$$

$$\bar{x} = 620$$

$$N = 25$$

We still want to know the probability of a score of 620 if the training course didn't contribute to the gain, i.e., **the probability of a score of 620 under the condition of the null hypothesis.**

- $H_0$ : The training course did not affect productivity (the 620 score was due to random chance);  $\mu_{\bar{x}} = \mu$ .
- $H_a$ : The training course affected productivity (the 620 score was a true gain);  $\mu_{\bar{x}} \neq \mu$ .

The new standard error is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{100}{\sqrt{25}} = \frac{100}{5} = 20$$

Then the z-value of 620 is:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{620 - 600}{20} = \frac{20}{20} = 1$$

Given the properties of the normal curve, we know that 68% of all means in infinite sampling will fall between  $\pm 1$  standard error (i.e., between 580 and 620), 95% will fall between  $\pm 1.96$  standard errors (i.e., approximately between 560 and 640), and 99% will fall between  $\pm 2.58$  standard errors (i.e., approximately between 540 and 660). The score of 620 has  $z = 1$  — it falls quite close to the not-trained group's mean of 600.

In terms of probabilities, consider the following:  $z=1$  has a  $p > 0.30$ . **Assuming the null hypothesis is true, our calculations show that the 620 score will appear more than 30% of the time due to random chance, which is a lot more than the 5% (at  $\alpha=0.05$ ) that we are willing to tolerate. As such, we cannot reject the null hypothesis: we do not have enough evidence to conclude that the gain in productivity of 20 points which the 25 employees demonstrated is statistically significant. In other words, we don't have enough evidence that the training course was effective.** (This doesn't mean that it didn't beyond a shadow of a doubt, just that *at this point in this particular study we don't have enough evidence to say it did.*)

We can also see the correspondence with confidence intervals:

- 95% CI:  $\bar{x} \pm 1.96 \times \sigma_{\bar{x}}$   
 $= 620 \pm 1.96 \times 20 = 620 \pm 39.2 = (580.8; 659.2)$

That is, we can be 95% certain that the average score for the population of employees who take the training course would be between roughly 581 points and 659 points. **The average general score of 600 points is a plausible value for  $\mu_{\bar{x}}$ , which is consistent with our decision to not reject the null hypothesis.**

Again, Example 8.2 is a heuristic device, used only to explain the logic of hypotheses testing. Of course, normally we wouldn't have information about population parameters and we will be using sample statistics (i.e., we would use not only the sample mean  $\bar{x}$  but also the sample standard deviation  $s$ , to calculate the estimated sampling distribution  $s_{\bar{x}}$ ). (Not to mention that we would have two different standard deviations, one for the trained group and one for the not-trained group of employees.) As you learned in the previous chapter, this moves us from using the  $z$ -distribution to the  $t$ -distribution with given degrees of freedom. Recall that with a sample size of about 100 — i.e., with  $df=100$  — the two distributions converge.

Here then is a quick-and-dirty method you can use as a preliminary indication of whether something will be statistically significant. Since  $z=1.96$  corresponds to 5% probability (2.5% in each tail), and  $z=2.58$  corresponds to 1% probability (0.5% in each tail), even without knowing the exact  $p$ -value associated with a given  $z$ -value, you can

guess that getting a  $z < 1.96$  will be non-significant while a  $z > 1.96$  will be significant at  $\alpha = 0.05$ ; similarly, getting a  $z > 2.58$  will be statistically significant at  $\alpha = 0.01$ <sup>4</sup>. As samples used in sociological research are commonly of  $N > 100$ , the same insight applies to the corresponding  $t$ -values with  $df \geq 100$ . Understand, however, that this is not an official way to test hypotheses or report findings: to do that, **you always need to report the exact  $p$ -value associated with a  $z$ -value or a  $t$ -value with given  $df$** <sup>5</sup>.

**One-tailed tests.** Finally, a note on *one-tailed tests*. While at the beginner researcher level, I advise you against using them yourself, it is not a bad idea to know they exist and what they are. Briefly, the idea is that if we have a good reason to suspect not only a difference/effect but a difference/effect with a specific direction (i.e., positive or negative), we can specify the hypotheses accordingly. To use Example 8.2(A) again, say, we think there is no possibility that the training course *decreased* productivity scores. Then we can state the hypotheses as:

- $H_0$ : The training course either did not affect productivity or *decreased* it;  $\mu_{\bar{x}} \leq \mu$ .
- $H_a$ : The training course *increased* productivity;  $\mu_{\bar{x}} > \mu$ .

This is a stronger claim (that's why it needs to be well-justified) — we test not a difference (that can be either positive or negative) but an *increase*. Thus, we move the

4. Obviously, for negative  $z$ -values we'll have all these in reverse:  $-z > -1.96$  will be non-significant and  $-z < -1.96$  will be significant, etc.

5. You can find a handy online  $p$ -value calculator of  $t$ -values here:  
<https://goodcalculators.com/student-t-value-calculator/>.

significance level to only *one* of the tails, as it were, the positive (right) tail, so instead of 2.5% being there, 5% are.

This change in probability essentially “moves” the  $z$ -value corresponding to significance closer to the mean; now a smaller  $z$ -value will have the  $p$ -value necessary to achieve statistical significance. To be precise, 5% (2.5% in each tail) corresponded to  $z=1.96$ ; all 5% in the *right* tail corresponds to  $z=1.65$ <sup>6</sup>. This obviously “lowers the bar” of achieving statistical significance *without changing the level of significance  $\alpha$  itself*, and makes rejecting the null hypothesis easier, hence my description of the two-tailed test as more conservative (and my insistence on using it instead of a one-tailed test).

Before we move on to the last section of this theoretical chapter, the promised warning about the meanings of the term *significance*.

**Watch out!! #15 ... for Mistaking Statistical Significance for Magnitude or Importance**

If you have been paying attention, you have learned by now that statistical significance has a very narrow meaning. To have a statistically significant result simply means that the probability of observing our sample statistics (or

6. You can check it here by selecting "up to Z": <https://www.mathsisfun.com/data/standard-normal-distribution-table.html>.

difference, or effect, etc.) as they are, given that the null hypothesis is true, is small enough to be (highly) unusual, to be so relatively rare as to indicate what we have is not a result of random sampling variation but of untrue null hypothesis.

None of this says *anything* about how *big* a difference/effect is — in fact it can be quite small, and still *statistically* significant, given large enough sample size and other study specifications<sup>7</sup>

Similarly, many people unfamiliar with statistics take statistical significance to mean that the finding are of significant *importance*. Again, nothing about statistical significance confers great meaning to or implies importance of statistically significant findings. One can study an objectively trivial/unimportant issue and have statistically significant findings of no relevance to anyone whatsoever.

To conclude, keep these distinctions in mind — between the conventional usage of the word *significant* (meaning either important, or big) and *statistical* significance — both when interpreting and reporting results and when reading and evaluating existing research.

7. This is actually one of the reasons some have called for abandoning *p*-values, statistical significance, and hypothesis testing whatsoever, because statistical significance is not indicative of effect size and is frequently over-stated to mean more than it does; at the same time over-reliance on *p*-values decreases attention to effect size, careful study design, context, etc..

When testing hypotheses, I defined the significance level as sort of probability of being wrong we are willing to tolerate. This implies that a likelihood of making an *erroneous* decision about the null hypothesis (to reject it or not) exists. The next and final section deals with just that.





---

## 8.5 Errors of Inference

Making decisions about hypotheses is inference based on evidence and logic. Inference, however, doesn't come with a guarantee of being right — in fact, it is guaranteed that being right all the time is impossible. All the evidence and logic in the world will not be enough to ensure 100 percent certainty of making the right decision simply by the probabilistic nature of statistical inference. As long as we work with samples to estimate populations, some amount of uncertainty will be unavoidable — or it wouldn't be called *inference* but *knowing*.

Logically speaking, since we have *two* options given a null hypothesis (to reject or not to reject), we can make *two* types of mistakes. One is to be wrong about rejecting the null hypothesis, the other to be wrong about *not* rejecting it.

You might be rolling your eyes at this — well *duh!* — but bear with me: these really are the two types of statistical error, imaginatively called *Type I* and *Type II*.

**If we reject a true null hypothesis, we commit a Type I error. If we fail to reject a false null hypothesis, we commit a Type II error.** Before I even explain these further, make a mental note that since we *either* reject *or* fail to reject a null hypothesis — one *or* the other — **at any given time we can only make *only one* of the two types of errors**. If you rejected your null hypothesis, the *only*

error you could have committed is Type I; if you did *not* rejected your null hypothesis, the *only* error you could have made is Type II.

The trick is that we never know if we have made an error or not. (If we knew, we wouldn't be making it in the first place, right?) We only know that the possibility that we have made an error exists. However, as with everything about inference we have discussed so far, what we can do is to quantify the uncertainty as best as we can.

Table 8.1 summarizes the errors of inference based on the (unknown) real situation and the (uncertain) decision we have made about it, through an analogy of a criminal trial. The null hypothesis then stands for “innocent” (no effect/difference/association, etc.) while the alternative hypothesis stands for “guilty” (there is an effect/difference/association, etc.).

*Table 8.1 Errors of Statistical Inference*

	Reality: Guilty	Reality: Innocent
Reject $H_0$ : Innocent ? Guilty Verdict	Correct Decision ( $1-\beta$ )	Type I Error ( $\alpha$ )
Fail to Reject $H_0$ : Innocent ? Innocent Verdict	Type II Error ( $\beta$ )	Correct Decision

Recall that to reject the null hypothesis, we had to have a test with a  $p$ -value lower than the pre-selected level of significance  $\alpha$ , i.e.,  $p \leq \alpha$ . The level of significance amounted essentially to how much probability of being

wrong we were able to tolerate (so as long as the probability of having the observations we did, given a true null hypothesis — i.e., the  $p$ -value — was less than that, we would be fine).

Now consider that I just defined Type I error as the probability that we are wrong about rejecting a true null hypothesis — and *ta-dam!* — **Type I error is exactly equal to  $\alpha$ , the significance level!** The great thing about it is that it is not only precise, it is also utterly under our control as we are the ones to decide how much error (regarding “convicting an innocent”) we want to tolerate. If we want a smaller such chance, we can just raise the bar — so that only the smallest  $p$ -values can pass under the lowest possible  $\alpha$ <sup>1</sup> Make sure you don not confuse  $p$  and  $\alpha$ , especially in that  $p$  does not show the probability of being wrong. Even the significance level is not the *true* error rate (Selkke, Bayarri & Berger, 2001), as you can see here if you're curious: <https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-correctly-interpret-p-values>.[/footnote].

On the other hand, **when we fail to reject a false null hypothesis** (i.e., when we “let a guilty person go free as if innocent”), **we make a Type II error, called  $\beta$ .** At the same time, as you can see in Table 8.1, **the probability to correctly reject a false null hypothesis is a neat  $1-\beta$**  (after all, the decision has only two options), **known as the power of the test.**

Unfortunately, there is no way for us to directly control  $\beta$ ; your best bet is to have a large sample size, which increases the test's power (to detect an

effect/difference/"guilt") where it truly exists, and thus indirectly decreasing  $\beta$ .

*Well then, you might logically ask, why don't we just decrease both Type I and Type II errors? I am afraid you cannot do that: **Type I and Type II errors are opposites, and as such there is a trade-off between them.*** Think about it: if you hate the thought of convicting an innocent, and say you would never do it, you will end up deciding "innocence" all the time, thus inevitably at some point letting a criminal go. If you decide that you hate letting criminals go, you can convict everyone, but then of course, eventually you will inevitably end up convicting an innocent.

In other words, the harder you make it to reject a null hypothesis/"to convict" (by making  $\alpha$  the lowest possible), the higher the chances you will commit Type II error, failing to reject a false null hypothesis (and you will let a criminal slip free). The easier you make it to reject a null hypothesis/"to convict" (by making  $\alpha$  as high as you want), of course the higher the odds of committing Type I error, rejecting a true null hypothesis (and convicting an innocent).

In summary, the errors of inference are unavoidable: every time we make a decision about the null hypothesis one way or the other, we run the risk of making *one* of the statistical errors. With a careful selection of  $\alpha$  and a comfortably large sample size, making an error should not worry you too much -- but do not forget that it is a distinct possibility.

I end this chapter with a warning.

**Watch out!! #16 . . . for Mixing Up Your Error Concepts**

The statistical errors presented in this chapter aside, you might recall that we discussed two other error concepts, the *random error* and the *standard error*. Make a note about all three: **1) the random error, 2) the standard error, and, 3) the Type I error and Type II error of statistical inference. They are all different concepts.**

As a brief reminder, the random error is an inevitable corollary of sampling and reflects the fact that a sample is different from the population from which it was taken; the standard error is simply a formula for the standard deviation of the sampling distribution; and finally, the Type I and Type II statistical errors apply to decisions about the null hypothesis during testing.

Now that you know how hypothesis testing works *in principle*, let's get us some variables' associations tested with their appropriate tests, in Chapter 9 and Chapter 10.



