(Adobe Stock)

chapter 2

Data

If statistics is a living being, then data is its lifeblood. Just like a person has to feed their own body properly for it to function, we must input the appropriate data to the correct type of statistical analysis in order to make statistics work the way it was intended. This chapter will define what data are, review the different types of data that exist, and explain why these things are important to know.

Data is just a fancy word for "information." Let us imagine that I want to gather a lot of information about the students enrolled in my statistics class. I could ask them all to answer these questions:

(a) What year in school are you?

(b) What is your age?

(c) Have you ever taken a statistics course before?

(d) Do you prefer chocolate or vanilla ice cream?

When I gather their responses all together, I will have plenty of information about the class. I can do some very interesting things with this information, which we will get to in later chapters.First, consider some of the different characteristics of the information that I have.

Notice that the first question will probably have four possible answers. I will probably get some who answer that they are in their freshman year, some will be sophomores, some will be juniors, and some will be seniors. It is less likely, but possible, that I will have some high school students who are earning college credits. It is also less likely, but possible, that I will have some graduate students or college graduates who are taking the course to fulfill some requirement for a job or a program in which they are enrolled. Therefore, that first question has some limitations in the ways it can be answered.

In contrast, the second question is asking something with more possibilities. Age (in Earth years) starts at zero and then goes on up past 100 in some cases. There are more answer options to this question than are possible with my first question about what year in school a person is.

In the third and fourth questions I have asked something in such a way that I will get only two answers. Some students will respond to the statistics question with "yes," and some will respond with "no." Similarly, some will prefer chocolate over vanilla, and vice versa.

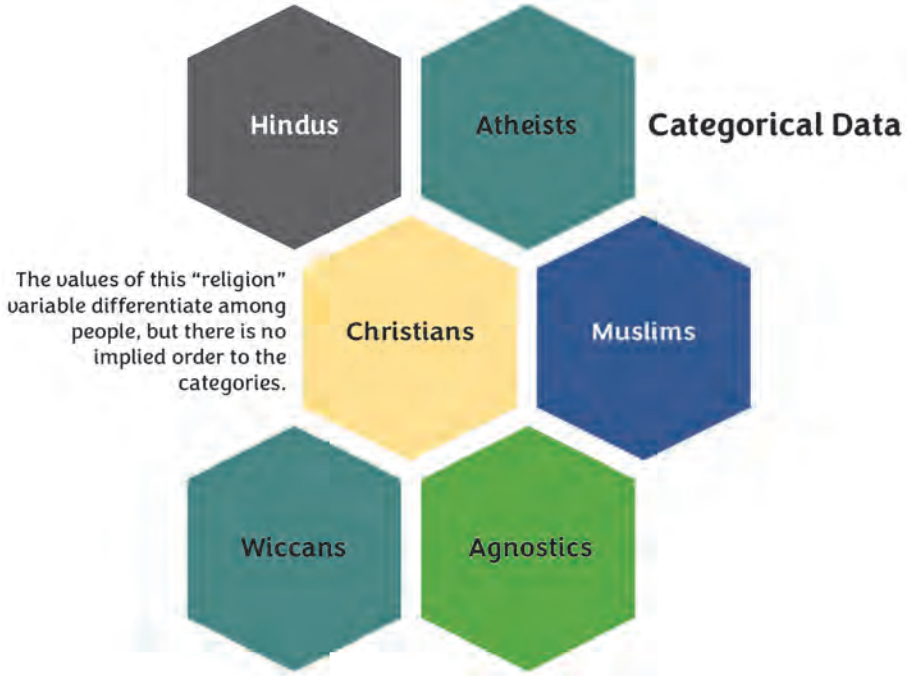
These different properties of information have names and can be broadly characterized. That is what we will go over next. Information (or data) like this can be described broadly as categorical or continuous.Whether something has categorical or continuous properties will influence the types of statistical tests we can do with it, so this is important to understand before we start moving forward.

2.1 Categorical Data

Categorical data are bits of information that have no inherent order to them. They are also sometimes called "nominal" or "discrete" data. This is information that is used to differentiate between whatever it is describing, but not anything more than that. For example, the fourth question above, "Do you prefer chocolate or vanilla ice cream?" is a question that would get categorical answers. It would help me to differentiate those who prefer chocolate ice cream from those who prefer vanilla ice cream. There is no inherent order to these two answer options, because there is no reason to treat chocolate preference as superior to vanilla preference (despite the researcher's own preferences), or vice versa. Specifically, these sorts of data are **dichotomous**, which means that they can take on only two values. Dichotomous (or "binary") variables will produce categorical data, because their values can be one of only two possibilities. Of course, some categorical data can take on dozens of categories (e.g., the country where a person currently resides is a variable that can take on hundreds of values, but it is still categorical).

There are lots of naturally occurring categorical data. We could, for example, be interested in surveying everybody's religious affiliation. We would get some Atheists, Buddhists, Christians, Hindus, Jews, Muslims, and so on. Given only these affiliations, there is no need to put them in any kind of order, and there is no intuitive order in which they should belong, because these are just ways of differentiating characteristics about the people who answered the question.

