1.3: Two Realms of Statistics- Descriptive and Inferential

Learning Objectives

- Define data
- Define descriptive statistics
- Distinguish between a sample and a population
- Define biased sample
- Introduce sample statistics and population parameters
- Define inferential statistics

Descriptive Statistics and Inferential Statistics

As we have discussed, our lives and the scientific method involve significant amounts of observation and experimentation. During both of these processes, we are gathering information. **Data** refers to information that has been collected from observation, experimentation, surveying, historical records, etc. We study the collected data to understand what is happening around us. Looking for patterns in raw data can be difficult, especially if we have many observations and measurements. In this course, we will review and develop various ways to summarize and visualize raw data. Numbers that are used to summarize and describe data are called **descriptive statistics**. In order to understand a data set sufficiently, we must use several descriptive statistics.This need will be explored in depth in Chapter 2.

? Text Exercise 1.3.1

Data and descriptive statistics are closely related to each other and are sometimes confused. For each of the following claims, identify the data and any descriptive statistics. Note that sometimes the data is implied as opposed to directly given.

1. The average score on Exam 3 was 76% for this statistics course last semester.

Answer

The data is only referenced implicitly. Since we are looking at the average score on Exam 3 from last semester, the data would consist of all scores on Exam 3 from this course last semester. 76% provides a summary of the data and is a descriptive statistic.

2. We spent $\$4160$ on groceries and household goods last year.

Answer

Again the data is only referenced implicitly. Our data consists of the expenditures from the previous year related to groceries and household goods. These costs could be found on receipts, bank records, credit card statements, or some combination. The value $\$4,160$summarizes the data by summing all the values together and is thus a descriptive statistic.

3. We have four children, aged 2, 4, 6, and 9. The oldest is 9 years old, and the average age is 5.25 years.

Answer

The data explicitly consists of the ages of the four children: $\{2,4,6,9\}$.The oldest being 9 summarizes the data by providing us with the maximum value. Both the maximum and average values are descriptive statistics. Notice that descriptive statistics can be values in the original data but do not necessarily have to be.

We collect data each and every day to help us understand the world and act accordingly, but most of the time, our interest lies beyond understanding just the collected data. We hope to generalize, to use descriptive statistics from our collected data to make a claim on a larger scale. This generalization process of extending claims to larger audiences is called **inferential statistics.** We are inferring statistics describing a data set we do not have based on a smaller data set that we do have. This process is necessary as it is impossible to collect exhaustive data for most situations and research questions. Even when collecting exhaustive data is possible, we still utilize inferential statistics to help mitigate costs while balancing accuracy.

? Text Exercise 1.3.2

Consider the difficulty of collecting all the necessary data in the following situation and the need to generalize from the possible data.

The National Election Commission has hired us to examine how U.S. citizens feel about the fairness of the voting procedures in the U.S.

Answer

To ask every single U.S. citizen how he or she feels about the fairness of the voting procedures is practically impossible. U.S. citizens live throughout the entire world. Even if all the contact information could be gathered, we could not guarantee a response whether we visit, email, or call. The time and financial costs associated would be prohibitive. Perspectives could change by the time the data collection is completed. Inferential statistics will be necessary. We will need to determine which U.S. citizens to collect data from and use that data to estimate the views of the entire country.

Populations and Samples

In inferential statistics, we draw inferences (conclusions) about large sets of data using data from a small subset of those same subjects or events. The entire data set from all subjects/events of interest is the **population**. Any smaller subset of the population data set is the **sample**. Samples are used to gain insight into the population from which it originated.

In the previous example, the population we are interested in consists of hundreds of millions of U.S. citizens. Those who we actually interviewed would constitute our sample. We would probably sample a few thousand U.S. citizens drawn from the hundreds of millions that make up the population. When choosing a sample, ensuring that one type of citizen does not have more representation than another is crucial. For example, something would be wrong with our sample if the sample happened to be made up entirely of Florida residents. A sample exclusively composed of Floridians should not be used to infer the attitudes of other U.S. citizens. The same problem would arise if the sample were comprised only of Republicans. When these types of situations occur, we say that our sample is **biased**; it over-represents or under-represents a relevant segment of the population of interest.