---

# Chapter 9 Testing

## Associations I: Difference of Means, F-test, and $\chi^2$ Test

All the theory you had to suffer through in Chapter 8 (and all other theoretical chapters) was for the purposes of what we will do in this chapter and the next. All your efforts in introductory statistics will culminate in your ability to test bivariate associations for statistical significance — i.e., to make statistical inference about populations based on random samples.

Recall that we ended Chapter 7 with the knowledge that we describe/examine potential bivariate associations 1) between a discrete and a continuous variable through boxplots and difference of means, 2) between two discrete variables through contingency tables and difference of proportions, and 3) between two continuous variables through scatterplots and the correlation coefficient  $r$ , in a given dataset.

In this chapter and the next you will learn how to test these three types of bivariate associations for statistical significance, i.e. to check whether they can be generalizable to the population of interest. The current chapter is devoted to the first two types of bivariate associations. Chapter 10, the last chapter in this book,

offers a preliminary first glimpse into a powerful technique for multivariate inference (that can be used for variables at any level of measurement), called statistical regression — albeit we only cover the continuous two-variable case to serve as introduction.

Now that you know how hypothesis testing works, most of the associations testing will seem straightforward and somewhat formulaic: pose hypotheses, test hypothesis, make a decision regarding hypotheses, interpret findings in a substantive manner. The only thing that differs is the tests, as different type of associations generally require different tests. Regression is the one procedure that adds more, as it were, to this predictable pattern, but we will deal with it when we get there.

And then you will be done. So what are you waiting for? Gird up your loins for this last final push and let's get it over with!



---

## 9.1 Between a Discrete and a Continuous Variable: The t-test

For this part, you need to recall (from Section 7.2.1, <https://pressbooks.bccampus.ca/simplestats/chapter/7-2-1-between-a-discrete-and-a-continuous-variable/>) how we described bivariate associations between two variables, one of which is treated as discrete and one as continuous. In this case we essentially compared the groups (categories of the discrete variable) by their mean (or median) value on the continuous variable. We examine the potential association between such variables visually through boxplots and numerically through a difference of means.

Now the question in front of us is: even if we do see a difference in the means of the different groups *in sample data*, how certain can we be that this association is real and reflective of the population? As we learned in Chapter 8, to answer this question, we need to test the difference for statistical significance.

We start with a few theoretical notes, which we will then apply to the example I used in Chapter 7 about the potential gender difference in average income. In this way we will be able to test whether the difference observed in the *NHS 2011* data (\$16,401 in favour of men to be precise) is statistically significant or not. In the latter half of this section we will see what happens when there are more than two groups' means to compare.

**Testing the difference of two means.** Recall from Section 8.3 (<https://pressbooks.bccampus.ca/simplestats/chapter/8-3-hypothesis-testing/>) that we tested whether the employees who took a training course indeed had a higher average productivity by simply calculating the  $z$ -value (or, using the estimated standard error, the  $t$ -value with a given  $df$ ) for the mean and then finding its associated  $p$ -value. We could then compare the  $p$ -value to the preselected  $\alpha$ -level and make a conclusion regarding the null hypothesis.

You will be happy to know that testing a difference of means follows the same principle: obtain the  $z$  (or rather, the  $t$ -value), get the associated  $p$ -value, compare to the  $\alpha$ . What is not the same is that now we are testing expressly a difference of two means — so we need the  $t$ -value for the *difference*. It turns out, we can calculate one as easily as ever, as long as we had the standard error of the *difference*<sup>1</sup>.

**The standard error of a difference of two means is a combination of their separate standard errors:**

$$\sigma(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}} = \text{standard error of the difference of two means}$$

where the subscripts refer to the first and second group being compared.

The  $z$ -value for a difference of two means follows the ordinary  $z$ -value formula, but with the *difference* taking the place of the single mean:

1. I hope you have not forgotten that  $z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$ , where the standard error  $\sigma_{\bar{x}} = \frac{\sigma}{N}$ .

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sigma(\overline{x_1} - \overline{x_2})}$$

However, under the null hypothesis we hypothesize there is no difference in the population means, as such  $\mu_1 = \mu_2$ , and thus  $\mu_1 - \mu_2 = 0$ . Accounting for that in the formula, along with substituting the standard error with its own formula from above, we get:

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

Finally, since we generally don't know the population parameters but work with sample data, we estimate the standard error  $\sigma$  with the sample standard error  $s$ , thus moving to the ***t-value through which we test the difference for statistical significance:***

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} = \text{t-test for the difference of means}^2$$

$$t = \frac{\overline{x_1} - \overline{x_2}}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

Note that unlike the single value case where the  $df=N-1$ , when working with a difference of means of two groups the  $df=N-2$ .

2. The more observant of you would notice that the squared standard deviations of the two groups, i.e., the  $s_1^2$  and  $s_2^2$  here are of course the groups' variances (which we need if we are to have them under the square root). In this version of the formula, the groups are taken to have *unequal* variances, which is a more conservative assumption than assuming the variances of the two groups are equal. If we have a good reason to assume *equal* variances, then  $s_1^2$  and  $s_2^2$  will just be the same (combined, or pooled) variance  $s^2$ , and the formula will look like this:

Before your eyes glaze over (completely), rest assured that SPSS calculates this for you; I only provide it here to show you that the logic of hypothesis testing is the same, only the formulas change to accommodate the testing of a *difference of means* rather than a single mean.

From this point on, it's easy: you only need to check the  $p$ -value of the  $t$ -value you have obtained (given the specific  $df$ )<sup>3</sup>, and compare it to the significance level, and *voila* — you have yourself a significance test!

Let's see how this all works out in an example. A few sections back I promised you to test the gender differences in average income, didn't I?

*Example 9.1 Testing Gender Differences in Average Income, NHS 2011*

As in Example 7.2 in Section 7.2.1, I use a random sample of about 3 percent of the entire *NHS 2011* data, this time resulting in  $N=21,902$ <sup>4</sup>.

3. You can do that through an online  $p$ -value calculator for the  $t$ -distribution like this one here: <https://www.socscistatistics.com/pvalues/tdistribution.aspx>.
4. Since I use a new random sub-sample of the data, you can consider this an indirect illustration of sampling variation. For comparison of sample statistics as well as variable description, refer back to Example 7.2

We are still interested in whether women and men on average earn differently per year, i.e., whether *gender* affects *income*:

- $H_0$ : The average annual income of women and men is the same,  $\mu_m = \mu_f$
- $H_a$ : The average annual income of women and men is different,  $\mu_m \neq \mu_f$

There are 11,323 women ( $N_f=11,323$ ) and 10,579 men ( $N_m=10,579$ ) in the sample. The men earn an average of \$48,113 ( $\bar{x}_m = 48113$ ) and women earn an average of \$31,519 ( $\bar{x}_f = 31,529$ ). The respective standard deviations are \$68214 for men ( $s_m = 68214$ ) and \$34,760 for women ( $s_f = 34760$ ).

The difference of means is therefore:

$$\bar{x}_m - \bar{x}_f = 48113 - 31519 = 16594$$

The question is whether this \$16,549 is due to sampling variation (i.e., statistically not different than a population difference of means of \$0), or unusual enough so that a population mean of \$0 to be unlikely (i.e., so the difference is statistically significant).

To test this, we need to calculate the standard error of the difference. Once we have the standard error of the difference, we can calculate the *t*-value.

The standard error of the difference is:

$$\frac{s_{\bar{x}_m - \bar{x}_f}}{\sqrt{\frac{s_m^2}{N_m} + \frac{s_f^2}{N_f}}} = \sqrt{\frac{68214^2}{10579} + \frac{34760^2}{11323}} = \sqrt{439848 + 106708} = 739$$

The  $t$ -value is then:

$$t = \frac{\bar{x}_m - \bar{x}_f}{(\bar{x}_m - \bar{x}_f)} = \frac{16594}{739} = 22.446$$

Given the large  $N$ , even just looking at the  $t$ -value should make it clear that the difference is statistically significant — after all, in a two-tailed test, the  $t$ -value is significant at 1.96 and on (for  $\alpha=0.05$ ) and at 2.58 and on (for  $\alpha=0.01$ ).

Still, this is not the way to report a test — this is: **With a  $t=22.447$ ,  $df=21,900$ , and  $p=0.000^5$ , and  $p<0.001^6$ , we have enough evidence to reject the null hypothesis. Indeed, we can conclude with 99.99% certainty that there is a statistically significant difference between the average annual income of men and women (i.e., that the difference exists in the population).**

We can check this with a confidence interval too, again substituting the difference in place of a single value<sup>7</sup>:

5. You can check this with a  $p$ -value calculator; SPSS reports it too.

6. That is, the probability to observe a difference of \$16,594 in the sample if there were no difference in the population is smaller than 0.1%.

7. I hope you remember that 95% CI:  $\bar{x} \pm 1.96 \times s_{\bar{x}}$ .

$$\begin{aligned}
 95\% \quad \text{CI:} \quad \bar{x}_m - \bar{x}_f \pm 1.96 \times s_{\bar{x}_m - \bar{x}_f} &= \\
 16594 \pm 1.96 \times 739 = 16594 \pm 1448 &= \\
 = (15145; 18043)
 \end{aligned}$$

That is, we can say that **the difference of average annual incomes between men and women will be between \$15,145 and \$18,043 with 95% certainty; or that 19 out of 20 such studies will find a difference of \$16,594  $\pm$  \$1,448.** (We also see the correspondence with hypothesis testing: since the interval does *not* contain 0, 0 is not a plausible value for the difference.)

Inference is not doing too badly, no?

Again, SPSS will provide all the calculations but I advise you to still test your understanding of the procedure with the following exercise.

### *Do It!! 9.1 Gender Differences in Age of Actors in Main Roles*

Studies find that due to the gendered social construction of aging (i.e., women are considered “older” and “mature” at younger ages than men), male actors are frequently paired with much younger female actors (Buchanan 2013; Follows 2015). For example, the Oscars average age of male and

female Academy Award nominees is telling: in the Best Actor category, the average age of men is 43.4 years while the average age of women is 37.2 years (Beckwith & Hester, 2018 [http://thedataface.com/2018/03/culture/oscar-nominees-age]).

Let's say that you want to investigate this phenomenon yourself. You randomly select 100 male and 100 female academy award nominees, and calculate their age at nomination for an Academy Award. You find that men's average age is 45 years and women's is 36 years, with standard deviations of 15 years for men and 20 years for women. Test the hypothesis that the average age for women is different from that of men for the population of all Best Actor/Actress Oscar nominees. Create a 95% CI for the difference to see its correspondence with the hypothesis test.

Now that you understand the principle of testing the difference of two means, let's see what we can do about non-binary discrete variables, in the next section. The SPSS guidelines for doing a *t*-test are below.

#### *SPSS Tip 9.1 The t-test*

- From the *Main Menu*, select *Analyze*, and from



the pull-down menu, click on *Compare Means* and *Independent Samples T Test*;

- Select your continuous variable from the list of variables on the left and, using the top arrow, move it to the *Test Variable(s)* empty space on the right;
- Select your discrete variable from the list of variables on the left and, using the bottom arrow, move it to the *Grouping Variable* empty space on the right;
- Click on *Define Groups*, and in the new window, keep *Use specified values* selected; in the empty spaces for *Group 1* and *Group 2*, enter the *numeric values*<sup>8</sup> corresponding to the two categories of your discrete variable; click *Continue*.
- In the *Independent Samples T Test* window click *Options...*; you can request specific confidence interval in the new window (the default is 95%); click *Continue*;
- Click *OK* once back to the *Independent Samples T Test* window.
- SPSS will produce two tables in the *Output* window: a *Group Statistics* one (where you can see sample size, the mean, standard deviation, and standard error for each group (category in the discrete variable), and an *Independent Samples Test* one (where you can find the *t*-value, *df*, *p*-value, mean difference, standard

8. That would be the "code" -- for example, *gender* may be coded as "1 female, 2 male", or "0 male, 1 female", etc., depending on the dataset. You have to know this beforehand; if unsure, go back to Variable View and check.

error of the difference, and the requested confidence interval)<sup>9</sup>.

9. The table provides two versions of the test: *with* and *without* equal variances assumed.

Which one you should use depends on the size of the two groups' variances. If the variance of one groups is twice (or more) as big as the other group's variance (like in Example 9.1 above, where the men's variance was much larger than the women's one), use the test results in the bottom row, "equal variances not assumed". If the two groups' variances are relatively similar, you can use the top row, "equal variances assumed". You don't have to decide on your own, as SPSS provides a convenient indication for which one is better to use, under *Levene's Test/F* for comparing variances. If the *F*-test is significant (i.e.,  $p \leq 0.05$ ), the variances are too different and using the bottom row is better; if the *F*-test is non-significant (i.e.,  $p > 0.05$ ) you can assume the variances are equal and use the top row of results.

---

## 9.2 Between a Discrete and a Continuous Variable: The $F$ -test

When the discrete variable of interest has more than two categories, we can no longer use the simple  $t$ -test presented in the previous section. While we can still use a boxplot chart for visualizing the association between the two variables — where instead of two boxplots, we will have as many boxplots as there are groups (categories of the discrete variable) — we no longer have only one difference to test.

