CHAPTER 1 Introduction to Data

Chapter Outline

1.1 Why Study Statistics?

Learning Objectives

- Recognize what we mean by the term statistics.
- Become familiar with several examples of how statistics are used in the real world.

What is Statistics?

Statistics is one of the most honestly useful math topics you are likely to study. Nearly every kind of occupation and human activity can benefit from an application of statistics. In the most general sense, statistics describes a set of tools and techniques that can be used to describe, organize, and interpret information or data. What are those data? They could be the scores of students on their final exam in history, the speed with which a new drug relieves headaches, the number of complaints received by customers, the free-throw percentage for different college basketball players, or the average price of going out for pizza on a Friday night.

Statistics is a tool that helps us understand the world around us and make better decisions with the information we have available to us. By knowing which products sell better at a particular time of year, for example, a business can make the best use of product placement and advertising. Knowing what time(s) of the day, week, or year are the busiest, a restaurant manager can efficiently schedule her employees so as not to waste labor costs. Coaches that know the statistics on their players can use that information to help them field stronger offensives or defensives.

Applications

One type of business that makes extensive use of statistics is insurance sales. Insurance companies are just like most companies from the standpoint that they are in business to make a profit for their investors. That does not mean that buying insurance is a bad idea for individual people, or that the companies deliberately overcharge their customers, but it does mean that the companies are very careful to charge enough for each policy to 'insure' that the company makes money overall.

How can the companies know for certain how many people are going to make claims against their insurance policies? Or how big their claims will be? They can't know for *certain* since they don't have a way to see the future, but they can get a very reliable idea of the average number of claims from a specific population of people through the use of **sample** groups and the application of probability and statistics.

Example A

Predicting the weather is a tricky job. There are a nearly infinite number of possible **variables** that can affect the temperature and chance of precipitation for any given day. Of course, a weatherman cannot possibly take *all* of these variables into account every time he/she makes a prediction, so he/she must identify the most influential variables and just watch them closely for each prediction. Suppose that according to records, it has rained an average of 5 days during the month of April for each year over the last 15 years. If it is currently April 25^{th} and there has been no rain, should the weatherman warn everyone to bring an umbrella to work for the next five days?

Maybe. However, the information regarding the average number of rainy days in April over the last 15 years probably won't have much to do with it. Although the history may be suggestive of a particular number of rainy days, it is certainly no guarantee of a specific result. If the weather conditions such as temperature, cloud cover, relative humidity, etc., are all conducive to rain, then he/she is likely to predict rain, but the fact that there are only five days left is certainly no assurance that there *must* be rain all five final days so that the average will be fulfilled.

Example B

Suppose a car insurance company reviews the police records for thousands of speeding tickets and minor car accidents over a ten-year period, and notes the following:

TABLE 1.1:

	Speeding Tickets	**"Fender Benders"**
Boys ages 16 - 23	4,532	1,725
Girls ages 16 - 28	1,242	1,715

Would it make sense for the company to charge the same rates for boys and girls?

It certainly does not look like it. According to the statistics, boys are nearly four times as likely to drive over the speed limit, and although there were slightly fewer recorded accidents for girls than boys, note that the age range for the girls was greater than for the boys. The greater age range suggests that there may have been more girls actually driving than boys, yet they ended up in nearly the same number of accidents!

However, it is extremely important to note that without data regarding the actual number of boys and girls in each group, we can't really get a good feel for the overall increased likelihood of boys making claims.

Lesson Summary

Statistics is about how to think clearly about data. There is no question that a little practice in learning how to think statistically will help you see the world more clearly and accurately – at least when it comes to making sense of the data that surrounds us. Our objectives in this course are to help you identify the kinds of questions that can be answered by statistics, to show you the tools you can use to organize and summarize your data, and to help you practice the most important skill of all: to clearly interpret what those data are saying.

1.2 Classifying Variables

Learning Objectives

- Distinguish between quantitative and categorical variables.
- Understand the concept of a population and the reason for using a sample.
- Distinguish between a statistic and a parameter.

Introduction

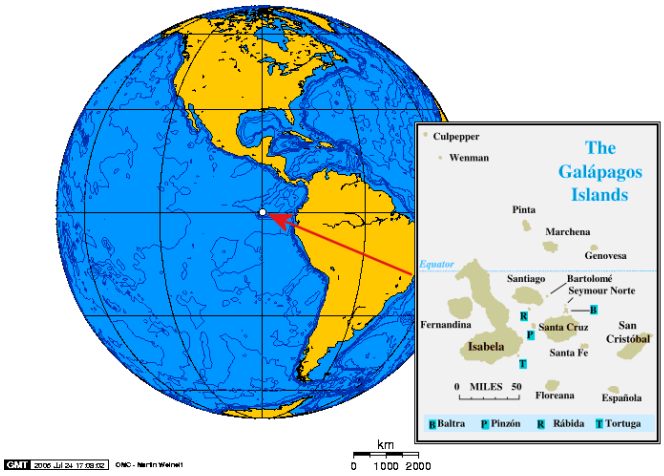
Data in its original form, just a list of numbers, names, letters, colors, etc., is known as **raw data**, and is often not particularly useful without some kind of organization. Without some sort of context and some level of organization, data can seem like just a bunch of meaningless values.

