1.2 An Introduction to Statistics

The field of statistics is the art and science of collecting, presenting, analyzing, and interpreting data.

It's not just mathematical formulas and recipes. It's partly an "art" because judgment, experience, and intuition play a role.

We can make a distinction between **descriptive** and **inferential** statistics:

- **Descriptive statistics** are methods for summarizing and organizing the information in a dataset.
- **Inferential statistics** are methods for estimating and drawing conclusions from a dataset.

We need some basic ideas about datasets.

1.2.1 Terminology

Team	Captain's Gender	Wins	Rank	Winning Percentage
Dragonborn	Male	10	1	0.667
Sprites	Female	9	2	0.600
Enchanters	Female	7	3	0.467
Trolls	Male	4	4	0.267

Table 1.1: Data for an intramural league; see Table 6 on page 16 of the textbook.

In Table 1.1, the teams are the **elements**, and the row for each element is an **observation** (see Table 1.2).

Team	Captain's Gender	Wins	Rank	Winning Percentage
Sprites	Female	9	2	0.600

Table 1.2: Sprites is an element with an observation.

Captain's Gender, Wins, Rank, and Winning Percentage are the **variables** in this dataset.

1.2.2 Classifying Variables

The ways we analyze data depend on the kind of data associated with the variable. We distinguish between **qualitative** and **quantitative** variables.

Qualitative variables may be classified into categories, but their values have no numerical meaning.

Quantitative variables take numerical values, and at least some operations of arithmetic have meaning for these values.

Not all numbers are quantitative.

A driver's license number, or the number on the back of a jersey, is a qualitative variable, since it's just a "label". There would be no purpose in, say, adding or dividing such numbers.

Qualitative variables can be further classified as **discrete** or **continuous**.

The possible values of a **discrete variable** are a collection of separate points on the number line. This is often because the variable represents a count of something, so that the possible values are whole numbers, like 0, 1, or 2. Discrete variables cannot represent fractional or decimal numbers, such as 1.7 or 2/3.

The possible values of a **continuous variable** form an interval on the number line with no space between points. For example, if a quantity is measured in seconds, or in centimeters, fractional values are certainly possible.

Referring back to Table 1.1, the variables can be classified as follows:

- Captain's Gender: qualitative
- Wins: quantitative, discrete
- Rank: unknown

  Why is rank marked as unknown?

  The book calls this variable qualitative, because the difference in rank doesn't actually measure a meaningful difference between the teams. However, there are valid arguments as to it being either qualitative or quantitative.

- Winning Percentage: quantitative, continuous

  Why is winning percentage continuous?

  A winning percentage in this case could be anywhere between 0 and 1, a hallmark property of a continous variable.

1.2.3 Statistical Inference

We need to use inferential statistics when we want to draw conclusions about a **population** from a **sample**.

A **population** is the collection of all elements of interest for a particular study.

A **sample** is the subset of the population from which information is actually collected.

Example: Exercise 53

A psychologist is concerned about the health of veterans returning from war. She examines 20 veterans and assesses whether they show signs of post-traumatic stress disorder. Identify the population and the sample.

The population is the veterans returning from war.

The sample is the 20 veterans examined by the psychologist.

A characteristic of a population, such as the mean or median value of some variable for the members of that population, is called a **parameter**. A characteristic of a sample is called a **statistic**, and is what the study of statistics is named after.

Thus, we collect sample data to compute statistics when it is impractical to collect data from an entire population and compute a parameter.

The accuracy with which statistics can be used to estimate parameters is a major theme of the second half of this course.

1.3 Gathering Data

1.3.1 Sampling Methods

When choosing a sample from a population, we want to try to get a sample with similar characteristics to the population.

Ideally, we choose a **random sample**, which means that every element of the population has an equal chance of being selected.

There are sampling methods that resemble random sampling and may be appropriate in certain situations:

- **Systematic sampling:** Suppose there is a list of all elements of a population. Choose a number k with the idea that one out of every k elements will be sampled. Then choose a number in the range from 1 to k for the starting position.

Example: Systematic Sampling

Let $k = 20$, and choose 7 in the range from 1 to 20. We will put the 7^{th} element in the list in our sample, followed by the 27^{th}, 47^{th}, 67^{th}, and so on.

- **Stratified sampling:** Divide the population into subgroups of interest, such as age group or state of residence. Then take a random sample from each group.

- **Cluster sampling:** Divide the population into **clusters**, where a cluster is a collection of elements that would be convenient to sample together, such as students in the same class, or people living on the same block. Then choose some of the clusters at random and include the elements in those clusters in the sample.

A Word of Caution: Convenience Sampling

A method to avoid, to the extent that you can, is **convenience sampling**: choosing the sample that is easiest to study. A way to think about this is to compare the **target population** to the **potential population**.

The **target population** is all elements of the population of interest, and the **potential population** is the elements of the target population that could actually be sampled by the sampling method used.

When the target and potential population are different, there is **selection bias**.

Example: Exercise 27

Brandon is trying to estimate the proportion of all college students who are physically fit. He obtains a sample of students working out at the gymnasium on Monday night.

The target population is all college students, but the potential population is only the college students working out at the gym on a Monday night. This sampling method is a clear example of selection bias.

1.3.2 Experimental and Observational Studies

In these studies, the value of a **predictor variable** is used to try to explain the values of a **response variable**.

Example: Exercise 35

A sociologist is interested in whether large families (at least four children) attend religious services more often than other families do.

Predictor: the size of the family

Response: attendance at religious services

In an **experimental study**, the researchers can control the value of the predictor variable.

In an **observational study**, the researchers can only observe the value of the predictor variable. With the given example of family size and worship, and for many other areas of study in real life, it is often impractical or unethical to go beyond observational studies.

In experimental studies, there are often **treatment** and **control** groups. The researchers make some purposeful intervention for the treatment group, such as administering the drug under study, but not the control group, which will often be administered a placebo drug.

In observational studies, there can be **case** and **control** groups, where the elements of the case group share some characteristic the researchers wish to study, and the elements of the control group do not have this characteristic.

1.3.3 Important Experimental Concepts

Randomization is the practice of randomly assigning participants in an experiment to either the treatment or the control group. This is done to minimize the chance that the two groups will be significantly different from each other in other ways.

Replication is the practice of including enough participants in each group of the study so that meaningful differences between the groups can be detected.