Inferential statistics consists of mathematical frameworks that convert information about a sample into intelligent estimations about the population from which the sample was drawn. Our estimations depend on how representative our sample is of the population. How can we ensure that our sample is a good, unbiased representation? While the task is impossible without perfect knowledge, we can address the concern by building inferential statistics around random sampling. We trust a large enough, random sample to represent different segments of society in close to the appropriate proportions and that any bias in the sample is purely by chance.

? Text Exercise 1.3.3

Consider the difficulty of collecting all the necessary data in the following situation and the need to generalize from a sample. How could we construct a random sample and estimate the value of interest?

We are interested in examining the average number of math classes taken by current graduating seniors at U.S. colleges and universities during their four years in college.

Answer

Our population consists of just the graduating seniors throughout the country. This is still a large set since there are thousands of colleges and universities, each enrolling many students. In 2022, over 2 million bachelor's degrees were granted in the United States. The cost to examine the transcript of every college senior would be prohibitive. We must construct a sample of college seniors and then make inferences to the entire population based on what we find.

To make a sample, we might first choose some public and private colleges and universities across the United States. Then we might sample 50 students from each of these institutions. Suppose that the average number of math classes taken by the people in our sample was 3.2. Then we might speculate that 3.2 approximates the number we would find if we had the resources to examine every senior in the entire population. But, we must be careful about the possibility that our sample is non-representative of the population. Perhaps we chose an overabundance of math majors or chose too many technical

institutions that have heavy math requirements. Such bad sampling would make our sample unrepresentative of the population of all seniors.

Building from this example, we mentioned that over 2 million bachelor's degrees were awarded in the United States in 2022. Since this figure describes the population, it is what we would call a **parameter**. Furthermore, we collected a sample and calculated that the average number of math classes was 3.2 per student. This figure describes the sample, referred to as a **statistic**. To summarize, a population is described by parameters, while a sample is described by statistics.

? Text Exercise 1.3.4

Identify the population and the sample, then reflect on whether the sample will likely yield the desired information.

1. A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well.

Answer

The population consists of all students in the class. The sample comprises the 10 students sitting in the front row. The sample is not likely to be representative of the population. Those sitting in the front row tend to be more interested in the class and perform higher on tests. The sample may perform at a higher level than the population.

2. A coach is interested in how many cartwheels the average college freshman at his university can do. Eight volunteers from the freshman class stepped forward. After observing their performance, the coach concluded that college freshmen can do an average of 16 cartwheels in a row without stopping.

Answer

The population is the class of all freshmen at the coach's university. The sample is composed of the 8 volunteers. The sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman: people who cannot do cartwheels probably did not volunteer!

? Text Exercise 1.3.5

Determine when descriptive and inferential statistics are being utilized. Assess the quality of the inference.

A quick Google Maps search showed that there were 20 Chick-fil-A restaurants open in Kansas in May of 2024. This means there were about 20 per state for a total of 1000 Chick-fil-A restaurants in the United States.

Answer

It appears that we are interested in the total number of Chick-fil-A restaurants in the United States. To guess the number that characterizes the population (the United States), a sample (Kansas) was taken, and the number of Chick-fil-A restaurants in the sample was determined. Summarizing the sample data with the number 20 would be classified as a descriptive statistic. Estimating the total number of Chick-fil-A restaurants in the United States to be $20 \cdot 50 = 1000$ belongs to the realm of inferential statistics. Descriptive statistics merely describes data, while inferential statistics makes informed guesses about what goes beyond the collected data. The inference is dubious. Kansas is a state with a relatively small population. A better sampling option would be to randomly pick more states. Indeed, if we cared about this situation, inferential statistics would be unnecessary. The desired information is readily available through Chick-fil-A itself; there were over 3000.