Figure 2-1 Categorical Data Describe Only Differences



Other types of categorical data:

- Political party
- Zip code
- City of one's birth
- Smoker or non-smoker
- Male or female
- The color of the tassel worn at graduation
- And many others

For each of these, there is no inherent order to the categories within them. Democrats and Republicans are different parties. Zip codes specify different postal areas, but even though we could put them in numerical order, that does not actually mean that the zip code 80909 is better, bigger, and so on, than is zip code 60525. (Note that we could possibly start to rank some of these based on other information about them, such as which political party was founded first, but that is different information from using the label just to distinguish among the political orientations.)

Also notice that some of these data have more options than others. For example, most people probably see themself as either a smoker or a non-smoker. There are just two options to this one. However, the city of one's birth could take on thousands of values. The possible values of a variable are often called the **levels** of a variable. For example, if I asked a group of people what their sex was at birth, there would be two levels: male or female (excluding the rare cases with genetic disorders leading to ambiguity in this regard). If I ask their city of birth, I have one variable with thousands of levels. There are other types of variables that have additional properties which we will start to explore next.

2.2 Continuous Data

Data that are continuous have additional properties that allow us to do more than just differentiate whatever they label. There are many characteristics that are important to account for in continuous data, which we will now explore.

2.2a Ordinal Data

The first characteristic that all continuous data must possess is that there must be an inherent order to the data. For example, if we know for sure that 30 cm is shorter than 40 cm, then the variable "length measured in cm" has ordinal properties. However, some data have only ordinal properties to them, and no other properties (besides distinguishing them from the other values of the variable).

For example, take the Olympic medal system. The first-place winner gets a gold medal, the second-place winner gets a silver medal, and the third-place winner gets a bronze medal. This is an excellent example of ordinal data because there is clear ranking to the system. If two Olympians started up a conversation with us at bar, and one said, "I won a silver medal," while the other one said, "I won gold," we would know who got the higher rank, because somewhere at some point a group of people decided that this is how the Olympic medal system would work.

Figure 2-2 The Olympic Medal System Is Ordinal

In other words, ordinal data do more than just differentiate between their levels—they also rank the values in some order.

Other Types of Ordinal Data:

- High School Class Rank
- Rank in the Military
- Ranking one's top 5 favorite movies
- The order in which my children were born
- The stops that a train makes on its route
- Shoe size

Having inherent order is a useful property in data, but it lacks other information that may be very helpful. For example, what we do not know about the Olympic medal system is the distance between the rankings. Let us try a real-life example to explain this issue:

In the 2016 Summer Olympics in Brazil, Katie Ledecky of U.S.A. competed against several other swimmers in the Women's 800 Meter race. If I were to use only ordinal data to explain the outcome, I would say that Ledecky won gold, that Jazmin Carlin (United Kingdom) won silver, and that Boglárka Kapás (Hungary) won bronze.

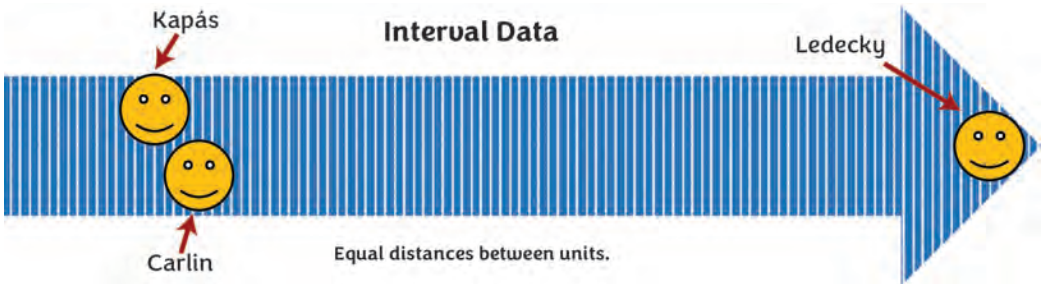
We could then know who did better than who. However, we could not then answer other questions like *how much* better Ledecky did than the other swimmers. It would sound a little silly to say that Ledecky performed "two medals better" than Kapás. To know how much better Ledecky did than Kapás, my data would need to have another property—equal intervals between the values, which is the next property we will examine.

2.2b Interval Data

Interval data have ordinal properties, too. However, in addition to differentiating between things (categorical) and having an implied order to the values they can take on (ordinal), they also have equal *intervals* between the units of measurement. For example, time measured in seconds (on the planet Earth, anyway) possesses this characteristic (and others we will cover here shortly) and can help us to have a better understanding of Katie Ledecky's swim time in the 800 Meter race.

Ledecky's finishing time for the race was 8:04.79. Carlin's finishing time was 8:16.17. Kapás's finishing time was 8:16.37. In other words, Katie Ledecky's swim time was enormously better than the silver medalist's, but the silver medalist beat the bronze medalist by just a fraction of a second. Now that is some interesting information!

Figure 2-3 Interval Data Has Equal Distances Between its Units of Measurement



Another example of data that have interval properties is I.Q. scores. If Harry has an I.Q. score of 112, and Jin has an I.Q. score of 114, we can state exactly *how much higher* Jin's score is than Harry's. It is better by 2 points. We could also know that the difference between an I.Q. score of 110 and 100 is the same as the difference between 90 and 100. This is all because there are equal intervals between the units of measurement. This characteristic of equal intervals between the units of measurement is especially useful in statistics, because it allows us to calculate many of the things we will need to know, like means and standard deviations, but we will cover that all in Chapters 4 and 5.

