7.1: Basic Concepts of Statistics

Learning Objectives

- Understand the basic terminology used in statistics
- Understand the difference between populations and samples
- Classify data as categorical or quantitative

Introduction

Like most people, you probably feel that it is important to "take control of your life." But what does this mean? Partly it means being able to properly evaluate the data and claims that bombard you every day. If you cannot distinguish good from faulty reasoning, then you are vulnerable to manipulation and to decisions that are not in your best interest. Statistics provides tools that you need in order to react intelligently to information you hear or read. In this sense, Statistics is one of the most important things that you can study.

To be more specific, here are some claims that we have heard on several occasions. (We are *not* saying that each one of these claims is true!)

- 4 out of 5 dentists recommend Dentyne.
- Almost 85% of lung cancers in men and 45% in women are tobacco-related.
- Condoms are effective 94% of the time.
- Native Americans are significantly more likely to be hit crossing the streets than are people of other ethnicities.
- People tend to be more persuasive when they look others directly in the eye and speak loudly and quickly.
- Women make 75 cents to every dollar a man makes when they work the same job.
- A surprising new study shows that eating egg whites can increase one's life span.
- People predict that it is very unlikely there will ever be another baseball player with a batting average over 400.
- There is an 80% chance that in a room full of 30 people that at least two people will share the same birthday.
- 79.48% of all statistics are made up on the spot.

All of these claims are statistical in character. We suspect that some of them sound familiar; if not, we bet that you have heard other claims like them. Notice how diverse the examples are; they come from psychology, health, law, sports, business, etc. Indeed, data and data-interpretation show up in discourse from virtually every facet of contemporary life.

Statistics are often presented in an effort to add credibility to an argument or advice. You can see this by paying attention to television advertisements. Many of the numbers thrown about in this way do not represent careful statistical analysis. They can be misleading, and push you into decisions that you might find cause to regret. For these reasons, learning about statistics is a long step towards taking control of your life. (It is not, of course, the only step needed for this purpose.) These chapters will help you learn statistical essentials. It will make you into an intelligent consumer of statistical claims.

You can take the first step right away. To be an intelligent consumer of statistics, your first reflex must be to question the statistics that you encounter. The British Prime Minister Benjamin Disraeli famously said, "There are three kinds of lies -- lies, damned lies, and statistics." This quote reminds us why it is so important to understand statistics. So let us invite you to reform your statistical habits from now on. No longer will you blindly accept numbers or findings. Instead, you will begin to think about the numbers, their sources, and most importantly, the procedures used to generate them.

We have put the emphasis on defending ourselves against fraudulent claims wrapped up as statistics. Just as important as detecting the deceptive use of statistics is the appreciation of the proper use of statistics. You must also learn to recognize statistical evidence that supports a stated conclusion. When a research team is testing a new treatment for a disease, statistics allows them to conclude based on a relatively small trial that there is good evidence their drug is effective. Statistics allowed prosecutors in the 1950’s and 60’s to demonstrate racial bias existed in jury panels. Statistics are all around you, sometimes used well, sometimes not. We must learn how to distinguish the two cases.

Basic Terms

In order to study and understand statistics, you must first be acquainted with the basic terminology.

Data

Data are the individual items of information such as measurements or survey responses that have been collected for a study or analysis.

Statistics

Statistics is a collection of methods for collecting, displaying, analyzing, and drawing conclusions from data.

There are 2 branches of statistics: descriptive and inferential.

Descriptive Statistics

Descriptive statistics is the branch of statistics that involves collecting, organizing, displaying, and describing data.

Inferential Statistics

Inferential statistics is the branch of statistics that uses probability to analyze, make predictions and draw conclusions based on the data.

We will mainly be exploring descriptive statistics in this class. To learn more about the methods of inferential statistics, you should take a course in introductory statistics.

Before we begin gathering and analyzing data we need to characterize the **population** we are studying. If we want to study the amount of money spent on textbooks by a typical first-year college student, our population might be all first-year students at your college. Or it might be:

- All first-year community college students in the state of California.
- All first-year students at public colleges and universities in the state of California.
- All first-year students at all colleges and universities in the state of California.
- All first-year students at all colleges and universities in the entire United States.
- And so on.

Population

The **population** of a study is the group the collected data is intended to describe.

Sometimes the intended population is called the **target population**, since if we design our study badly, the collected data might not actually be representative of the intended population.

Why is it important to specify the population? We might get different answers to our question as we vary the population we are studying. First-year students at Cal State Fullerton might take slightly more diverse courses than those at your college, and some of these courses may require less popular textbooks that cost more; or, on the other hand, the University Bookstore might have a larger pool of used textbooks, reducing the cost of these books to the students. Whichever the case (and it is likely that some combination of these and other factors are in play), the data we gather from your college will probably not be the same as that from Cal State Fullerton. Particularly when conveying our results to others, we want to be clear about the population we are describing with our data.

Example 7.1.1

A newspaper website contains a poll asking people their opinion on a recent news article. What is the population?

Solution

While the target (intended) population may have been all people, the real population of the survey is readers of the website.

If we were able to gather data on every member of our population, say the average (we will define "average" more carefully in a subsequent section) amount of money spent on textbooks by each first-year student at your college during the 2019-2020 academic year, the resulting number would be called a **parameter**.

Parameter

A **parameter** is a value (average, percentage, etc.) calculated using all the data from a population.