Testing multiple means for statistical significance is done through a version of a test called an  $F$ -test. This  $F$ -test tests whether the means of several groups<sup>1</sup> are all equal (versus at least one of them not being the same as the rest) through an analysis of variance (aka ANOVA).

At this point you might feel like a treatment of the topic of the kind I offered about the  $t$ -test above would be a tad too much, and you will be correct: providing the full-on technical details and the formula of the  $F$ -test is beyond the scope of this book.

Briefly, the ANOVA  $F$ -test calculates a ratio of variances (between groups to within groups, in terms of sums of

1. Note that "several groups" includes the two-groups case as well: you *could* test the significance of a difference between the means of two groups with an  $F$ -test too (it will just provide less information).

squares): the larger the ratio, the more evidence there is against the null hypothesis, and vice versa. The  $F$ -test statistic follows an  $F$ -distribution (not discussed here), which provides the  $F$ -value with its  $p$ -value, which is then compared to the  $\alpha$ -level and interpreted in the usual way. Example 9.2 illustrates.

*Example 9.2 Education Differences in Average Income, NHS 2011*

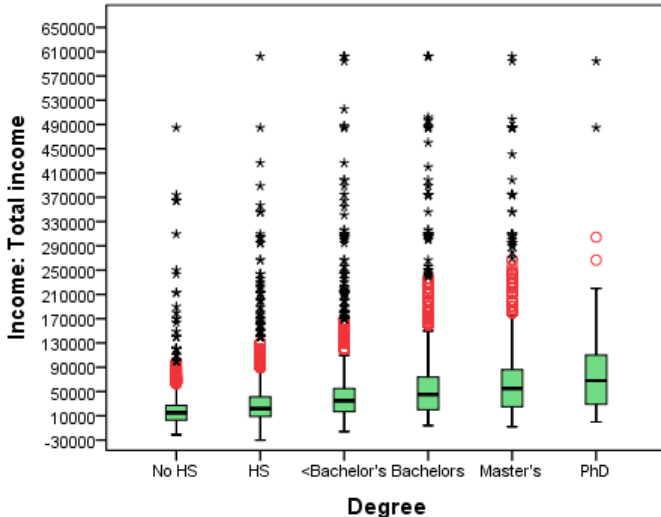
Presumably, college is worth it. You delay your full entry into the labour force and instead invest in your education, with the hope that you will then be able to have a better — and *better-paying* job.

Let's examine this questions then — do higher educational degrees translate into higher average income? — using about 3 percent random sample of the *NHS 2011* data. The variable *income* is the same one I used in previous occasions (i.e., *total income* in *NHS 2011*). The groups to compare are the categories of a variable called (highest) *degree*. The variable *degree* is a recoded version of the *NHS 2011's highest certificate, diploma or degree*. I recoded the original variable's thirteen categories in *degree's* six: 1) no high school, 2) high school, 3) certificate or diploma below Bachelor's, 4) Bachelor's, 5) Master's<sup>2</sup>

2. This category includes certificates above Bachelor's, and medical, dentistry, and veterinary degrees., and 6) PhD.

A brief descriptive investigation of the data reveals that the average income reported by the six education groups *looks* different: \\$19,433 for respondents without a high school degree, \\$30,455 for respondents with a high school degree, \\$41,971 for respondents with more than a high school but less than a Bachelor's degree, \\$60,360 for respondents with a Bachelor's degree, \\$71,593 for respondents with a Master's degree, and \\$93,924 for respondents with a PhD. This potential positive association (more education, more income) is also reflected in the boxplots in Figure 9.1. While there are outliers with extremely high average income in all groups (the most extreme were even truncated at the top), the median and the outlier-less maximum income increase from left to right with the increase of highest degree.

Figure 9.1 Average Income by Highest Degree, NHS 2011



Are these differences statistically significant? In other words, are the differences observed in the sample a result of regular sampling variation, or reflective of differences in the population?

- $H_0$ : The average income of all six education groups is the same.
- $H_a$ : The average income of some of the education groups is different from others.

SPSS reports a larger between-groups than within-groups variance;  $F=413.535$  with  $p<0.001$ . **With the probability of observing such differences between the groups in the sample — had there been no difference in the population (i.e., under the null hypothesis) — less than 1 in a thousand, we reject the null hypothesis and conclude that the differences in average income of groups with different highest degrees are statistically significant.**

Before we turn to testing associations between two discrete variables, the SPSS Tip 9.1 below lists the steps of the  $t$ -test and ANOVA  $F$ -test procedures.

#### *SPSS Tip 9.2 The F-test*

- From the *Main Menu*, select *Analyze*, and from

the pull-down menu, click on *Compare Means* and then *One-Way ANOVA*;

- Select your continuous variable from the list of variables on the left and, using the top arrow, move it to the *Dependent List* empty space on the right;
- Select your discrete variable from the list of variables on the left and, using the bottom arrow, move it to the *Factor* empty space on the right; click OK.
- The *Output* window will present a *Oneway ANOVA* table, listing a breakdown of variances (by sums of squares), and most importantly, the resulting *F*-statistics and *p*-value.





---

## 9.3 Between Two Discrete Variables: The $\chi^2$ , Part 1

As in the previous section, here you need to recall how we examine potential association between two variables both treated as discrete (Section 7.2.2, <https://pressbooks.bccampus.ca/simplestats/chapter/7-2-2-between-two-discrete-variables/>). We described such associations through contingency tables, reporting differences of proportions as appropriate.

We can start with the simplest, binary case: when the discrete variables have two groups each. Then we compare the groups of interest (categories of one variable) on one of the categories of the other variable. (The example in Chapter 7 we used was to compare the percentage of first-year students who like the campus cafeteria to the percentage of second-year students who do.)

**The  $t$ -test for testing difference of two proportions.** When we have only two proportions (or percentages) to compare, we can actually use the same  $t$ -test we used for testing differences of means, again treating the *difference* as a single, normally distributed statistic. Since we have categorical variables, however, and no standard deviations/variances, we resort to measuring population variability by  $\pi(1-\pi)$  and sample variability by  $p(1-p)$ <sup>1</sup>. (See Section 6.7.2, <https://pressbooks.bccampus.ca/simplestats/chapter/>

1. Do not forget that  $p$  here stands for *proportion*, not *probability*/ $p$ -value.

[6-7-2-confidence-intervals-for-proportions/.](#)) We can thus simply substitute that into the formula for  $z$ :

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{N_1} + \frac{\pi_2(1-\pi_2)}{N_2}}}$$

where, of course, under the null hypothesis  $(\pi_1 - \pi_2) = 0$ . Then, using the sample proportions leaves us with  $t$ :

$$t = \frac{(p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}}$$

Again, under the null hypothesis the two groups' proportions are assumed to be the same so effectively we have:

$$t = \frac{(p_1 - p_2)}{\sqrt{p(1-p)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Let's revisit the cafeteria-preferences example from Section 7.2.2 to see how the  $t$ -test for testing difference of proportions works.

*Example 9.3 Do You Like the Campus Cafeteria? (A  $t$ -Test)*

In Chapter 7 we imagined that you asked 35 students in

your class<sup>2</sup> whether they liked the campus cafeteria: 12 of your classmates said yes (i.e., 34.3 percent), 7 (out of 15) first-years and 5 (out of 20) second-years (46.7 percent of all first-years and 25 percent of all second-years, respectively).

We want to know whether the observed in the sample difference in proportions ( $0.467 - 0.25 = 0.217$ ) is statistically significant: can it be generalized to a larger student population, or is it due to a regular sampling variability?

- $H_0$ : The proportion of first year students who like the cafeteria is the same as the proportion of second year students who do;  $\pi_1 = \pi_2$ .
- $H_a$ : The proportion of first year students who like the cafeteria is different than the proportion of second year students who do;  $\pi_1 \neq \pi_2$ .

Substituting these numbers in the formula we have:

$$t = \frac{(p_1 - p_2)}{\sqrt{p(1-p)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} = \frac{0.467 - 0.25}{\sqrt{0.343(1-0.343)\left(\frac{1}{15} + \frac{1}{20}\right)}} = \frac{0.217}{0.162} = 1.34$$

**With a  $t=1.34$ ,  $df=34$ , and  $p=0.189$  (i.e.,  $p>0.05$ ) we *fail* to reject the null hypothesis: at this point we do not have enough evidence to conclude there is a difference between the proportions of first and second year students who like the campus cafeteria. The 21.7 percentage points difference is not statistically**

2. Note that this of course is not a random sample; we are using it here only for illustrating how hypothesis testing works so we are effectively pretending it is random. In a real-life study, you should not use non-probability samples for statistical inference.

significant, and has a high enough probability of being due to random chance.

We can check this with a confidence interval too:

- 95% CI:

$$\frac{(p_1 - p_2) \pm 1.96 \times \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}}{0.217 \pm 0.316} = \frac{0.217 \pm 1.96 \times \sqrt{\frac{0.467(0.533)}{15} + \frac{0.25(0.75)}{20}}}{0.217 \pm 0.316} = (-0.099; 0.533)$$

**In other words, the difference between the proportion of first years and the proportion of second years who like the cafeteria could be anywhere between -9.9 percentage points and 53.3 percentage points with 95% confidence (or 19 out of 20 such samples will have a difference within this pretty large interval).** The difference can be in favour of second years or in favour of the first years (notice the negative lower bound); it can even be 0. Thus, **since a difference of 0 (i.e., no difference) is a plausible value, we cannot reject the null hypothesis. We conclude that we do not have enough evidence of an association between year of study and opinion on the campus cafeteria.**

Admittedly, the formulas look scary but if you have followed through the example above, you have seen by now the actual calculation is quite simple. You can try it out and see for yourself.

*Do It! 9.2 Vegetarianism/Veganism among Canadian and International Students*

Imagine you are interested in exploring whether there is a difference between Canadian and international students in your university when it comes to dietary preferences like vegetarianism and veganism. With your institution's registrar's assistance, you take a random sample of 100 students and poll them on 1) whether they are a Canadian or an international student, and 2) whether they are vegetarian/vegan or not.

You find that you have 70 Canadian and 30 international students in your sample. Out of the Canadian students, 15 (or 21.4 percent) are vegetarian or vegan; out of the international students 5 (or 16.7 percent) have such dietary restrictions.

Check if the observed *in the sample* difference in proportions is generalizable to the larger student population by testing the hypothesis whether dietary preferences are associated with country of origin. Create a 95% confidence interval for that difference, and substantively interpret what you have found with both the *t*-test and the confidence interval.

Useful hint 1: Among the 100, there are 20 vegan/vegetarian students in total.

Useful hint 2: You can find the *p*-value of your *t*-statistic here: <https://www.socscistatistics.com/pvalues/tdistribution.aspx>.

Of course, discrete variables do not have to be binary:

they can have more than two categories each. Just like in the case of a continuous and a discrete variables' association discussed in the previous section where non-binary variables required the use of an  $F$ -test, there is a different test for testing the association between any two discrete variables, regardless of their respective number of categories (i.e., not just binary ones).

**The  $\chi^2$ -test for testing associations between discrete variables.** The  $\chi^2$ -test<sup>3</sup> (or Pearson's  $\chi^2$ -test) is based on a **comparison between the *observed* and the *expected* cell values in a contingency table.**

The observed values are the cell counts you see in a contingency table given a specific dataset. The expected values, on the other hand, are the counts we would *expect* to see *if there were no pattern/association in the data*. In other words, the test effectively compares the sample to a null-hypothesis-like hypothetical distribution of the observations across the cells. Thus, logically, **if there is a relatively large difference between the observed and the expected values, we can take that as evidence *against* the null hypothesis and reject it. If, however, the difference between observed and expected values is relatively small, the evidence against the null hypothesis will be insufficient and we would *fail* to reject it.**

The actual way the  $\chi^2$  is calculated is this:

3. This is the small-case Greek letter  $\chi$ . It is pronounced [KHAI], but since it is transliterated as *chi*, many people incorrectly pronounce it as [CHAI] or even [CHEE]. The test itself is called chi-squared test (again, pronounced as [KHAI- squared] not [CHAI- or CHEE-squared]).

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where  $f_o$  is the observed frequency (count) and  $f_e$  is the expected frequency count of a given cell.

The formula looks more complicated than it is (don't they always?) — it only asks us to calculate the difference between the observed and the expected count *for each cell*, square it and divide it by the expected count. Once we have done this for all cells, we need only add the resulting numbers together to get the  $\chi^2$ .

Considering that the  $\chi^2$  is then a sum of as many numbers as there are cells, the larger the table (i.e., the more rows and columns there are), the bigger the resulting  $\chi^2$  will be. To account for that, the  $\chi^2$  too has degrees of freedom, where the  $df = (\text{rows}-1)(\text{columns}-1)$ . The  $\chi^2$  follows a  $\chi^2$ -distribution, which too provides a  $p$ -value given specific  $df$ .

**The hypothesis testing then follows the same steps as the  $t$ -test and the  $F$ -test: obtain  $\chi^2$ -value with specific  $df$ , find its associated  $p$ -value, and finally compare the  $p$ -value to the pre-selected significance level. If  $p < \alpha$ , reject the null hypothesis.**

To demonstrate, we will first do a *one-way*  $\chi^2$  calculation, i.e., based on the frequency distribution of just *one* variable. (Of course, if tabulated, this would not be considered a contingency table but a frequency table.)