Other Examples of Interval Data:

- SAT scores
- GRE scores
- Golf score

One thing to note with interval data is that they do not need to have a meaningful starting point. Just as a golf score of zero does not mean that one had zero strokes of the club, and it is impossible to get a score of zero on the SAT (the lowest score is 400), it turns out that only some kinds of data have meaningful starting points.

I.Q. scores are on an arbitrary scale. That is, the average I.Q. score is 100, and it is 100 just because somebody decided to call 100 the average. There is no reason it could not have been any other number. Additionally, it is impossible to score an I.Q. score of 0, because it does not actually mean anything. We cannot measure a total lack of intelligence in a human, so there is no 0 on the I.Q. scale. However, some data do have meaningful zero points, and that is a very nice characteristic. We call data that has a meaningful zero point **ratio** data.

2.2c Ratio Data

For example, time measured in seconds can be considered ratio data because it has all of the properties of interval data, but it also can be zero, in theory. At least it starts at zero, and then goes up. But an I.Q. score cannot be zero, because that number does not exist on the scale, and it would not actually represent any real phenomenon. I.Q. does not start at zero and then go up. The scale does not exactly "start" anywhere.

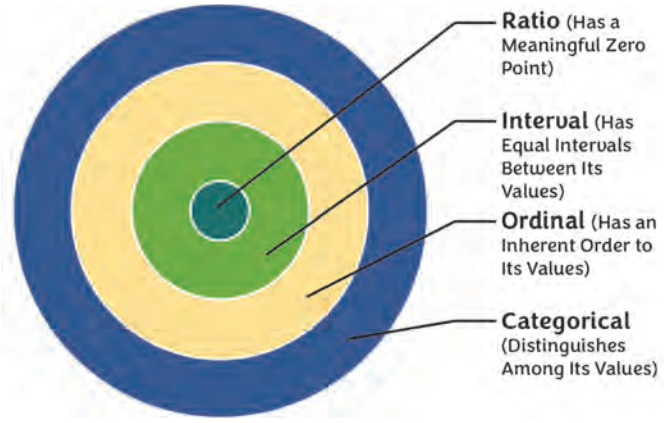
Another example of ratio data is the Kelvin scale for temperature. Temperature is measured in degrees, and these incrementally increase on for infinity (in theory). They originate at a true zero point that means there is an absolute lack of motion in the molecules of the substance being measured (that is the definition of heat). Because zero on this scale means a total lack of heat (at least, in theory), then Kelvin is a ratio scale.

Other Examples of Ratio Data:

- Speed in miles per hour
- Blood alcohol content
- Number of cigarettes smoked in a year
- Daily salt intake
- Population of a city

Note that data with ratio properties must also have interval and ordinal properties. Interval data must also have ordinal properties, but not necessarily ratio properties. Additionally, all types of data must be able to distinguish each value they could be from the other values they could be (like categorical data do, at a minimum). Otherwise, it would not really be information.

Figure 2-4 All Data Distinguish Among Values, but Not All Data Have Other Properties



2.2d Psychological Scales

Generally, most psychological scales are treated as interval scales. That is because psychological constructs rarely have a real-world zero point. Think about mood, for example. If we were to design a measure of how happy someone is, could we really measure zero happiness? Would that mean that someone completely lacks happiness, or would it mean that someone is actually depressed? If we are actually measuring the presence of depression, then that is not really a lack of happiness, but maybe the presence of something else, so there probably is not a meaningful zero point on that psychological scale. Similarly, what would a score of 0 on an I.Q. measure mean? Would it mean that the respondent lacks any intelligence whatsoever? Can we even measure a total lack of intelligence? To take the test, the person must at least be able to communicate, so that *per se* indicates some intelligence. Therefore, an intelligence test likely has only interval properties.

Notice that we also typically assume that most psychological scales have equal distances between the units. For example, if I ask someone to tell me how well their last date went on a scale of 1 to 10, we treat that as an interval scale, but some people would argue that it is not truly interval, because do we really have an objective measure of a mental phenomenon? How do I know a score of 8 for one person means the same thing as it does to another person?

The truth is that I do not. We usually treat these sorts of psychological scales as having interval properties, but we must take that with a grain of salt. Psychological phenomena are challenging to measure at all.

Furthermore, things can start to get muddy very quickly if we split hairs on the data's characteristics. For example, would we consider a person's birthdate ratio data? Someone might argue that year is ratio data because there was a year zero on our calendar, but someone else could counter that the year zero in our calendar is just arbitrary and does not actually mean that there was no year before it. Maybe it is interval data, then. Another person might say, "But guys. The Big Bang." If we subscribe to the idea that time did have a starting point billions of years ago, then time is ratio data, but if the universe is constantly expanding more and more rapidly, then what is that doing to time? Maybe there are not equal intervals between years even.

Although this is maybe an interesting philosophical and astrophysical discussion (see Rovelli, 2018), for the purposes of a study that is interested in how birth year might be related to some other variable, then it makes no real difference if time has been slowing down or speeding up or whatever since the Big Bang. In that case, we can just calm down and treat the variable as ratio data.