Data can be classified into two general types, **quantitative** and **qualitative**. There are a number of ways to group or organize each type of data to make it more useful.

In this lesson, you will be introduced to some basic vocabulary of statistics and learn how to distinguish between different types of variables. We will use the real-world example of information about the Giant Galapagos Tortoise.

The Galapagos Tortoises

The Galapagos Islands, off the coast of Ecuador in South America, are famous for the amazing diversity and uniqueness of life they possess. One of the most famous Galapagos residents is the Galapagos Giant Tortoise, which is found nowhere else on earth. Charles Darwin's visit to the islands in the 19^{th} Century and his observations of the tortoises were extremely important in the development of his theory of evolution.



The tortoises lived on nine of the Galapagos Islands, and each island developed its own unique species of tortoise. In fact, on the largest island, there are four volcanoes, and each volcano has its own species. When first discovered, it was estimated that the tortoise population of the islands was around 250,000. Unfortunately, once European ships and settlers started arriving, those numbers began to plummet. Because the tortoises could survive for long periods of time without food or water, expeditions would stop at the islands and take the tortoises to sustain their crews with fresh meat and other supplies for the long voyages. Also, settlers brought in domesticated animals like goats and pigs that destroyed the tortoises' habitat. Today, two of the islands have lost their species, a third island has no remaining tortoises in the wild, and the total tortoise population is estimated to be around 15,000. The good news is there have been massive efforts to protect the tortoises. Extensive programs to eliminate the threats to their habitat, as well as breed and reintroduce populations into the wild, have shown some promise.

TABLE 1.2:

Island or Volcano	Species	Climate Type	Shell Shape	Estimate of Total Population	Population Density (per km^2)	Number of Individuals Repatriated*
Wolf	becki	semi-arid	intermediate	1139	228	40
Darwin	microphyes	semi-arid	dome	818	205	0
Alcedo	vanden-burghi	humid	dome	6,320	799	0
Sierra Negra	guntheri	humid	flat	694	122	286
Cerro Azul	vicina	humid	dome	2.574	155	357
Santa Cruz	nigrita	humid	dome	3,391	730	210
Española	hoodensis	arid	saddle	869	200	1,293
San Cristóbal	chathamen-sis	semi-arid	dome	1,824	559	55
Santiago	darwini	humid	intermediate	1,165	124	498
Pinzón	ephippium	arid	saddle	532	134	552
Pinta	abingdoni	arid	saddle	1	Does not apply	0

**Repatriation is the process of raising tortoises and releasing them into the wild when they are grown to avoid local predators that prey on the hatchlings.*

Classifying Variables

Statisticians refer to an entire group that is being studied as a population. Each member of the population is called a **unit**, or **subject**. In this example, the population is all Galapagos Tortoises, and the units are the individual tortoises. It is not necessary for a population or the units to be living things, like tortoises or people. For example, an airline employee could be studying the population of jet planes in her company by studying individual planes.

A researcher studying Galapagos Tortoises would be interested in collecting information about different characteristics of the tortoises. Those characteristics are called **variables**. Each column of the previous figure contains a variable. In the first column, the tortoises are labeled according to the island (or volcano) where they live, and in the second column, by the scientific name for their species. When a characteristic can be neatly placed into well-defined groups, or categories, that do not depend on order, it is called a **categorical variable**, or **qualitative variable**.

The last three columns of the previous figure provide information in which the count, or quantity, of the characteristic is most important. For example, we are interested in the total number of each species of tortoise, or how many individuals there are per square kilometer. This type of variable is called a **numerical variable**, or **quantitative variable**. The figure below explains the remaining variables in the previous figure and labels them as categorical or numerical.

TABLE 1.3:

Variable	Explanation	Type
Climate Type	Many of the islands and volcanic habitats have three distinct climate types.	Categorical
Shell Shape	Over many years, the different species of tortoises have developed different shaped shells as an adaptation to assist them in eating vegetation that varies in height from island to island.	Categorical
Number of Individuals Repatriated	There are two tortoise breeding centers on the islands. Through these programs, many tortoises have been raised and then reintroduced into the wild.	Numerical

Population vs. Sample

We have already defined a population as the total group being studied. Most of the time, it is extremely difficult or very costly to collect all the information about a population. In the Galapagos, it would be very difficult and perhaps even destructive to search every square meter of the habitat to be sure that you counted every tortoise. In an example closer to home, it is very expensive to get accurate and complete information about all the residents of the United States to help effectively address the needs of a changing population. This is why a complete counting, or **census**, is only attempted every ten years. Because of these problems, it is common to use a smaller, representative group from the population, called a **sample**.