*Example 9.4 Do You Like The Campus Cafeteria? (Univariate  $\chi^2$ -Test)*

To use the imaginary data from before, we had 12 people who admitted liking the campus cafeteria food out of the 35 polled. (Since we are interested only in one of the variables, here we ignore whether the students who like the cafeteria are first- or second-years.) As such, we have the following table:

*Table 9.1 Approval of the Campus Cafeteria, Observed Count (Univariate)*

<b>Yes</b>	12
<b>No</b>	23
<b>Total</b>	35

If you did not know anything about the campus cafeteria and had no observations about it whatsoever — i.e., had you been an impartial observer, as it were — wouldn't you expect to see an approximately 50/50 split of the 35 students into the two categories? After all, there are only two groups, and an unbiased (random) distribution would be exactly like everyone flipping a coin as a manner of deciding in which group they end up. Thus, **the expected count here is simply  $N$  divided by the number of groups/categories** (denoted by  $k$ ):



$$f_e = \frac{N}{k} = \frac{35}{2} = 17.5$$

Table 9.2 adds the expected count in brackets next to the observed count.

*Table 9.2 Approval of the Campus Cafeteria, Observed and Expected Count (Univariate)*

<b>Yes</b>	12	(17.5)
<b>No</b>	23	(17.5)
<b>Total</b>	35	

Then, according to the formula, this is what we have for each of the two groups:

- **Yes-group:**  

$$\frac{(f_o - f_e)^2}{f_e} = \frac{(12 - 17.5)^2}{17.5} = \frac{30.25}{17.5} = 1.73$$
- **No-group:**  

$$\frac{(f_o - f_e)^2}{f_e} = \frac{(23 - 17.5)^2}{17.5} = \frac{30.25}{17.5} = 1.73$$

Finally, to get the  $\chi^2$  we only need to add these two numbers together:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(12 - 17.5)^2}{17.5} + \frac{(23 - 17.5)^2}{17.5} = 1.73 + 1.73 = 3.46$$

The degrees of freedom in a one-way  $\chi^2$ -test is  $k-1$ , where  $k$  is the number of categories/groups. In this case we have  $k=2$ , so  $df=1$ .

**With a  $\chi^2=3.45$ ,  $df=1$ , and a  $p=0.06^4$  (i.e.,  $p>0.05$ ), we fail to reject the null hypothesis. At this time, we do *not* have enough evidence to conclude that the observed distribution of the students is unusual enough to suggest a pattern which is different than a random variation of a 50/50 split. As such, this distribution is *not* statistically significant — we cannot conclude that the students lean one way or the other in their opinion about the campus cafeteria.**

Calculating a two-way  $\chi^2$  — by far the more often used one as it tests associations between *two* variables — is just as easy, even if it involves calculating more numbers (since in the bivariate case we have more cells; four at the minimum, given a  $2 \times 2$  cross-tabulation). The next section is devoted to that.

4. You can check the significance of any  $\chi^2$  with a convenient online calculator, like this one here: <https://www.socscistatistics.com/pvalues/chidistribution.aspx>.

---

## 9.4 Between Two Discrete Variables: the $\chi^2$ , Part 2

Calculating a two-way  $\chi^2$  is only marginally more complicated than the one-way case we examined in the previous section, as Example 9.5 demonstrates.

*Example 9.5 Do You Like The Campus Cafeteria? (Bivariate  $\chi^2$ -Test)*

While we already know that year of study and opinion on the campus cafeteria are not statistically associated from the  $t$ -test in Example 9.3, I will further use the imaginary data in the original contingency table from Example 7.3 to demonstrate a two-way  $\chi^2$ -test. This was the table we had in Section 7.7.2.

*Table 9.2 (A) Do You Like The Campus Cafeteria? (Revisited)*

	First Year Students	Second Year Students	Total
YES	7	5	12
NO	8	15	23
Total	15	20	35

Our hypotheses are:

- $H_0$ : Liking the cafeteria or not is not associated with one's year of study; first- and second-year students are equally likely to like the cafeteria, or  $\pi_1 = \pi_2$ .
- $H_a$ : Liking the cafeteria is associated with one's year of study; first-year students and second-year students differ in their liking of the cafeteria, or  $\pi_1 \neq \pi_2$ .

To compute the  $\chi^2$ , we need the expected count for each cell. Unlike the one-way  $\chi^2$  case, however, determining the expected count in a contingency table is a bit more complicated than dividing the  $N$  on the number of groups and expecting the same (expected) number in each cell. Instead, we multiply the respective group/category sizes (i.e., the row total and the column total at the margins) and divide the product by  $N$  (the full total)<sup>1</sup>

$$f_e = \frac{N_j \times N_k}{N}$$

1. We do that to account for the different group/category sizes.:

where  $j$  is the size of the respective group and  $k$  is the size of the respective category<sup>2</sup>.

Thus we have the following:

- First-years who said *Yes*:

$$f_e = \frac{N_j \times N_k}{N} = \frac{15 \times 12}{35} = 5.14$$

- Second-years who said *Yes*:

$$f_e = \frac{N_j \times N_k}{N} = \frac{20 \times 12}{35} = 6.86$$

- First-years who said *No*:

$$f_e = \frac{N_j \times N_k}{N} = \frac{15 \times 23}{35} = 9.86$$

- Second-years who said *No*:

$$f_e = \frac{N_j \times N_k}{N} = \frac{20 \times 23}{35} = 13.14$$

Table 9.2 (B) adds the expected count in brackets next to the observed count.

*Table 9.2 (B) Do You Like The Campus Cafeteria?  
(Observed and Expected Frequencies)*

2. Recall that to differentiate between the groups/categories of the two variables, we refer to one variable having groups and the other having categories: so that we can say we compare the groups of one variable on the categories of the other.

	<b>First Year Students</b>	<b>Second Year Students</b>	<b>Total</b>
<b>YES</b>	7 (5.14)	5 (6.86)	12
<b>NO</b>	8 (9.86)	15 (13.14)	23
<b>Total</b>	15	20	35

Now we only need calculate the four elements of the  $\chi^2$  and add them altogether at the end.

- First-years who said *Yes*:  

$$\frac{(f_o - f_e)^2}{f_e} = \frac{(7 - 5.14)^2}{5.14} = 0.67$$
- Second-years who said *Yes*:  

$$\frac{(f_o - f_e)^2}{f_e} = \frac{(5 - 6.86)^2}{6.86} = 0.5$$
- First-years who said *No*:  

$$\frac{(f_o - f_e)^2}{f_e} = \frac{(8 - 9.86)^2}{9.86} = 0.35$$
- Second-years who said *No*:  

$$\frac{(f_o - f_e)^2}{f_e} = \frac{(15 - 13.14)^2}{13.14} = 0.26$$

Finally,

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 0.67 + 0.5 + 0.35 + 0.26 = 1.78$$

The degrees of freedom are,

again,  $df=(rows-1)(columns-1)$ , so  
 here  $df=(2-1)(2-1)=1(1)=1$ .

That is, with  $\chi^2=1.78$ ,  $df=1$ , and  $p=1.18$  (i.e.,  $p>0.05$ ), we **do not** have enough evidence to reject the null hypothesis. At this time, we *cannot* claim there is an association between year of study and opinion on the cafeteria, i.e., the 0.217 difference in proportions we observe in the sample (7/15 versus 5/20, or 0.467 versus 0.25) is *not* statistically significant.

Of course, we already knew this from the  $t$ -test in Example 9.3<sup>3</sup>, so no surprises here.

The imaginary example above serves well as a work-through for calculating  $\chi^2$ , but we can do better — an example using real, random-sample data and a large  $N$  is in order.

If you recall, in Section 7.7.2 we also explored gender differences in the ability to speak an Aboriginal language using *APS 2012* (Statistics Canada, 2019) data. Armed with knowledge about the  $\chi^2$ , now we can finish that investigation.

3. You may find it curious to know that the correspondence of results between the  $t$  and the  $\chi^2$  goes even further: in the binary variables' case, squaring the  $t$ -value will give you exactly  $\chi^2$ :  $t^2=\chi^2$ . In our examples,  $t=1.34$ , and  $1.34^2=1.79$  which, if it was not for rounding, would be the same as  $\chi^2$ . Even their respective degrees of freedom are the same, 1.8. This of course is not the case when at least one of the discrete variables has more than two categories.

*Example 9.6 Testing Gender Differences in the Speaking Aboriginal Language Ability among Indigenous Canadians , APS 2012*

Our exploration in Section 7.2.2 left us with the following table.

*Table 9.3 Speaking Aboriginal Language Ability by Gender, APS 2012 (Revisited)*

Lang. - Speaking Aboriginal language * Sex of respondent Crosstabulation					
			Sex of respondent		Total
			MALE	FEMALE	
Lang. - Speaking Aboriginal language	Yes	Count	4877	5672	10549
		% within Sex of respondent	41.4%	45.0%	43.3%
	No	Count	6898	6933	13831
		% within Sex of respondent	58.6%	55.0%	56.7%
Total	Count		11775	12605	24380
	% within Sex of respondent		100.0%	100.0%	100.0%

Our hypotheses are:

- $H_0$ : Gender and the ability to speak an Aboriginal language are not associated; women and men are equally likely to speak an Aboriginal language, or  $\pi_f = \pi_m$ .
- $H_a$ : Gender and the ability to speak an Aboriginal language are associated; women and men are not equally likely to speak an Aboriginal language, or  $\pi_f \neq \pi_m$ .



SPSS calculates  $\chi^2$  as 31.78. With  $\chi^2=31.78$ ,  $df=1$ , and  $p<0.001$ , we have enough evidence to reject the null hypothesis and conclude that Indigenous women and men tend to differ in their ability to speak an Aboriginal language. The 3.6 percentage points difference (i.e., 45 percent minus 41.4 percent) in favour of women being more likely to speak an Aboriginal language is statistically significant and therefore generalizable to the larger Indigenous population.

I “cheated” out of presenting the actual calculations in the example above to give you the opportunity to do it on your own. Use it as an exercise in practicing your understating of the  $\chi^2$  and  $t$  statistical significance tests.

*Do It! 9.3 Testing Gender Differences in the Speaking Aboriginal Language Ability among Indigenous Canadians, APS 2012*

Using the information presented in Table 9.3 above, 1) calculate the expected frequencies for each cell and compute the  $\chi^2$ ; and 2) do a  $t$ -test on the difference of proportions and create a 95% confidence interval for the difference, to observe the correspondence between the different tests.

Finally, lest I leave you with the impression that there is no difference between using a  $t$ -test and a  $\chi^2$ -test, let's consider a case where both variables have more than two categories, next.

---

## 9.5 Between Two Discrete Variables: the $\chi^2$ , Part 3

We definitely need to use the  $\chi^2$  for testing contingency tables when at least one of the variables has more than two categories, as we no longer have only two proportions to consider.

### *Example 9.7 Citizenship and Education, NHS 2011*

A lot has been written about Canada's selective immigration practices: the Canadian government is committed to getting "the best and the brightest" immigrants through a point system which awards more points the more education the prospective immigrant has. [CITATIONS] Be that as it may, how does the rest of the Canadian population (the one born in Canada) compare to the supposedly highly-educated foreign-born? With the help of *NHS 2011* (Statistics Canada, 2019), we can find out. (Note that once again, I will use about 3 percent random sub-sample of the data, for an  $N=21,577$ .)

For this example I use the variable *citizenship* which has three categories: "born in Canada", "naturalized Canadian",

and “not a Canadian citizen”. For education, I use the same recoded variable I used in Example 9.2 in Section 9.2 earlier, namely *degree*. Degree has six categories, ranging from (1) “no high school degree” to (6) “PhD” (for full category listing, see Example 9.2).

Table 9.4 cross-tabulates citizenship and degree in a busy-looking 3×6 table (that’s 18 cells!).

*Table 9.4 Degree by Canadian Citizenship Status (NHS 2012)*

Degree * Citizenship status and type						
		Citizenship: Citizenship status and type - Summary				
			Canada, by birth	Canada, by naturalization	Not a Canadian citizen	Total
Degree	No HS	Count	3307	746	210	4263
		% within Citizenship: Citizenship status and type - Summary	20.4%	18.6%	15.8%	19.8%
	HS	Count	4326	895	307	5528
		% within Citizenship: Citizenship status and type - Summary	26.7%	22.3%	23.1%	25.6%
	>HS, <Bachelor's	Count	5630	1214	342	7186
		% within Citizenship: Citizenship status and type - Summary	34.7%	30.2%	25.7%	33.3%
	Bachelor's	Count	1998	673	273	2944
		% within Citizenship: Citizenship status and type - Summary	12.3%	16.7%	20.5%	13.6%
	Master's (+medical,etc.)	Count	893	451	179	1523
		% within Citizenship: Citizenship status and type - Summary	5.5%	11.2%	13.5%	7.1%
	PhD	Count	72	42	19	133
		% within Citizenship: Citizenship status and type - Summary	0.4%	1.0%	1.4%	0.6%
Total		Count	16226	4021	1330	21577
		% within Citizenship: Citizenship status and type - Summary	100.0%	100.0%	100.0%	100.0%

What do we see? Let’s carefully examine the evidence<sup>1</sup>.