2.2e Note of Caution

A statistician ultimately makes the decision about whether the analysis they choose is most appropriate for the data they have, and so it is vital that they take note of the properties of their data. For example, although ordinal data are continuous, the lack of interval properties often makes it more appropriate to use tests set up for categorical data. Additionally, even though some variables are continuous, if there are very few values they could take on, a researcher may wish to use a statistical test that is normally reserved for categorical data. In any case, researchers must make these decisions carefully once they design their research projects, and if one is not sure what to do, they should consult with someone who has a lot of experience with statistics.

Practice

What sort of scale would be most appropriate for these continuous variables? Explain why:

1. Alphabetized last names.*
2. Speed in miles-per-hour.†
3. Year based on the current calendar.‡
4. Body mass index.§
5. Barometric pressure.¶
6. Religions ranked by the year in which they were first founded.**

---

* The alphabet, when used to determine the order of things like last names, certainly contains ordinal properties, and probably also interval properties if each letter is considered a unit. There is no meaningful zero, so it is not ratio data.

† Speed in this context is certainly ratio data as there are equal intervals between units (miles per hour), and there is a meaningful zero. One could be traveling zero miles per hour.

‡ Our current calendar certainly has ordinal and interval properties, but it may not be ratio data. Was there a year zero? Was there a beginning to time at all? Depends on whom we ask. It is probably fine to treat it as ratio data.

§ Body mass index cannot be zero. It is an index of the ratio of height to weight, so there is order to it, but I think most people would argue that it does not have equal intervals.

¶ Barometric pressure would be ratio data, as it can have a zero point, and has equal intervals between the units of pressure.

** Even though religion itself is categorical, Timmy specifically wants to rank them using other information about the religions (not the qualitative belief systems). So, Timmy wants to create an ordinal dataset from the year the religion was founded, which is ratio data.

2.3 Independent and Dependent Variables, and Constants

We have discussed "variables" a little bit already, but let us clarify what that means now, and then we will explore what different variables there are.

A **variable** is any phenomenon that is likely to differ or "vary" among the units being measured. In psychology, we usually study people, so people are the units. All sorts of things can vary among people. That is also why psychology is so interesting! A person can differ from their roommate in age, ethnicity, religious background, favorite food, birth order, parental education, ambition, personality, and thousands of other things that are potentially measurable. Those are all variables, because they can vary among the units (i.e., people) we use to measure them.

A variable is in contrast to things that are **constant**. A constant is something that is not expected to vary among the units that are measured. For example, if I survey a student and their roommate about something, the units I surveyed are both human beings (at least that is a pretty good guess). So, their species would be a constant. Perhaps a researcher wants to study Catholics' attitudes about the current Pope. As long as the researcher recruits only Catholics into her study, then Catholicism would be a constant, because by design, the researcher did not expect or allow the respondents' religion to vary.

Researchers often expect lots of things to remain constant in their experiments, and good researchers make efforts to ensure that some potential variables do remain constant. A great example of this is in an experimental study where the researcher wants to focus on changing only one thing in the conditions of the study, but nothing else. Imagine a researcher who wants to know how anxious people feel after viewing a scene from a scary movie. He already knows that they will feel different levels of anxiousness—that is sort of the point. What he needs to do, though, is make sure that the only source of fluctuation in anxiety is the movie scene, and nothing else. That means he needs to make sure everyone sees the movie scene in the same room, sitting in the same chair, at the same volume, and having slept around the same amount the night before, and so on. If anything else could be affecting the fluctuation in anxiety, the researcher needs to either account for it (by measuring it), or exclude it (by making it a constant). If half of the participants see the movie scene before lunch, they may report greater average anxiety than the half of participants who see it after lunch, and that would potentially lead to a Type I Error.

Researchers want a lot of things to be constant in their research, but then they want to focus on the variability that occurs in the variables that interest them. There are two main types of variables that are vital to understand in research methods and statistics. They are important for statistics, because some equations have a specific place that counts for one of them rather than the other. If we get them wrong, we may commit a Type I or Type II Error.

In a simple study, the **independent variable** is the thing that varies among the units (usually people) that is also theoretically what causes the change in the **dependent variable**. The dependent variable is what is thought to vary due to changes in the independent variable. In other words, the independent variable is thought to have an effect on the dependent variable. The dependent variable "depends" on the independent variable, in theory. Let us dissect this a little more.

The independent variable can be identified in many ways. Often, the clearest way to do this is to look at the study's description and ask, "What does the researcher think is causing some change?" For example, say that Professor J. wants to see how an intervention affects bullies' perceptions of their behavior. In this case, Professor J. thinks that it is the intervention that is making a change in the bullies' perceptions, and so the intervention is the independent variable.

Although the best definition of the independent variable is that it is thought to cause the change in the other variable, there are also some other helpful ways to identify it. One clear sign that something is an independent variable is if it is something that the researcher has introduced to the units (usually people) in their study. Professor J., for example, brought the intervention to the sample of bullies. They did not have it before, and they did not think it up—Professor J. is the one who introduced it into the sample. That is a pretty good sign that it is the independent variable.