We seldom see parameters, however, since surveying an entire population is usually very time-consuming and expensive, unless the population is very small or we already have the data collected.

Census

A survey of an entire population is called a **census**.

You are probably familiar with two common censuses: the official government Census that attempts to count the population of the U.S. every ten years, and voting, which asks the opinion of all eligible voters in a district. The first of these demonstrates one additional problem with a census: the difficulty in finding and getting participation from everyone in a large population, which can bias, or skew, the results.

There are occasionally times when a census is appropriate, usually when the population is fairly small. For example, if the manager of Starbucks wanted to know the average number of hours her employees worked last week, she should be able to pull up payroll records or ask each employee directly.

Since surveying an entire population is often impractical, we usually select a **sample** to study.

Sample

A **sample** is a smaller subset of the entire population, ideally one that is fairly representative of the whole population.

We will discuss sampling methods in greater detail in a later section. For now, let us assume that samples are chosen in an appropriate manner. If we survey a sample, say 100 first-year students at your college, and find the average amount of money spent by these students on textbooks, the resulting number is called a **statistic**.

Statistic

A **statistic** is a value (average, percentage, etc.) calculated using the data from a sample.

Example 7.1.2

A researcher wanted to know how citizens of Brea felt about a voter initiative. To study this, she goes to the Brea Mall and randomly selects 200 shoppers and asks them their opinion. 60% indicate they are supportive of the initiative. What is the sample and population? Is the 60% value a parameter or a statistic?

Solution

The sample is the 200 shoppers questioned. The population is less clear. While the intended population of this survey was Brea citizens, the effective population was mall shoppers. There is no reason to assume that the 200 shoppers questioned would be representative of all Brea citizens.

The 60% value was based on the sample, so it is a statistic.

Try It 7.1.1

To determine the average length of trout in a lake, researchers catch 20 fish and measure them. What is the sample and population in this study?

Answer

The sample is the 20 fish caught. The population is all fish in the lake. The sample may be somewhat unrepresentative of the population since not all fish may be large enough to catch the bait.

Try It 7.1.2

A college reports that the average age of their students is 28 years old. Is this a statistic or a parameter?

Answer

This is a parameter, since the college would have access to data on all students (the population).

Classifying Data

Once we have gathered data, we might wish to classify it. Roughly speaking, data can be classified as **categorical data** or **quantitative data**.

Categorical and Quantitative Data

- **Categorical (qualitative) data** are pieces of information that allow us to classify the objects under investigation into various categories. They are measurements for which there is no natural numerical scale, but which consist of attributes, labels, or other non-numerical characteristics.
- **Quantitative data** are responses that are numerical in nature and with which we can perform meaningful arithmetic calculations.

Example 7.1.3

We might conduct a survey to determine the name of the favorite movie that each person in a math class saw in a movie theater. Is the data collected categorical or quantitative?

Solution

When we conduct such a survey, the responses would look like: *Top Gun: Maverick*, *Doctor Strange in the Multiverse of Madness*, or *Turning Red*. We might count the number of people who give each answer, but the answers themselves do not have any numerical values: we cannot perform computations with an answer like "*Turning Red*." This would be categorical data.

Example 7.1.4

A survey could ask the number of movies you have seen in a movie theater in the past 12 months (0, 1, 2, 3, 4, ...). Is the data collected categorical or quantitative?

Solution

This would be quantitative data since the responses are numerical. We could perform meaningful arithmetic calculations on the data such as finding the average number of movies that people saw in a movie theater in the last year.

Other examples of quantitative data would be the running time of the movie you saw most recently (131 minutes, 126 minutes, 100 minutes, ...) or the amount of money you paid for a movie ticket the last time you went to a movie theater ($10.50, $13.75, $16, ...).

Sometimes, determining whether or not data is categorical or quantitative can be a bit trickier.

Example 7.1.5

Suppose we gather respondents' ZIP codes in a survey to track their geographical location. Is the data collected categorical or quantitative?

Solution

ZIP codes are numbers, but we can't do any meaningful mathematical calculations with them (it doesn't make sense to say that 92806 is "twice" 46403 — that's like saying that Anaheim, CA is "twice" Gary, IN, which doesn't make sense at all), so ZIP codes are really categorical data.

Example 7.1.6

A survey about the movie you most recently attended includes the question "How would you rate the movie you just saw?" with these possible answers:

1 - It was awful
2 - It was just OK
3 - I liked it
4 - It was great
5 - Best movie ever!

Is the data collected categorical or quantitative?

Solution

Again, there are numbers associated with the responses, but we can't really do any calculations with them: a movie that rates a 4 is not necessarily twice as good as a movie that rates a 2, whatever that means; if two people see the movie and one of them thinks it stinks and the other thinks it's the best ever it doesn't necessarily make sense to say that "on average they liked it."

As we study movie-going habits and preferences, we shouldn't forget to specify the population under consideration. If we survey 3-7 year-olds the runaway favorite might be *Turning Red*. 13-17 year-olds might prefer *Doctor Strange*. And 33-37 year-olds might prefer *Top Gun*.

Try It 7.1.3

Classify each measurement as categorical or quantitative:

a. Eye color of a group of people
b. Daily high temperature of a city over several weeks
c. Annual income

Answer

a. Categorical
b. Quantitative
c. Quantitative