1. Do not forget to focus on the percentages, not the number count in each cell! Recall that you can only compare relative frequencies (relative to group size, that is).

While all citizenship groups follow a similar vertical “spread” (i.e., relatively few people without degrees, most people with high/secondary school and some post-secondary school certificates and some sort of Bachelor’s degrees, then decreasing proportions in the higher education categories), this is *not* what we are interested in. Recall that we are looking for a pattern between the two variables’ categories/groups — we are comparing groups on their levels of education.

As such, we see that fewer naturalized Canadians (18.6 percent) and fewer still non-Canadian citizens (15.8 percent) have no degrees compared to Canadian citizens (20.4 percent). Furthermore, in the three highest education categories (Bachelors, Master’s, and PhD), both naturalized Canadians and non-Canadian citizens outperform those born in Canada (while the non-Canadian citizens even outperform naturalized Canadians in turn): 16.7 percent of naturalized Canadians and 20.5 percent of non-Canadian citizens have Bachelor’s degrees compared to only 12.3 percent of Canadians born in the country; 11.2 percent of naturalized Canadians and 13.5 percent of non-Canadian citizens have Master’s degrees compared to only 5.5 percent of the ones born in Canada; and, finally, 1 percent of naturalized Canadians and 1.4 percent of non-Canadian citizens have PhD’s compared to 0.4 percent of those born in Canada.

Thus, the table suggests a pattern — Canadians born elsewhere and non-Canadian citizens seem to have more education than the Canadian-born. Whether this pattern showing difference in proportions in the education degrees among the different citizenship status groups is statistically significant (i.e., generalizable to the Canadian population) remains to be checked — through a  $\chi^2$ -test.

These are our hypotheses:

- $H_0$ : Citizenship status and educational degree are not associated; Canadian-born, naturalized citizens, and non-Canadian citizens are on average similarly educated, and are equally likely to be highly educated.
- $H_a$ : Citizenship status and educational degree are associated; Canadian-born, naturalized citizens, and non-Canadian citizens have different levels of education on average, and are not equally likely to be highly educated.

I would guess you would rather not calculate the expected frequencies and their differences from the observed frequencies for all 18 cells (but if you want to do it, who am I to stop you), so I'll report the SPSS output instead.

**With  $\chi^2=449.543$ ,  $df=10^2$ , and  $p<0.001$ , we have enough evidence to reject the null hypothesis and conclude that citizenship status and educational degree are statistically significantly associated: people born in Canada, naturalized Canadians, and non-Canadian citizens differ in their levels of education.** It seems indeed that Canadians born in the country are on average less educated than both naturalized Canadians and non-Canadian citizens, perhaps as a result of the selective criteria for Canadian immigration.

**Important conditions for using the  $\chi^2$ -test.** For the  $\chi^2$ -test to work properly, two conditions must be met:  
1) the expected count should not be less than 1 for any of

$$2. Df=(rows-1)(columns-1)=(6-1)(3-1)=5(2)=10.$$

the contingency table cells; and 2) no more than 20 percent of the cells should have an expected count less than 5. SPSS warns you about violations of these conditions in the output; if you are not using SPSS you should make sure the conditions are met before proceeding with analysis. Either way, if these conditions are not met, you should not use the  $\chi^2$ -test and consider a different type of testing instead (not discussed here).

Finally, a brief word of warning.

**Watch Out!! #17 ... for Identifying The Wrong Pattern**

Once again, the warning is about how to read a contingency table *in light of an association between two variables*. The pattern (association) in which we are interested and the one we test is a comparison between the groups of one variable on the categories of the other variable. Thus, looking at how the observations are divided within each group is only marginally relevant to the research question, and does not contribute to analyzing the association in question.

In Example 9.7 above, all immigration status groups were divided relatively similarly across the educational categories but, as interesting as you may find this “pattern”, that is *not* an indication of an association — comparing the percentages/proportions of the different groups in the same category is. In other words, in that example we were

interested in whether there was a *difference in percentages/proportions among the Canadian-born, naturalized Canadians, and non-Canadians citizens* with no education, or with high school degree only, or with some college degree or certificate only, or with Bachelor's degree, etc. We were *not* interested in what percentage of Canadian-born (or naturalized, or non-Canadian citizens) have no degree, *and* what percentage have high school degree, *and* what percentage have some college degree or certificate, etc. (if you recall, the latter add up to a 100 percent, and can be referred to as how the observations are spread across categories *within* each group).

As in Section 7.2.2, what it comes down to is knowing which way to read the table, according to the research question you have<sup>3</sup>

We finish the chapter with the tips on using SPSS for  $\chi^2$ -testing.

3. I remind you again of the rule of thumb: if the groups you are comparing are in the *columns*, and the percentages down the columns add to 100 percent, then look at and compare the percentages/proportions on the same row. If the groups you are comparing are in the *rows*, and the rows add up to 100 percent, then compare the percentages down the same column.. To use the language of causality, **to the extent that you can identify an independent and a dependent variable, to examine an association between the variables you will be looking to compare the groups of the independent variable on the categories of the dependent variable.**



SPSS Tip 9.3 The  $\chi^2$ -test

- From the *Main Menu*, select *Analyze*, and from the pull-down menu, click on *Descriptive Statistics* and then *Crosstabs*;
- From the variable list on the left, select your variable (the independent variable, with groups to be compared) and, using the bottom arrow, move it to the *Column(s)* empty space on the right;
- From the variable list on the left, select your variable (the dependent variable, on whose categories you will compare the groups) and, using the bottom arrow, move it to the *Row(s)* empty space on the right<sup>4</sup>;
- Click on *Statistics* and select *Chi-square* at the top of the new window, click *Continue*;
- Once back in the *Crosstabs* window, click *Cells*; in the new window keep *Observed*<sup>5</sup> in *Counts* selected, and further select *Column* in *Percentages*; click *Continue*;
- Once back to the *Crosstabs* window, click *OK*.

4. Again, the convention is to put the independent variable in the columns and the dependent variable in the rows. This is not a hard-set rule, however, and it is perfectly acceptable to do it the opposite way. The only thing that is *not* a matter of preference is for which percentages you should ask, *columns* or *rows*. *If your independent variable is in the columns, you need column percentages to compare, if your independent variable is in the rows, you need row percentages to compare*. In this latter case, this is a hard-set rule, and if you violate it, you will not be able to properly identify -- and test -- the association you might be investigating.
5. Note that from here you can also request *Expected* counts if you would like to check them at any point.

- SPSS will provide the requested output in the Output window: a contingency table followed by a  $\chi^2$ -test table, containing the  $\chi^2$ -value,  $df$ , and  $p$ -value<sup>6</sup>.

With this, we turn to our last remaining topic: the testing and investigation of the association between two continuous variables in Chapter 10, next.

6. Note that the table contains more than just the  $\chi^2$ -test; discussing the rest of the tests is beyond the intended scope of this book.

---

# Chapter 10 Testing

## Associations II: Correlation and Regression

This is it: you are finally here, reading the *last chapter*. (And after nine chapters, what's just one more?) This is not a heavy chapter as some of the others but regression is sufficiently different from the type of testing about which you learned in Chapter 8 to deserve a heads-up — so if you find yourself despairing at some point, just remind yourself that *this is it*; once you've learned *this*, you will have a passing knowledge about *how*, *what for*, and *why* statistics is used in sociological research, and you will also be able to do some basic analysis on your own! — and you'll be done in no time.

Pep talk aside, for this chapter you should review/recall Chapter 7, and specifically Section 7.2.3 (<https://pressbooks.bccampus.ca/simplestats/chapter/7-2-3-between-two-continuous-variables/>) on examining bivariate associations between two variables, treated as continuous for the purposes of statistical analysis. We did that visually through a scatterplot (with a line of best fit) and numerically through the coefficient of correlation called Pearson's  $r$ .

In this chapter, you will learn what  $r$  actually is, and that

it has its own  $t$ -test and a  $p$ -value to test its significance. In addition, I will present a relatively brief and basic introduction into the topic of regression, a powerful and versatile technique with truly impressive number of applications which readily allows for doing *multivariate* analysis.

After all, recall that when we do bivariate analysis, we ignore the complexity of the real world where variables are/may be tangled in a veritable web of almost endless interrelationships. With bivariate analysis we ignore all that to focus on how just *two* variables are statistically associated. But because of that, we cannot say anything about *causality* as we cannot account for additional variables that could serve as alternative explanations to what we observe. And while multivariate regression cannot *completely* do that either (in the social sciences establishing causality is a pretty tall order), with careful assumptions and the right specifications, it can help bring us more than a few steps further in that direction.

Of course, even if I haven't already told you, you would have been able to tell by now that multivariate regression analysis falls beyond the scope of what we do here. What follows is a necessary stepping stone, however; once you have the right idea about how regression works with two continuous variables, everything else *regression* follows the same basic principle and thus can be built on top of the foundation you will have by the end of this chapter (and book!).

So, ready? Let's go then! The end is just several sections away!

---

## 10.1. Correlation

You will recall from Section 7.2.3 that we use the coefficient of correlation (Pearson's)  $r$  to examine associations between two continuous variables. The correlation coefficient  $r$  varies between -1 and 1. The closer it is to either, the stronger the correlation, and the closer it is to 0, the weaker the correlation<sup>1</sup>.

Where does  $r$  come from though? What does it actually measure? I doubt you have lost sleep wondering about these questions which I left unanswered in Chapter 7, but here is your chance to learn this anyway (think of it as closure of sorts).

The correlation coefficient is, essentially, a ratio of the variabilities of the two variables<sup>2</sup>

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

1. The sign of  $r$  is there *only* to indicate the direction of the association: positive or negative, nothing else. Thus this is a reminder not to use  $r$ 's sign as a measure of magnitude or strength of the association. Thus, for example, -0.9 is a stronger association than 0.2 because -0.9 is closer to -1 than 0.2 is to 1. (In fact, 0.2 is much closer to 0, or no association.) That is, a strong negative correlation is *stronger* than a weak positive one, despite that  $-0.9 < 0.2$ .
2. To be precise, the ratio is between the covariance of  $x$  and  $y$  (i.e., their joint variability,  $s_{xy}$ ) and the product of their separate variances  $s_x$  and  $s_y$ :

or

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

if we apply it to a population instead of a sample. (Here  $\rho$  is the small-case Greek letter  $r$ , pronounced [ROH].)

**The easiest way to calculate  $r$  between a variable  $x$  and a variable  $y$  is through the distances of the observations from the means of the two variables, or more accurately, the sums of squares<sup>3</sup> before adding to turn them all positive, otherwise they'd cancel each other upon summation.** See Section 4.3 (<https://pressbooks.bccampus.ca/simplestats/chapter/4-3-variance/>) for details.) :

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}}$$

From Section 4.3, we know that  $\Sigma(x - \bar{x})^2$  is the sum of squares of the variable  $x$  (so,  $SS_x$ ); by analogy,  $\Sigma(y - \bar{y})^2$  will be the sum of squares of the variable  $y$  (so,  $SS_y$ ). When the distances between an observation and the two means are “cross-multiplied” before summing (like in the

3. Recall that the sum of squares was the numerator in the formulas for the variance and the standard deviation. We take the distances of the observations from the mean, square them, and then add them altogether. (We square them

numerator), they are called the sum of products ( $SP_{xy}$ ).

Thus we can restate the formula above in the following simplified (and easier to remember) way<sup>4</sup>  $r$  exist. All of them calculate the same  $r$ , but are just restated in different term. The two "versions" presented in the text above are the simplest. For example, one of the most common ways to express  $r$  you may find elsewhere (but which is rather hard on the eyes and for purposes of calculation by hand) is this:

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{N\sum x^2 - (\sum x)^2}(N\sum y^2 - (\sum y)^2)}$$

:

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

Example 10.1(A) provides an empirical application of  $r$ 's calculation.

*Example 10.1(A) Education and Parental Education, GSS 2018*

4. Note that other "versions" of the formula for

Table 10.1 lists the years of schooling (our variable  $y$ ) of seven respondents in the *GSS 2018* (NORC, 2019) and the years of schooling of their respective fathers (our variable  $x$ )<sup>5</sup>. While inference with  $N=7$  is not a serious proposition, the small observation count allows for a quick calculation for demonstration purposes only. (After all, we already know the correlation coefficient of these exact same two variables from Section 7.2.3; there the SPSS-calculated  $r$  was equal to 0.413.)

The rest of the columns in Table 10.1 list the necessary computations (obtaining distances from the mean, squaring distances, summing distances, etc.) to produce  $SS_x$ ,  $SS_y$ , and  $SP_{xy}$ .

*Table 10.1 Calculating Pearson’s  $r$*

5. Here *parental education* is the independent variable and *respondent’s education* is the dependent variable, so they are denoted as  $x$  and  $y$ , respectively, according to convention.



