1.2: Tenets of Statistics

There are three main tenets of statistics: Variability, Probability, and Uncertainty. Let's review some key terms and concepts that connect to and help us understand these three tenets.

Variability

In statistics, data are collected about variables. **Data** is the plural term used to refer to multiple bits of information together. The term **data set** can also be used to refer to data from one or many variables. Datum is the singular form of data which is used to refer to one piece of information. **Variability** is a broad term that refers to the fact and ways in which things can differ. A **variable** is anything that is measured which is not always the same. For example, we could collect data about the rainfall in inches for each day of January. The variable measured would be rainfall. Though it is possible for the amount of rain measured each day to be exactly the same, it can (and likely will) be at least a little different on some days compared to others. Thus, the thing being measured (rainfall) is considered a variable. The data for this variable would be all the numbers for the amounts of rainfall each day. For example, it might have rained 0.00 inches on the 1st, 1.25 inches on the 2nd, 0.40 inches on the 3rd , 1.25 inches on the 4th and so on. These numbers (0.00, 1.25, 0.40, and 1.25, respectively, are the first four pieces of data in the data set).

In contrast, a **constant** is anything which is measured that cannot or does not vary. Whenever something being measured is always the same or did not differ at all among the instances being measured, it is considered a constant. For example, if a person grew up in a town called Statistonia, and each day you collected data on where they grew up, the data would be the same every day. Every day the data for the variable *Childhood Town* would be *Statistonia*. It doesn't matter how many times you ask it, the data are always the same so there isn't much you can do or say with these data. This makes constants quite limited in their utility and, thus, they are rarely of interest to statisticians when testing hypotheses.

Instead, statisticians are particularly interested in examining data collected about variables. The nature of variables means that they might have interesting patterns on their own or in relation to other variables. Let's consider an example. We can measure the variables income and happiness to see if there is any pattern between them. If we measure income and level of happiness among many people, we will likely see that income differs from person to person and that happiness also differs from person to person. Some incomes may be quite low relative to others while others are quite high. We might see a pattern where incomes are generally lower than $60,000 but that a few people make well over $100,000. We would likely also see variability in happiness. For example, if individuals rated their happiness on a scale from 1 (not at all happy) to 10 (extremely happy), we might see that most people rated their happiness somewhere between a 4 and a 6 but that a few people rated their happiness below this range and a few others rated their happiness above this range. With these kinds of descriptions, we are summarizing some patterns we see in the data for each variable. Already, this is interesting but let's take it a step further. We may be curious whether income relates to happiness such as whether happiness tends to be higher when income is higher. Because both things which were measured varied, statistical techniques to assess this proposed relationship are possible (and we will learn about them and how to use them in Chapter 12).

However, patterns like these cannot be discerned when looking at constants on their own nor in relation to variables. Let's take the same example but consider what would happen if income was a constant. Suppose that when the data were collected, income was a constant $60,000 meaning every person made this amount. There isn't much we can say other than "the income was a constant $60,000." That's it. Even if happiness varied, we couldn't say whether there was a relationship between income and happiness because such a pattern cannot be established or tested when either or both of the things which were measured were constants; we can only establish or test patterns between and among variables. Therefore, the field of statistics is primarily focused on variables.

Probability

In statistics, data are collected from samples. Thus, we can say that statistics is the study of data from samples. However, these data are often collected from samples in the hopes of understanding populations. The distinction between populations and samples is, therefore, important for understanding why and how probability is central to much of statistics.

Populations

A **population** includes all cases that comprise a specified group. The term **case** is used to refer to a single member of a population. Another way of saying this is that a population includes all the examples of who or what we are trying to understand. For example, a statistician might be interested in understanding college students. The defining characteristic that identifies who the statistician wants to understand is "college student." Therefore, the population would be comprised of every college student that exists. However, it is often difficult and is sometimes impossible to collect data from a population. Think of how difficult it would be to

identify and collect data from every single college student that exists. Even if we could identify and locate every college student, for practical and ethical reasons, we are unlikely to be able to get data from all of them. If even one college student cannot be located or doesn't want to provide information about themselves, the data would be incomplete. Therefore, it is far more common to have data from samples than from populations.