Additionally, if the researcher **manipulates** a variable, that is a clear sign that it is also the independent variable. A manipulation is when a researcher makes something different among participants on purpose, so that they will get different experiences in the research. The simplest example of a manipulation is when a researcher gives one pill or treatment to some of the participants but gives another pill or treatment to the rest of the participants. The researcher not only introduces the variable to the sample, but they also changed it among the sample on purpose. It is also what they think will have an effect on the dependent variable (treatment outcome, for example), and so it is the independent variable.

Note that sometimes researchers are not able to control or even introduce some independent variables, so it is best to get proficient at identifying independent variables by their role in the study: whether it is the thing being changed or the thing causing the changing of the other thing. That is the clearest way of identifying the independent variable.

For example, say a researcher wants to know how a person's genetic sex affects their career aspirations, if at all. In this instance, the researcher has no ability to (ethically) manipulate the genetic sex of the people in her sample. The variable varies naturally. Additionally, this researcher did not "introduce" genetic sex into the sample, because it was already present among the sample for years before the researcher recruited the sample. Still, even though the researcher did not introduce it and cannot manipulate it, it is still what she thinks influences the other thing she is measuring, which is career aspirations.

The independent variable is believed to have an effect on the dependent variable. That is, the dependent variable is thought to be affected by the other variable in the study. That is the clearest definition of the dependent variable, but there are also some other things to look at that may help in making a determination. The dependent variable is also usually the thing that the researcher measures after having inputted the independent variable. Thus, chronologically speaking, the independent variable comes first, and then the dependent variable is measured.

For example, do career aspirations usually get determined before or after genetic sex is determined? Obviously, genetic sex is determined at the moment of conception, before the resulting human has self-awareness, and so genetic sex clearly comes before career aspirations. Thus, this is another strong clue that genetic sex is the independent variable, and career aspirations is the dependent variable in this particular example. To summarize how these variables work together, this mnemonic should help:

Independent variable → dependent variable

If we just remember, "The independent variable theoretically causes a change in the dependent variable," we should have little trouble identifying which is which. We can also remember, "The independent variable comes before the dependent variable chronologically."

Practice

Identify the independent and dependent variables in the following studies:

1. Dr. L. has some volunteers drink alcohol and others drink a **nocebo**, and then assesses their driving abilities on a simulator.*
2. Dr. B. wants to see if elementary schoolers are more extraverted in 3rd grade or 4th grade.†
3. Dr. O. is curious about whether dog-owners differ from cat-owners in their parenting styles.‡
4. Dr. I. asks volunteers to read one version of two statements and then quizzes them on the content to see which statement may improve retention.§

* The condition of alcohol or placebo is the independent variable. Dr. L. introduces the condition of either alcohol or placebo, and also manipulates which one the participants receive. Dr. L. also theorizes that the condition they receive is what affects driving ability. So, driving ability is the dependent variable.

† Dr. B. believes that it is the grade level that may influence extraversion, and so grade level is the independent variable, even though Dr. B. does not introduce or manipulate it. Extraversion is the dependent variable

‡ This one is tricky. The independent variable is what sort of pet the person owns. Although it does not clearly precede the dependent variable–parenting style–it is what the researcher seems to theorize comes first, or at least to cause the variation in the dependent variable. Of course, this one could easily be reversed if the researcher believes that parenting style predicts type of pet

§ The version of the statement is the independent variable, and retention is the dependent variable. The version of statement is clearly manipulated by the researcher, and that is what is presumed to affect retention

2.4 Populations and Samples

While we are distinguishing the kinds of data that we can gather, it is also worth spending some time on whence the data come. Let us first learn some new vocabulary that will help us out:

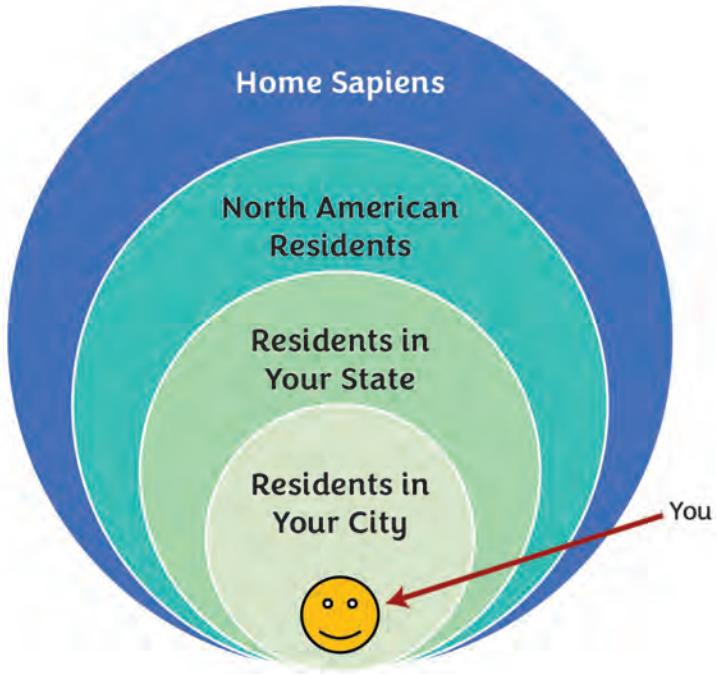
A **population** is every single unit that shares a definable characteristic. To simplify that a little bit, let us focus just on what we usually do in the social sciences. Psychologists, for example, study people almost exclusively. We are interested in all sorts of different kinds of people: young people, incarcerated people, sick people, influential people, pregnant people, people who have been through trauma, people who are impoverished, people who go to college, and so on. Each of these characteristics, and countless more, describes a **population**. When a researcher is interested in a particular characteristic about people, they are truly interested in every single person with that characteristic. If Professor Vivek. studies serial killers, he really wants to learn about *all* serial killers. If Dr. Updike studies African American racquetball players, she is curious about *everyone* who could fit that description. They are interested in *populations*.