$x$	$y$	$(x - \bar{x})$	$(x - \bar{x})^2$	$(y - \bar{y})$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
12	8	$(12-12.4) = -0.4$	$-0.4^2 = 0.2$	$(8-13.6) = -5.6$	$5.6^2 = 31.4$	$(-0.4)(5.6) = -2.2$
6	12	$(6-12.4) = -6.4$	$-6.4^2 = 41$	$(12-13.6) = -1.6$	$-1.6^2 = 2.6$	$(-6.4)(1.6) = -10.2$
12	19	$(12-12.4) = -0.4$	$-0.4^2 = 0.2$	$(19-13.6) = 5.4$	$5.4^2 = 29.2$	$(-0.4)(5.4) = -2.2$
16	16	$(16-12.4) = 3.6$	$3.6^2 = 13$	$(16-13.6) = 2.4$	$2.4^2 = 5.8$	$(3.6)(2.4) = 8.6$
15	12	$(15-12.4) = 2.6$	$2.6^2 = 6.8$	$(12-13.6) = -1.6$	$-1.6^2 = 2.6$	$(2.6)(-1.6) = -4.2$
12	12	$(12-12.4) = -0.4$	$-0.4^2 = 0.2$	$(12-13.6) = -1.6$	$-1.6^2 = 2.6$	$(-0.4)(-1.6) = 0.6$
14	16	$(14-12.4) = 1.6$	$1.6^2 = 2.6$	$(16-13.6) = 2.4$	$2.4^2 = 5.8$	$(1.6)(2.4) = 3.8$
$\bar{x}$ 12.4	$\bar{y}$ 13.6		<b>SS<sub>x</sub>=63.7</b>		<b>SS<sub>y</sub>=79.7</b>	<b>SP<sub>xy</sub>=19.3</b>

Then, according to the formula for  $r$  we have:

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{19.3}{\sqrt{63.7 \times 79.7}} = \frac{19.3}{71.3} = 0.271$$

Obviously, this  $r=0.271$  is not the same as the SPSS-produced  $r=0.413$  we had from Section 7.2.3; in fact, it

would be very surprising if they were the same, considering the former is based on  $N=7$  while the latter is based on  $N=1,687$ . The exact value of  $r$  in the above calculation ( $r=0.271$ ) doesn't matter, and doesn't serve any purpose and shouldn't be interpreted as it exists only as the end result of our demonstration.

Fancy trying it out on your own?

#### *Do It! 10.1 Calculating Pearson's $r$*

Here are 7 more cases from the same *GSS 2018* dataset. Fill out the table fully and produce  $r$ .

$x$	$y$	$(x - \bar{x})$	$(x - \bar{x})^2$	$(y - \bar{y})$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
12	12					
12	14					
13	13					
13	16					
14	20					
20	16					
21	18					
$\bar{x}$	$\bar{y}$					
=	=		$SS_x =$		$SS_y =$	$SP_{xy} =$

Even if we dismiss the value of the  $N=7$  coefficients and go back to  $r=0.413$  based on  $N=1,687$ , we still want to know if this correlation *as observed in the sample* is statistically significant (i.e., generalizable to the population). Thus, we need to test  $r$ , and we do that through a  $t$ -test.

**The  $t$ -test for Pearson's  $r$**  is given by the following formula:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

with  $df=N-2$ .

*Example 10.1(B) Testing the Education and Parental Education Correlation, GSS 2018*

As usual, it helps to know what we are testing exactly:

- $H_0$ : There is no correlation between parental and offspring education;  $\rho=0$ .
- $H_a$ : There is a correlation between parental and offspring education;  $\rho \neq 0$ .

Then, for  $N=1,687$  and  $r=0.413$ , we have:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{0.413\sqrt{1687-2}}{\sqrt{1-0.413^2}} = \frac{0.413(41.1)}{0.911} = 18.633$$

**With  $t=18.633$ ,  $df=1,685$ , and  $p=0.00001$  (i.e.,  $p=0.00001<0.5$ ), we can reject the null hypothesis that parental and offspring education are not correlated. At this time, we have enough evidence to conclude that there is a moderately weak ( $r=0.413$ ), statistically significant**

**correlation between parental education and offspring education in the US population<sup>6</sup>**

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{0.271\sqrt{7-2}}{\sqrt{1-0.271^2}} = \frac{0.271(5)}{0.963} = 1.407$$

In this case, we could interpret the results like this: "With  $t=1.407$ ,  $df=5$ , and  $p=0.218$  (i.e.,  $p=0.218>0.5$ ), we cannot reject the null hypothesis that parental and offspring education are not correlated. At this time, we do not have enough evidence to conclude that there is a statistically significant correlation between parental education and offspring education in the US population." However, we cannot trust this "inference" as it is only based on  $N=7$ .

With this, we can have established (with 99% certainty) that parental education and offspring education are correlated. Considering that parents tend to have their schooling done before their children have theirs, on average, it is also reasonable to assume that parental education affects offspring education (and not vice versa)<sup>7</sup>.

6. Purely for demonstration purposes, we could also calculate the  $t$  for the 7 respondents whose responses we used to calculate  $r=0.271$ :
7. In terms of establishing causality, we are limited by the bivariate case we have: it is entirely possible (and expected) that other things affect offspring education too, not just their parents' education. As well, it is possible than something else (for example, wealth, income, socioecoomic class, etc.)

Wouldn't then be good to know *exactly how much* effect parental education has on offspring education? That is, wouldn't you like to know that if a father had one more year of schooling compared to another father, how much more schooling the child of the former would be expected to have compared to the child of the latter? One type of regression — called *linear regression* — can tell us just that.

might be affecting both parental and offspring education, rendering the effect of parental education on offspring education spurious. These type of considerations are exactly the purpose of multivariate analysis, but since we are dealing with bivariate analysis here, we have to leave these considerations aside. I bring them up here to remind you not to forget them in the discussion that follows, which will focus on the two variables at hand.

---

## 10.2 Basics of Linear Regression

You may find it surprising but you already *have* an idea about linear regression from Section 7.2.3. Again, when describing and examining the association between two continuous variables, we can use the correlation coefficient  $r$  and a scatterplot plotting the observations in a coordinate system. To visualize the linear relationship between the variables, we could also add a *line of best fit* to the scatterplot. The line of best fit is actually also called a *regression line*; and regression itself is based upon the concepts of correlation and variance, with which you are already familiar.

You might be asking yourselves at this point what regression adds to the analysis of two continuous variables, or in other words, why do we even need it — don't we already have Pearson's  $r$  for that? As you will see in the examples below, **linear regression allows us to precisely calculate and predict a change in the *dependent* variable that is due to the *independent* variable.**

What we say in this case is that **the independent variable *explains* a percentage of the variance of the dependent variable.** Think about it this way: the dependent variable varies due to arguably many causes (i.e., independent variables), which affect it to a different extent and which each explain some part of its total variance. **Through linear regression, we are able to quantify to what extent an independent variable explains the variability of the dependent variable, i.e.,**

**to what extent it affects it**<sup>1</sup>. To take the example about parental and offspring education from the previous section, doing a regression analysis on these two variables would allow us to predict how much more education a respondent is expected to have for one more year of schooling for the parent<sup>2</sup> (father, in our case), and what percent of respondent's schooling is explained by the years of education of the parent.

How does linear regression do all that? To put it simply, through the regression line (of best fit), or more precisely, through the way the regression line is created.

**The linear function.** How do you draw a line? The simplest method requires exactly two pieces of information: a starting point of the line, and an indicator of slope (so that you know whether the line is straight, sloping upward, or sloping downward). This is captured in the following formula:

$$y = \alpha + \beta x$$

where  $\alpha$  is the line's starting point and  $\beta$  is the slope of the line. The two variables,  $x$  and  $y$ , are the independent and the dependent variable, respectively: we know this because

1. Multivariate regression thus allows for direct comparisons of the size of the independent variables' effects. In the bivariate case, we only focus on the effect of *one* independent variable, without considering and accounting for others -- which is not something you should do in a real-life social science research, especially in terms of causal analysis. Again, the bivariate case serves only as an illustration/introduction to the expansive topic of regression in general.
2. Or, to put it differently, if one father has one more year of schooling than another father, how much more schooling the offspring of the first would be expected to have in comparison to the offspring of the second.



the formula establishes  $y$  as a *function* of  $x$  (i.e., if we know  $\alpha$  and  $\beta$ , we can calculate  $y$  for any value of  $x$ ).

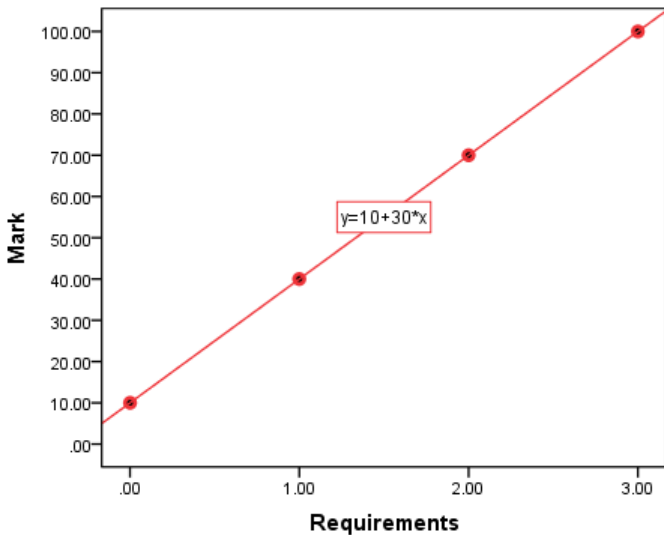
Let's take a brief example.

*Example 10.2 Class Assignment Mark*

Imagine you are given a written, take-home assignment in some class. Your professor has stipulated that there are three part of the assignment, each worth 30 points, and that you would receive 10 points just for turning in your work.

In this case, your assignment mark is entirely a function of your submitted work. You will be getting 10 points to start with, then 30 points for fulfilling each of the three requirements. The class grades on the submitted assignments could thus be 10 points (0 completed requirements), 40 points (1 completed requirement), 70 points (2 completed requirements), and 100 points (3 completed requirements). Figure 10.1 plots this.

*Figure 10.1 Assignment Mark as a Function of Completed Requirements*



As you can see, the relationship between the two variables, *assignment requirements completed* and *assignment mark*, is simply

$$y = 10 + 30x$$

as helpfully shown in the graph itself. This is a summary form of having to write out all the observations:

- when  $x=0$ ,  
 $y = 10 + 30x = 10 + 30 \times 0 = 10 + 0 = 10$

- when  $x=1$ ,  

$$y = 10 + 30x = 10 + 30 \times 1 = 10 + 30 = 40$$
- when  $x=2$ ,  

$$y = 10 + 30x = 10 + 30 \times 2 = 10 + 60 = 70$$
- when  $x=3$ ,  

$$y = 10 + 30x = 10 + 30 \times 3 = 10 + 90 = 100$$

In the example above  $\alpha=10$  and  $\beta=30$ : the line starts at  $x=0$  and  $y=10$ , and for each additional unit of  $x$  (i.e., each additional requirement completed),  $y$  increases by 30 points.

In fact, these are the exact definitions of  $\alpha$  and  $\beta$ . That is,  **$\alpha$  is the value of  $y$  when  $x=0$ , also called *Y-intercept*** (as it shows where the regression line crosses the vertical  $Y$ -axis), and  **$\beta$  is the *slope*, also called the *regression coefficient*, i.e., the amount of change in the dependent variable  $y$  expected for every unit change in the independent variable  $x$  (or simply, the size of the effect of  $x$  on  $y$ ).**

Let's now take a look at the regression model in detail.

3. Of course, to draw a line you only really need *two* points. Thus if you only take  $x=0/y=10$  and  $x=3/y=100$  and connect these points with a line, the line will also pass through  $x=1/y=40$  and  $x=2/y=70$ . This is a useful property if you need to draw a line by hand.



---

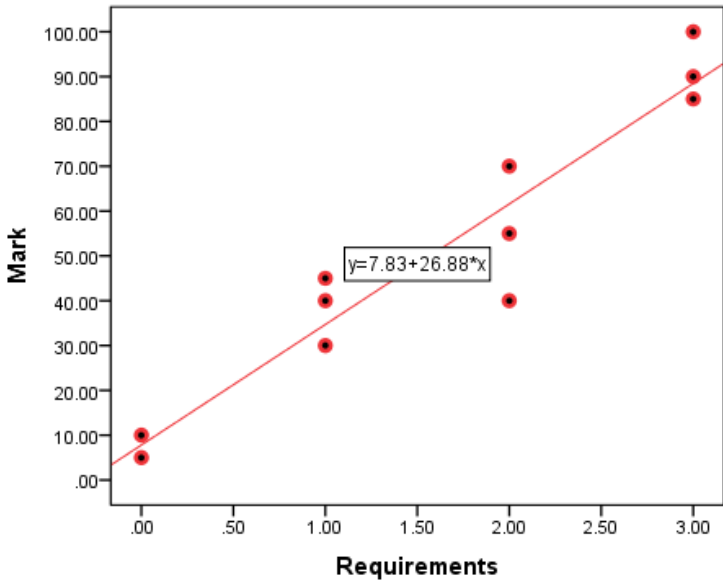
## 10.2.1 The Linear Regression Model and the Line of Best Fit

You might have noticed that there was no uncertainty of any kind in the Example 10.2 about the assignment requirements and mark in the previous section. The line in that case represented a *deterministic* relationship —  $x$  fully determined  $y$  (i.e.,  $x$  fully explained the variability of  $y$ ) — hence all the observations were on the line itself.