Samples

A **sample** includes some, but not all, cases that comprise a specified group. Another way of saying this is that a sample includes some of the cases from the population we are trying to understand. By definition, then, sample data are incomplete because data are not available and known for at least one member of the population of interest. Depending on the size of the population and the ability to identify each case and collect data, data may only be available from a relatively small sample of population members. For example, though there may be millions of college students, data might only be collected or available from 250,000, 10,000 or maybe even only 50 of them. Each of these would be samples of the population of college students. When a result is summarized from the full population, it is referred to as a **parameter**, similar to how a result summarized from a sample is referred to as a statistic. You can remember these terms alliteratively: Populations provide parameters while samples submit statistics. Because statistics are yielded from sample data, which are incomplete representations of population data, statisticians can only deduce what is probably true about populations rather than what is definitely true. Therefore, the theme of probability appears throughout statistics.

Probability refers to how likely something is to occur or be true. Because data from samples are both more available than those of populations and have the limitation of being incomplete, the field of statistics developed techniques that use probabilities to estimate what is true based on sample data. These techniques and, thus, statisticians, will always have to deal with some ambiguity. Statistical procedures are used to estimate the probabilities that various observations in sample data represent realities in the populations. Therefore, probability is central to statistics.

Some statistical procedures are built from assumptions about the patterns of data in populations. These are referred to as parametric statistics. **Parametric statistics** refer to techniques that use data from samples that are assumed to have been drawn from populations which are distributed in specific ways. These distributions are known as normal distributions which we will learn about in detail in Chapter 5. Essentially, this means that the data are assumed to follow certain patterns in the population and that probabilities can be calculated based on those assumptions. We will learn how to use some parametric statistics in later chapters (such as Chapters 8, 9, and 10). There are also non-parametric statistics. **Nonparametric statistics** are analytic techniques that are used when data are not assumed to follow the normal distribution a priori. Instead, these techniques are used when either the assumptions of parametric statistics are violated or cannot hold true because of the types of data being used. We will learn about one such test in Chapter 14. For now, it is important to know that probabilities are important components of both parametric and non-parametric statistics.

> *There are no guarantees in Statistics.*

Uncertainty

Because statistics yields estimates of probabilities, it means there will always be some uncertainty. The fact that statistics is focused on data which vary and are incomplete (because they come from samples rather than populations), means that statisticians focus on what is probably true and cannot deduce what is definitely true. Statisticians, instead, use procedures to deduce what is likely true about populations and their parameters but cannot 100% guarantee that these things are true. For these reasons you will notice that statisticians often use hedging language. For example, a statistician may have strong evidence suggesting that those who study more tend to earn higher grades. However, even if the data are quite strong, they cannot guarantee that all individuals who study more will earn higher grades. Therefore, they are likely to say something like "Studying more tends to produce higher grades" or "Increasing your hours of studying will likely improve your grades" rather than something absolute such as "If you study more, you will get higher grades." This form of hedging language is an important way of communicating that every analysis has some limitations, that nothing in statistics is a guarantee, and that those who are skilled in statistics are always open to the possibility that there are disconfirming data out there.

Uncertainty is an important reality of working with data and statistics that some people find frustrating, and for good reason. Of course, we would ideally like to have certainty and guarantees over probability but that is not always possible. This does not mean, however, that the field has nothing to offer and that we should discard our book and abandon our study of statistics here. Here's why: knowing what is probably true has value and can be useful in guiding decision-making.

Let’s take a quick but dramatic example. Suppose someone has a serious illness where the chance of dying within 6 months is 80% if left untreated. The 80% is a descriptive statistic that summarizes a probability. Suppose that two medications are available to treat the illness and a doctor must decide what to prescribe. Suppose the risk of death for those who took Medication A in a large study was only 30% and the risk of death was 80% in those who took Medication B, with no side effects reported in either group. Though there is still uncertainty because nothing in this example states what occurs 100% of the time, there is still information here that is useful for the doctor and patient as they make decisions about treatment options. Additional information about how similar the patient is to those in the study and statistical findings for other potentially relevant factors and studies would also be used to determine the treatment plan that is most likely to benefit the patient. When certainty is not possible, probability is the next most useful option and is more beneficial than being tasked with making decisions in the absence of any information.

Reading Review 1.1

1. What is the overarching goal of statistics?
2. Which tenet refers to the fact that statistics focuses heavily on data collected about things which are not the same for all cases?
3. Statisticians focus on data from _____, which include some, but not all, of the cases of interest.