The problem for researchers is that populations are often very large and/or hard to reach. It is, therefore, usually impossible (or impractical) to get information from everyone who conceivably fits within a population. For example, imagine for a moment that Professor Wu wants to see how to reduce cravings for cigarettes among heavy smokers, using some kind of new hypnosis technique. Professor Wu presumably wants to know how to reduce cravings in *every* heavy smoker on the planet, but of course there are so many heavy smokers that she has no possible chance of actually finding each one, asking them to be in her study, they also agreeing to do it, and then complying with the protocol. Professor W. has to do her research with only 2 years and with a budget of only $10,000, and so that is not enough time or resources to find everyone she would like to study. So instead, what she settles for is a smaller subset of the population, which is called a *sample*.

A **sample** is a subset of the population that is supposed to also represent that population in the study. In other words, the people who make up the sample are also members of the population—they share the characteristic that interests the researcher—but they are only some of the people who make up the entire population. Professor Wu might settle for reaching 300 heavy smokers and recruiting them to be in her study. Those 300 are not the only heavy smokers in the world, but if she carefully selects them, she can help ensure that those 300 are essentially representative of all other heavy smokers in the world. It will be fun to dive more into how she can do that, but let us first unpack the notion of populations and samples a little more.

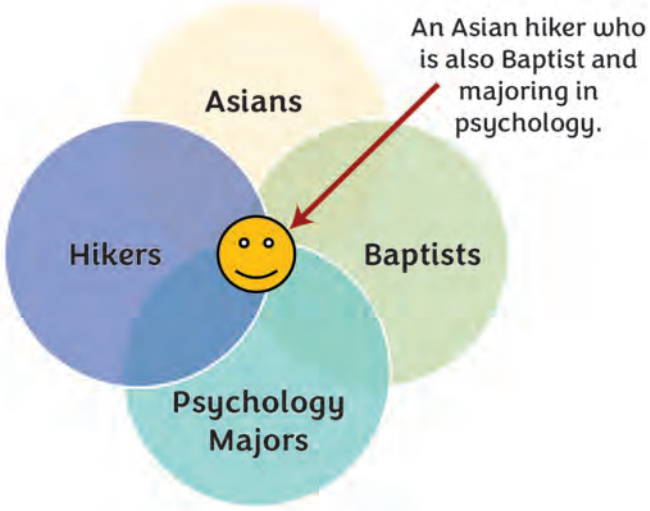
Let us think about ourselves for a moment (not in a conceited way). How many definable characteristics do we have that might interest a researcher? We all belong to a virtually infinite number of populations right now, I would wager. We are (presumably) both human beings, and lots of researchers are interested in how things affect all human beings. We must both be (at least semi-) literate if we are using this text, so we are literate human beings also. Our genders are characteristics that are interesting for many research purposes. Our type of employment is a definable characteristic that gets some researchers' tails wagging. Our marital status is interesting, and so on.

Figure 2-5 Everyone Belongs to Many Subpopulations Within Populations



Everyone belongs to many populations. Some of them are more exclusive than others. Many of the populations overlap with others. Additionally, each person may belong simultaneously to several populations that do not necessarily overlap with the others.

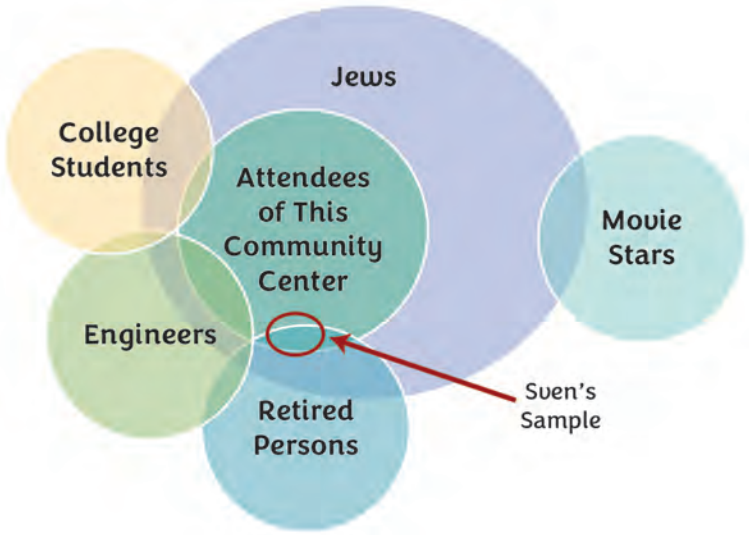
Figure 2-6 People May Belong to Many Populations That Do Not Overlap Very Much



We could easily map out things about a person that do not necessarily go together. Plenty of people who are in the same population in one area differ from each other in another area. Using the illustration above, there are plenty of hikers in the world, and most of them are not psychology majors. But maybe some people are both of these things simultaneously.