As such, this was not a typical situation and this was not a typical *regression* line. In reality, in statistical inference we deal with *probabilistic* associations, where the regression line does *not* capture all observations in itself but their *general* (on average) *trend*. That is, in a usual regression model situation, some observations will be above the line and some below it; thus some observations would be *underestimated* and others would be *overestimated* because **the line serves as a prediction** (an expectation, a summary, a trend) of the association. And as we know by now, predictions/estimations always contain a level of uncertainty.

Specifically, we cannot expect that a single independent variable  $x$  will explain away *all* variability in a dependent variable  $y$ ; there will always be some unexplained (by the regression model) variability left. Figure 10.2 illustrates.

*Figure 10.2 Assignment Mark as a Function of Completed Requirements (With Variance)*



In Figure 10.2 I have added seven more observations to the case we had in Figure 10.1 in the previous section, this time allowing for additional variability in the assignment marks: no longer is it enough to know the number of requirements completed to predict the assignment grade. (Imagine that the professor has started evaluating the completed requirements substantively, not just counting them: in this case while the number of requirements is still essential for the grade, *something else*<sup>1</sup> also affects the final assignment mark.)

An actual **regression model accommodates the uncertainty inherent in estimation through two interrelated concepts, *error of prediction* (a.k.a. statistical error) and *residuals*.**

1. This *something else* is an 'unobserved variable', or a variable not included in the model (even though we could speculate about it). This type of unobserved variable/s is the source for the unexplained variance in  $y$ .

**The *error of prediction* reflects the difference between the observations and the predicted values we would have if we had data about the population.** That is, if we imagined a line of best fit of the population<sup>2</sup>,  $\alpha + \beta x$ , the difference between our observations and that line would be:

$$y - (\alpha + \beta x) = \epsilon$$

= *error of prediction*<sup>3</sup>

That is, we need to include the error term in the regression model:

$$y = \alpha + \beta x + \epsilon$$

Considering that we pretty much never have information about the population, however, we can restate **the *sample regression model* like this:**

$$y = a + bx + e$$

**where  $a$  is the estimated  $\alpha$ ,  $b$  is the estimated  $\beta$ , and  $e$  is the estimated  $\epsilon$ , with all estimations based on sample data. Note that  $e$  here is called the *residual*, and it is not only the estimation of the unobservable error of prediction, but also simply the difference between an observation and its predicted value:**

$$y - (a + bx) = e$$

= *residual*

2. This line of course does not exist, it is a heuristic device.

3. This is the small-case Greek letter  $e$ ,  $\epsilon$  [EHpsilon].

Since  $a+bx$  is the regression line, or the prediction, it also stands for the predicted (estimated values), which we can, as usual, denote  $\hat{y}$ . Then, since

$$\hat{y} = a + bx$$

,

we also have

$$y - \hat{y} = e$$

or, again, that **the residuals are the difference between the observations and their predicted values.**

With this, we come at a full circle and the reason for all the notation and protracted explanations above (and here you thought I was subjecting you to all these equations without a purpose): in a graph, **the residuals are simply the distance between the observations and the regression line.** (In Figure 10.2 this is the empty space — the shortest distance — between an observation and the regression line.)

A comprehensive treatment of the residuals (through a full-blown analysis of variance) is beyond the scope of this book but they do help us understand the nature of the regression line and of the logic of regression in general. You see, **the regression line is called a line of best fit precisely because it *minimizes the residuals*** — it is created in such a way as to minimize the residuals (and therefore the error of prediction) and fit the data/observations as best as possible. Visually, this will mean that the line is drawn to pass *as close as possible* to all the observations.



In fact, **linear regression is also called *OLS regression*, which stands for *ordinary least squares***. The *least squares* concept comes from the fact that to minimize the distances of the observations to the prediction line, we need to first square them before adding them together<sup>4</sup> — just like we needed to do that in the calculation of the variance and the sum of squares (or the distances would cancel each other out)<sup>5</sup>.

But how do we ensure that the regression line minimizes the residuals? The next section explains.

4. I.e.,  $\sum (y - \hat{y})^2$ .

5. The *ordinary* part is there to differentiate between another regression version called *generalized least squares regression*, or *GLS* regression (not discussed here).



---

## 10.2.2 Elements of the Linear Regression Model

The secret to minimizing the residuals — and to ensuring the regression line is indeed *the best fitting* (to the data) line — lies in the way the elements of the line are calculated. The regression/prediction line is, after all, created through  $a$  and  $b$ , as I explained in Section 10.2:

$$\hat{y} = a + bx =$$

= *predicted values*

We can calculate  $a$  and  $b$  such that they minimize the residuals through the following formulas:

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{SP}{SS_x} =$$

= *slope, or regression coefficient*

$$a = \bar{y} - b\bar{x} =$$

= *Y-intercept, or constant*

where  $SP$  is, again, the sum of products,  $SS_x$  is the sum of squares for  $x$ , and  $\bar{x}$  and  $\bar{y}$  are the variable means of  $x$  and  $y$ , respectively.

As with the correlation coefficient  $r$ , once again, everything revolves around variances (and means)<sup>1</sup>.

1. So much so that the correlation coefficient  $r$  and the regression coefficient  $b$

An example will serve best to illustrate all this.

*Example 10.3 Assignment Requirements and Marks*

Here I continue with the fictitious data on which Figure 10.2 is based. In a “sample” of  $N=11$ , I have data about the “respondents” completed assignment requirements ( $x$ ) and their assignment marks ( $y$ ). In Table 10.3, I calculate the necessary means, sums of squares, and sum of products.

*Table 10.3 Assignment Requirements and Marks:  
Calculating  $a$  and  $b$*

are related:  $b = r \frac{s_y}{s_x}$  where  $s_y$  and  $s_x$  are, of course, the standard deviations of  $y$  and  $x$ , respectively.



This makes our **best-fitting/regression line** this:

$$\hat{y} = a + bx = 7.83 + 26.88x$$

... which is exactly what SPSS had already told us, if you care to go back to Figure 10.2 in the previous section and check.

You may or might not be impressed by this, but you certainly need to know how to interpret it. In this case the regression tells us that **a student who doesn't complete even one requirement of their assignment is expected to receive 7.83 points** (that's the constant, or Y-intercept); **further, for every requirement completed, their mark would increase by 26.88 points** (that's the regression coefficient). **That is, the effect of one completed requirement on the assignment mark is 26.88 points.**

We can also calculate the actual predicted values (which form the regression line itself):

- for  $x=0$ ,  

$$\hat{y} = 7.83 + 26.88 \times 0 = 7.83 + 0 = 7.83$$
 ;
- for  $x=1$ ,  

$$\hat{y} = 7.83 + 26.88 \times 1 = 7.83 + 26.88 = 34.71$$
 ;
- for  $x=2$ ,  

$$\hat{y} = 7.83 + 26.88 \times 2 = 7.83 + 53.76 = 61.59$$
 ;

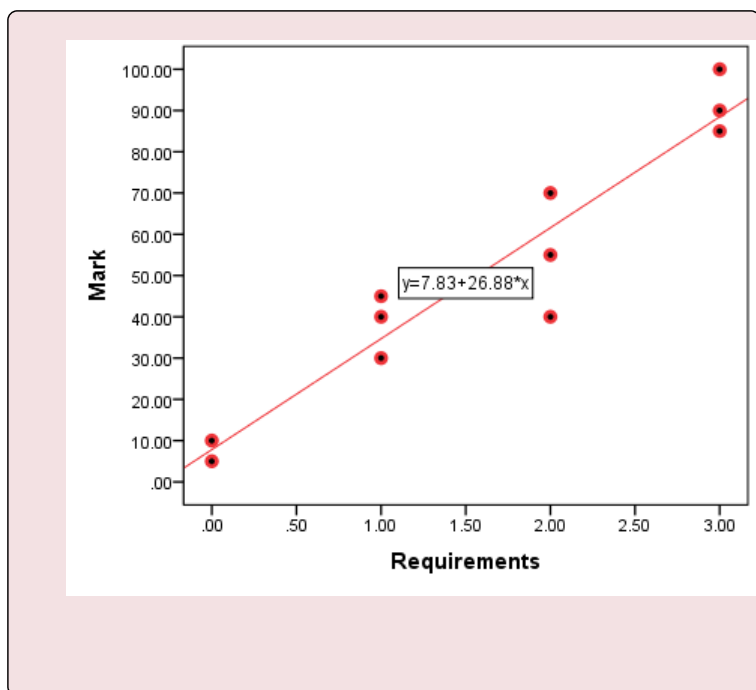
- for  $x=3$ ,  

$$\hat{y} = 7.83 + 26.88 \times 3 = 7.83 + 80.64 = 88.47$$

As you can see, these are different values than the ones we had in the deterministic version with which we started in Section 10.1 (i.e., 0 requirements = 10 points, 1 requirement = 40 points, 2 requirements = 70 points, 3 requirements = 100 points). The difference between the certainty of the deterministic version and the *uncertainty* of the current probabilistic version is the unexplained (by number of requirements) variance<sup>2</sup>. How much variance we *have* explained we will see in the next section. Before that, here is Figure 10.2 again so that you can pinpoint the predicted values for yourselves. (Hint: they're on the line.)

*Figure 10.2 Assignment Requirements and Mark (Redux)*

2. That is, in the deterministic version, we could say that  $y = \hat{y}$  (*reality = prediction*), or rather, that there is no prediction at all -- we know what the true relationship between the variables is as the assignment mark depends entirely on the number of fulfilled requirements. In the actual/probabilistic version,  $y = \hat{y} + e$  (*reality = prediction plus residual/error*), where the residual is what is left unexplained, or simply the difference between reality and prediction.



**Testing the regression coefficient for statistical significance.** Of course, as with any statistics obtained through a sample, we have to be able to check whether the regression coefficient is generalizable to the population, i.e., whether it is statistically significant. In other words, we have to examine the evidence whether the identified effect of the independent variable on the dependent variable exists in the population or whether it is a result of random sampling.

The significance test for  $b$  is your familiar  $t$ -test, given by the following formula<sup>3</sup>:

3. The population version is  $z = \frac{b}{\sigma_b}$ . Since we generally do not know  $\sigma_b$ , we



$$t = \frac{b}{s_b}$$

where  $s_b$  is  $b$ 's standard error.<sup>4</sup>

$$s_b = \sqrt{\frac{\frac{\sum(y - \hat{y})^2}{(N-2)}}{\sqrt{\sum(x - \bar{x})^2}}}$$

This can be simplified to be more user-friendly but then I will need to introduce additional concepts (like the *mean squared error* and the *standard error of the estimate*) which are not necessary for you at this stage and are therefore beyond the scope of this book. You will be happy to know that the hand calculation of  $s_b$  also falls in that category.:

The degrees of freedom for  $t_b$  are  $N-2$  in the bivariate case. We can see what we can do with the test in hypothesis testing, next,

substitute it with its estimate, the sample-based  $s_b$ . This of course also means we move to the  $t$ -distribution.

4. The standard error of  $b$  is calculated by this, admittedly scary-looking, formula:



---

## 10.2.3 Hypothesis Testing and Confidence Intervals for the Regression Coefficient

To test the regression coefficient  $b$  for significance we have the following hypotheses:

- $H_0$ : The independent variable  $x$  has no effect on the dependent variable  $y$  (i.e., the variables are not associated);  $\beta=0$ .
- $H_a$ : The independent variable  $x$  has an effect on the dependent variable  $y$  (i.e., the variables are associated);  $\beta \neq 0$ <sup>1</sup>.

**After calculating  $t_b$  with  $df=N-2$  and finding its associated  $p$ -value, we then compare the  $p$ -value to the pre-selected significance level  $\alpha$ . As usual, when  $p \leq \alpha$ , we reject the null hypothesis, and have enough evidence to deem the regression coefficient  $b$  statistically significant. If, on the contrary,  $p > \alpha$ , we fail to reject the null hypothesis and therefore conclude that at present there is no evidence to suggest an effect of  $x$  on  $y$ .**

Again, similarly to other statistics, we can **calculate**

1. Note that I am using causal language here with the assumption that the conditions for causality are met. There is a separate investigation. In and of itself, finding a significant effect of  $x$  on  $y$  does not itself establish that changes in  $x$  *cause* changes in  $y$ .

**confidence intervals for  $b$ , so that we can report the size of the effect with a specific level of certainty.** For example, the 95% confidence interval for the regression coefficient  $b$  is:

- 95% CI:  $b \pm 1.96 \times s_b$

To illustrate, let's revisit our example about the effect of parental education on their offspring education. (Don't worry, with  $N=1,686$  I will not offer you a calculation by hand: SPSS is there for us.)

*Example 10.4 Effect of Parental Years of Schooling on Respondent's Years of Schooling (GSS 2018)*

We already examined the association between parental and offspring education through the correlation coefficient  $r$  and found it to be moderately weak at 0.413, and statistically significant at  $\alpha=0.01$ . Can we do better, however, and estimate the effect of each additional year of parental schooling on the schooling of the respondents?