The fact that everyone on the planet simultaneously belongs to many different populations may also be a problem for researchers. For example, imagine that Sven is doing research involving people of the Jewish faith. He is interested in how they perceive the current political climate. He goes down to a local Jewish community center at 12:00 p.m. on a Tuesday and asks 20 of the people there to complete his survey. On the plus side, all 20 of the people in his sample also belong to the population in which he is interested—they are all of the Jewish faith. But what may be a problem is that they all may be from only certain subpopulations of people who are of the Jewish faith.

Figure 2-7 A Sample That Fits Within the Population, but Perhaps Too Narrowly



Because he went to only one recruitment center at one time of the day, it may be possible that his sample includes mostly retired people (who are not at work on a Tuesday at noon), or people who are even more similar to each other in their views than Jews are generally (perhaps they were meeting there for a specific interest group). The challenge for Sven is that his sample fits within his population of interest, but it may not include people who fit into other subpopulations within the larger population.

All of this information is important because researchers should clearly define the population that is relevant to their research, and then they need to take careful considerations to ensure that the *sample* they have in their study represents the *population* they want it to, but not necessarily others.

As a real-life example, when I was working on my dissertation (Ricks, 2015), the population I wanted to recruit was people who practice psychotherapeutic interventions within U.S. prisons. To make sure that the people responding to my survey were actually part of that population, the first question on the survey was something like, "How many hours of individual therapy do you do in an average week?" The second question was, "How many hours of group therapy do you do in an average week?" If they answered "0" to both of those questions, then they were not part of the population I wanted. I had advertised my study to thousands of people, and somehow a few people thought that they belonged to the population I wanted, when they really did not. Maybe they were administrators or something at the prisons, but they did not do any psychotherapy in their jobs, so they did not actually fit my population of interest. Had I kept them in my dataset, they would have answered the questions in ways that did not actually represent the population, because they did not belong to it. That is why I had those screening questions—so I could remove them from my dataset.

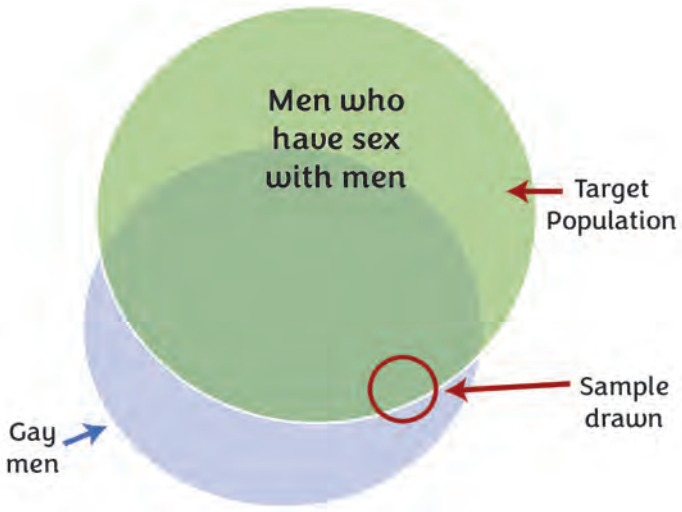
One reason that those administrators thought they were appropriate for my study was probably that they did not understand the **operational definition** I used for "prison psychotherapists." I knew what I meant, but perhaps they understood it to mean something else. My study was advertised for "prison psychotherapists," and that could mean lots of things to different people. Is that people with a social work degree who hold some self-help groups? Does it include clinical psychologists who just do research at the prison? Does it include psychiatrists who just prescribe medications to prisoners? If I do not set up clear criteria, then I leave all of this open to interpretation, which means that I may draw in participants from populations that I do not want to include in my research.

Operational Definition
A precise definition of the term or terms the researcher uses in their study

Consider this. Imagine that we want to study the safe sex practices of men in homosexual encounters. We advertise our study at gay community centers as a study of "gay men." What is interesting here is that by using these methods, we may be excluding the population that truly interests us. For example, in prisons there is a phenomenon referred to as "gay for the stay." This phrase describes when a man who identifies as normally heterosexual will have voluntary homosexual relationships while incarcerated, and then once he is released from prison, he ceases to have such encounters. In the literature, this phenomenon is sometimes also called "situational homosexuality" (e.g., Gibson & Hensley, 2013). This type of person does not identify as gay, and yet still engages in the behavior about which the researcher wishes to learn.

Another Venn diagram may help to illustrate this issue:

Figure 2-8 Sometimes the Sample Does not Match the Target Population



If gay men make up the blue circle, and men who have sex with men make up the green circle, then there is plenty of overlap, as one would expect. The overlap represents gay men who have sex with other men. The portion of the green circle that does not overlap the blue one represents men like those in prison who do not identify as gay, but who still engage in sexual behavior with other men. What is also interesting about this diagram is the sliver of the blue circle representing gay men that does not also overlap with men who have sex with men. This sliver represents men who identify as gay, but who do not actually have sex with other men, perhaps because they are married to a woman, they hold religious convictions that prohibit such behavior, they dislike physical touch from other people (e.g., Haphephobia), and many other reasons.

The challenge for the research is that if the researcher recruits only "gay men," they will also likely include some people from this sliver, who do not actually represent the population of interest. As an additional problem, the researcher is going to miss a portion of the people who fit into the population he really wants to study—men who have sex with men.

As we see, it is a vital early step in research to clearly identify the phenomena that interest the researcher, and then to use that information to inform their sampling. Note that there is not always a "correct" definition of these populations, so it can be tricky to come up with a good one for a study. When considering an operational definition, it is important to do a lot of reading about what other researchers have suggested as definitions, or what they have used in the past. That is usually a good place to start, and it helps with consistency among findings. In some cases, it may be most appropriate to have the participants self-identify their population, because definitions may not be generally agreed upon.

Practice

Write a clear operational definition one could use for each of these groups. Someone reading the definition should have a clear idea of whether they fit into the target population.

Example. **People taking antidepressants**–"Any person who has been prescribed a psychotropic medication for symptoms of depression by a medical professional, and who also has regularly taken these medications as prescribed for at least 6 months."

1. **Smokers**–
2. **Adolescents**–
3. **Juvenile probationers**–
4. **Bird watchers**–
5. **Pet owners**–
6. **Mexican Americans**–

2.5 Random Selection and Random Assignment

When we have the operational definition, we now know exactly whom to try to include in our study. That is a wonderful first step. The next thing we need to do is to make sure that the sample we want to recruit actually represents the population we think it does. As we explored above, we also want it to *proportionally* represent other subpopulations within the main population that interests us.

The way to ensure that a sample actually represents the population we want it to is to **randomly select** our sample from the population. That means that we take steps to ensure that every single person (or unit) in the population we carefully defined has an equal chance of being included in our sample. That is not very easy, depending on the study, but it is important that a researcher does what they can to make it happen. Here is a real example of what can be done:

For my dissertation (Ricks 2015), the population that interested me was people who professionally practice psychotherapy within a United States prison. In a *perfect* world, I would have had access to some large database of everyone who fits my criteria, and I could put each person's name on a card and then shuffle them and draw out the sample size I wanted. In the *real* world, I could not do that, because there is no such database. Instead, I spent about a year contacting departments of corrections in each state, and asking if they would be willing to send emails to all of their mental health staff. I also contacted professional organizations whose members likely are mental health professionals in corrections, and I contacted private companies who do contracted mental health work in prisons. Even though I could not officially randomly select prison therapists, I did what I could to make sure that as many prison therapists who fit my criteria could hear about the study and decide to participate.

Random assignment is another vital step in research. This step is used in situations when there is a manipulation, like a therapeutic intervention, that the researcher introduces. If the independent variable has two levels, for example, what they should do is randomly assign the participants in the sample to one of the levels. That means that everyone they recruited for the sample has an equal chance of being in any of the levels of the independent variable. Imagine that they want to know if Therapy A works better than Therapy B in treating anxiety. Let us then imagine that they do a good job of random selection. That is great, but if they do

not assign these participants randomly to either Therapy A or Therapy B, they may increase their risk of a Type I or Type II Error. Imagine that they assigned the first half of participants who signed up for the study to Therapy A, then the second half who signed up get Therapy B. What may have happened is that the people who were slightly more motivated to sign up for the study did so first, and then they all got grouped together into the same therapy. Now, when the researcher measures the dependent variable, they may find that those in Therapy A showed more improvement, but it may not have had anything to do with the therapy—it may have had more to do with the fact that they were generally more motivated people.

Random assignment is important to prevent these preexisting differences from affecting the outcome of the study. Its major advantage is that it virtually ensures that the groups to which a researcher assigns participants do not differ systematically, because the researcher shuffled them up randomly. If we flip a coin to decide who ends up in which group, then there is a very low possibility that all of the people who are very excited or motivated end up in the same group together—we leave that to chance.

Sometimes, we will find that we cannot randomly assign people, just like we cannot always randomly select our sample. For the same reason that we cannot always manipulate an independent variable's levels, we sometimes cannot randomly assign people to their conditions. For example, maybe Professor Updike suspects that people who live within city limits are more likely to vote than are people who live outside of city limits. Professor Updike can take some good steps to try and randomly select people for her study, but she cannot randomly assign people to live within or without city limits. She cannot say, "Ok, thanks Mr. Jones. For this study, would you please relocate your family to an address within the city limits for the duration of this study? That would be super." That may be super for her research, but it is impractical, and possibly unethical, to do something like that. Thus, even though Professor Updike wants to measure the effect of the independent variable, she will be unable to randomly assign people to one of its levels.

There are many important considerations for a researcher seeking data to analyze. The researcher will need to carefully consider how they are getting their data and what sorts of characteristics their data will have. These decisions should come up early in the research process, as they will affect everything moving forward, including the kinds of statistical analyses that can be used and how valid the study's findings are.

Summary

- Different sorts of data have different characteristics. Some possible values of a variable describe only differences, while others can be ranked, others can be measured by distances apart, and yet others also have a starting point that corresponds to some objective phenomenon.
- These characteristics of data are an important starting point toward knowing what type of statistical analysis is possible and appropriate to examine patterns within the dataset.
- Appropriate selection of the units from which to measure these data is also vital in the research process.