Again, we use data from the U.S. GSS 2018 (NORC, 2019). Our sample is  $N=1,686$ , and **our hypotheses are:**

- $H_0$ : Father's education has no effect on respondent's education;  $\beta=0$ .

- $H_a$ : Father's education has an effect on respondent's education;  $\beta \neq 0$ .

**The regression model is:**

years of schooling =  $y = a + bx + e = a + b(\text{years of parental schooling}) + e$

**Our predicted values are:**

predicted years of schooling =  $\hat{y} = a + bx = a + b(\text{years of parental schooling})$

Figure 10.3 plots the association and Table 10.4 show the relevant SPSS output.

Figure 10.3 *Linear Regression of Respondent's Years of Schooling and Father's Years of Schooling*

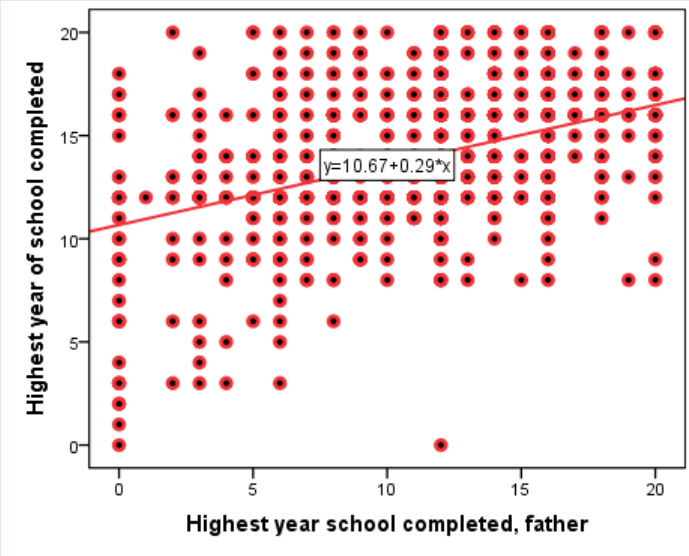


Table 10.4 Linear Regression of Respondent’s Years of Schooling and Father’s Years of Schooling

Coefficients <sup>a</sup>								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	10.672	.196		54.334	.000	10.287	11.058
	Highest year school completed, father	.290	.016	.413	18.607	.000	.260	.321

a. Dependent Variable: Highest year of school completed

That is, SPSS has calculated the constant (or Y-intercept)  $a$  and the regression coefficient  $b$  in such a way as to minimize the residuals:

- $a = 10.67$
- $b = 0.29$

Then, the predicted values (i.e., the regression line on Figure 10.3 above) are:

$$\text{predicted years of schooling} = \hat{y} = a + bx = 10.67 + 0.29(\text{years of parental schooling})$$

We also know that the standard error of  $b$  is  $s_b = 0.016$ , so

$$t = \frac{b}{s_b} = \frac{0.29}{0.016} = 18.607$$

2

**Thus, with  $t=18.607$ ,  $df=1,684$ , and  $p<\alpha=0.001$ , we can reject the null hypothesis. Our current evidence supports our hypothesis that father's education affects their offspring's education, on average. The effect is 0.29 years (or about 3.5 months) for every additional year of father's schooling, and it is statistically significant.**

As well, we can interpret **the confidence interval**:

- 95% CI:

$$b \pm 1.96s_b = 0.29 \pm 1.96(0.016) = 0.29 \pm 0.031 = (0.26; 0.32)$$

**Or, father's education's effect on offspring's education would be between 0.26 additional years and 0.32**

2. If you actually divide 0.29 by 0.016, you will end up with 18.125. The difference from 18.607 is due to rounding (as the standard error of  $b$  is rounded up to 0.016 from 0.01558...).

**additional years for every year of father's schooling with 95% certainty; in other words, the effect would be  $0.29 \pm 0.031$ , 19 out of 20 times.**

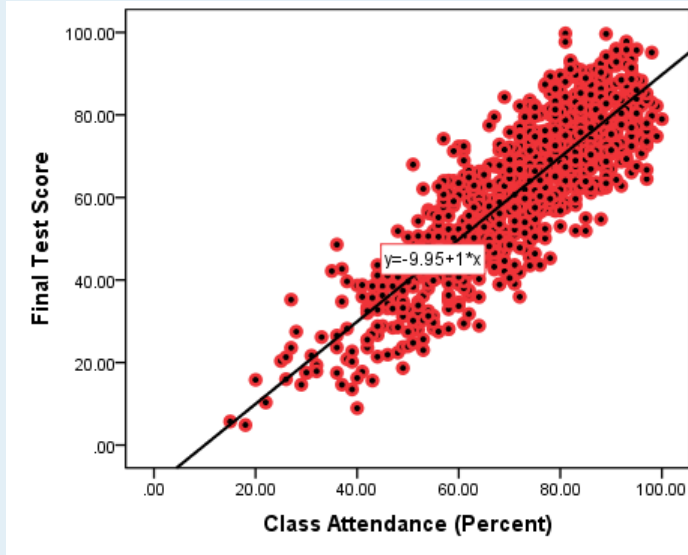
That's a lot more information than simply stating that the variables are associated based on the correlation coefficient!

Now let's make sure you understand how regression works and where the regression coefficients and line come from by interpreting regression output.

*Do It! 10.2 Class Attendance and Final Test Scores (Simulated Data)*

We are revisiting the simulated data on student class attendance (measured in percent of classes attended) and their final class scores.  $N=987$ . Start by stating your hypotheses, then, using the SPSS's output presented in Figure 10.4 and Table 10.5 below, write a paragraph interpreting what you have found, discussing the evidence presented regarding your hypotheses and your decision about them, etc. Include as much information as possible, and do not forget to justify your use of linear regression in this case.



*Figure 10.4 Class Attendance and Final Test Scores**Table 10.5 Class Attendance and Final Test Scores*

Coefficients <sup>a</sup>								
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		
	B	Std. Error	Beta			Lower Bound	Upper Bound	
1	(Constant)	-9.953	1.460		-6.816	.000	-12.818	-7.087
	AttendPer	.997	.020	.849	50.403	.000	.958	1.036

a. Dependent Variable: TestScore

a. Dependent Variable: TestScore

Finally, these are the steps through which the regression output is obtained in SPSS.

*SPSS Tip 10.1 Linear Regression*

- From the *Main Menu*, select *Analyze*, then from the pull-down menu, select *Regression* and click on *Linear*;
- Select your dependent variable from the list of variables on the left and, using the appropriate arrow, move it to the *Dependent* open space on the right;
- Select your independent variable from the list of variables on the left and, using the appropriate arrow, move it to the *Block 1 of 1* empty space on the right.
- You can click *OK* or, if you need a confidence interval for  $b$ , click on *Statistics*, and check off *Confidence intervals* in the new window (here you can also specify the confidence *Level* of the CI); click *Continue*;
- Once back in the original window, click *OK*.
- After the *OK*, SPSS will provide the output in the *Output* window. The relevant information we have discussed so far can be found in the last table called *Coefficients*.

SPSS provides several tables as the standard regression output. Beyond the *Coefficients* one, there are three other short tables: a *Variables Entered/Removed* (which lists the independent variable/s in the model and the dependent variable as a footnote), an *ANOVA* table (which presents

analysis of variance information that, as mentioned before, is outside the scope of this book), and a *Model Summary* table. We will take a brief look at that last table in the next section.



---

# 10.2.4 R-squared

In the previous section we established that the correlation coefficient  $r$  and the regression coefficient  $b$  are related:

$$b = r \frac{s_y}{s_x}$$

And how could they not be: if a slope exists, a correlation exists. As such, the standard regression output provided by SPSS includes a *Model Summary* table that lists the Pearson’s  $r$ . Table 10.6 below is the *Model Summary* table of the simulated-data class attendance/final class scores regression.

Table 10.6  $R$  and  $R^2$  for Class Attendance and Final Class Scores

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.849 <sup>a</sup>	.721	.720	9.55052

a. Predictors: (Constant), AttendPer

Pearson’s  $r$  (listed as  $R$ ) in this table is, of course, exactly the same as what the SPSS *Correlate* procedure provides. Squaring that number, however, provides us with a new and useful piece of information, sometimes called **the coefficient of determination, but more often simply referred to as  $R^2$** .

$$r \times r = R^2$$

**$R^2$  provides a measure of the proportion of the variability in the dependent variable explained by the independent variable**[footnote]Or, independent variables, in the case of multivariate regression.[/footnote] **in the model.**

$$R^2 = \frac{\text{explained variation of } y}{\text{total variation of } y}$$

Recall that regression's logic is based on minimizing residuals/errors and about explaining the variation of the dependent variable through information about the independent variable. In a deterministic case, the dependent variable will depend entirely on the independent one, and then we would have a correlation of 1 and  $R^2=1$ . However, with uncertainty and estimation, this is not the case — some variability of the dependent variable remains unexplained by the regression model (i.e., the independent variable).

Thus, one way to look at  $R^2$  is as an indication of *goodness of fit*: how close the observations are fitted around the regression line (i.e., how little variability is left unexplained). The larger  $R^2$  then, the better — as a large  $R^2$  would mean the model (the independent variable/s) explains a large proportion of the variability in the dependent variable.

As you can see in Table 10.6 above, the  $R^2$  of the class attendance/final test scores is:

$$r \times r = 0.849^2 = 0.721 = R^2$$

Or, class attendance explains 72.1 percent of the variability in final test scores, which is a lot, and quite good regression fit<sup>1</sup>.

Compare this to the *Model Summary* table of respondent's and father's years of schooling in Table 10.7 below.

*Table 10.7 R and R<sup>2</sup> for Respondent's and Father's Years of Schooling*

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.413 <sup>a</sup>	.170	.170	2.658

a. Predictors: (Constant), Highest year school completed, father

Unlike the very strong correlation of  $r=0.849$ , the moderately weak correlation coefficient  $r=0.413$  is already an indication of not that great a fit. Thus, the  $R^2$  of offspring and parental education is:

$$r \times r = 0.413^2 = 0.170 = R^2$$

That is, fathers' years of schooling explain only 17 percent of the variation of respondents' years of schooling. The biggest 'chunk' of the variation in schooling is left

1. Of course, this also means that  $(100-72.1=)$  27.9 percent of the variation in test scores is left unexplained by class attendance, i.e., is due to something else beyond class attendance.

unexplained, i.e., there are other factors influencing how much education one is expected to have, on average. Regardless, we should not dismiss parental education outright — it still has a statistically significant effect on offspring education (albeit not very strong).

. . . Or does it? Recall our discussion on causality. The fact that two variables are *statistically* associated does not necessarily mean that one *causes* the other to change (or, that it explains the other's variability). Working with only two variables prevents us from accounting for alternative explanations — i.e., of taking into account other factors, other variables, other effects. Luckily, regression has our backs. I leave you with how that happens in the next — and *final!* — section of this textbook.



---

## 10.3 What Lies Ahead: Multiple Regression

Don't worry, this is but a brief farewell. Do me a last favour and imagine we had more ideas about why students end up with different final test scores, or why people end up with different number of years of education. In other words, what else could possibly explain some of the variability in the dependent variables we investigated in the previous sections?

In the case of test scores, perhaps hours of independent study? Doing end-of-chapter exercises? How many classes in total the student is taking that semester? Does the student work for pay? Have they recently experienced problems related to their personal life? Do they have dependents of whom they have to take care at home? Is English their native language? Are they international students? What is their area of study? . . . And so on, and so on; I'm certain you can add more on your own.

In the case of years of schooling, perhaps the family's socioeconomic status? Wealth? Gender? Race/ethnicity? Citizenship status? Attitudes toward education? The presence of appropriate role models? Being passionate about a specific field of study? Go on, add your ideas to the list.

If there are so many factors that can affect a (dependent)

variable, how do we examine their individual effects? Bivariately, one by one? While this is a good first step (to establish *something* is going on), obviously that cannot be the end of our analysis. We *have* to be able to account for all of them *at the same time*, to compare their effects, and to create more complicated models which *together* to explain more variability in the dependent variable.

Multiple regression allows us to do just that. Instead of *one* independent variable  $x$ , we can consider *many* independent variables at the same time. Then, the effect of each single variable is provided *net* of the effects of the other variables (or we say that we *control for* the other variables), so that we can simultaneously take care of alternative explanations. In this way, a variable's effect on  $y$  may be decreased or increased (from what it used to be in the bivariate case), and its statistical significance may disappear (or even appear, in some cases). In any case, this effect would likely be 'truer' than the one obtained bivariately (though this of course depends on the choice of variable controls).

And this is where you will be going, if you choose to continue on the path of statistics enlightenment. If I said there is a lot more to learn it would be a gross understatement — but, given what statistics (*proper* use of statistics!) enables you to do in social research, it is absolutely and totally worth it.

If you choose not to continue on, then use what statistical knowledge you already have, and use it responsibly (great power, and all that). Either way, here you are, in the last section — you survived! (Possibly even with your sanity mostly intact.) Go celebrate!

With this, I bid you adieu.



---

## References



---

This is where you can add appendices or other back matter.