Errors in Sampling

We have to accept that estimates derived from using a sample have a chance of being inaccurate. This cannot be avoided unless we measure the entire population. The researcher has to accept that there could be variations in the sample due to chance that lead to changes in the population estimate. A statistician would report the estimate of the parameter in two ways: as a **point estimate** (e.g., 915) and also as an **interval estimate**. For example, a statistician would report: "I am 95% confident that the true number of tortoises is actually between 561 and 1075." This range of values is the unavoidable result of using a sample, and not due to some mistake that was made in the process of collecting and analyzing the sample. The difference between the true parameter and the statistic obtained by sampling is called **sampling error**. It is also possible that the researcher made mistakes in her sampling methods in a way that led to a sample that does not accurately represent the true population. For example, she could have picked an area to search for tortoises where a large number tend to congregate (near a food or water source, perhaps). If this sample were used to estimate the number of tortoises in all locations, it may lead to a population estimate that is too high. This type of systematic error in sampling is called **bias**. Statisticians go to great lengths to avoid the many potential sources of bias. We will investigate this in more detail in a later chapter.

Lesson Summary

In statistics, the total group being studied is called the population. The individuals (people, animals, or things) in the population are called units. The characteristics of those individuals of interest to us are called variables. Those variables are of two types: numerical, or quantitative, and categorical, or qualitative.

Because of the difficulties of obtaining information about all units in a population, it is common to use a small, representative subset of the population, called a sample. An actual value of a population variable (for example, number of tortoises, average weight of all tortoises, etc.) is called a parameter. An estimate of a parameter derived from a sample is called a statistic.

Whenever a sample is used instead of the entire population, we have to accept that our results are merely estimates, and therefore, have some chance of being incorrect. This is called sampling error.

Points to Consider

- How do we summarize, display, and compare categorical and numerical data differently?
- What are the best ways to display categorical and numerical data?
- Is it possible for a variable to be considered both categorical and numerical?
- How can you compare the effects of one categorical variable on another or one quantitative variable on another?

Review Questions

1. In each of the following situations, identify the population, the units, and each variable, and tell if the variable is categorical or quantitative.

1. (a) A quality control worker with Sweet-Tooth Candy weighs every 100th candy bar to make sure it is very close to the published weight.
   (b) Doris decides to clean her sock drawer out and sorts her socks into piles by color.
   (c) A researcher is studying the effect of a new drug treatment for diabetes patients. She performs an experiment on 200 randomly chosen individuals with type II diabetes. Because she believes that men and women may respond differently, she records each person's gender, as well as the person's change in blood sugar level after taking the drug for a month.
2. In Physical Education class, the teacher has the students count off by two's to divide them into teams. Is this a categorical or quantitative variable?
3. A school is studying its students' test scores by grade. Explain how the characteristic 'grade' could be considered either a categorical or a numerical variable.

1.3 Levels of Measurement

Learning Objective

- Understand the difference between the levels of measurement: nominal, ordinal, interval, and ratio.

Introduction

This lesson is an overview of the basic considerations involved with collecting and analyzing data.

Levels of Measurement

In the first lesson, you learned about the different types of variables that statisticians use to describe the characteristics of a population. Some researchers and social scientists use a more detailed distinction, called the **levels of measurement**, when examining the information that is collected for a variable. This widely accepted (though not universally used) theory was first proposed by the American psychologist Stanley Smith Stevens in 1946. According to Stevens' theory, the four levels of measurement are nominal, ordinal, interval, and ratio.

Each of these four levels refers to the relationship between the values of the variable.

Nominal measurement

A **nominal measurement** is one in which the values of the variable are names. The names of the different species of Galapagos tortoises are an example of a nominal measurement.

Ordinal measurement

An ***ordinal measurement*** involves collecting information of which the order is somehow significant. The name of this level is derived from the use of ordinal numbers for ranking (1^{st}, 2^{nd}, 3^{rd} etc.). If we measured the different species of tortoise from the largest population to the smallest, this would be an example of ordinal measurement. In ordinal measurement, the distance between two consecutive values does not have meaning. The 1^{st} and 2^{nd} largest tortoise populations by species may differ by a few thousand individuals, while the 7^{th} and 8^{th} may only differ by a few hundred.

Interval measurement

With **interval measurement**, the distance between any two values has a specific meaning. An example commonly cited for interval measurement is temperature (either degrees Celsius or degrees Fahrenheit). A change of 1 degree is the same if the temperature goes from 0° C to 1° C as it is when the temperature goes from 40° C to 41° C. In addition, there is meaning to the values between the ordinal numbers. That is, a half of a degree has meaning.

Ratio measurement

A **ratio measurement** is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind. A variable measured at this level not only includes the concepts of order and interval, but also adds the idea of 'nothingness', or absolute zero. With the temperature scale of the previous example, 0° C is really an arbitrarily chosen number (the temperature at which water freezes) and does not represent the absence of temperature. As a result, the ratio between temperatures is relative, and 40° C, for example, is not twice as hot as 20° C. On the other hand, for the Galapagos tortoises, the idea of a species having a population of 0 individuals is all too real! As a result, the estimates of the populations are measured on a ratio level, and a species with a population of about 3,300 really is approximately three times as large as one with a population near 1,100.

Comparing the Levels of Measurement

Using Stevens' theory can help make distinctions in the type of data that the numerical/categorical classification could not. Let's use an example from the previous section to help show how you could collect data at different levels of measurement from the same population. Assume your school wants to collect data about all the students in the school.

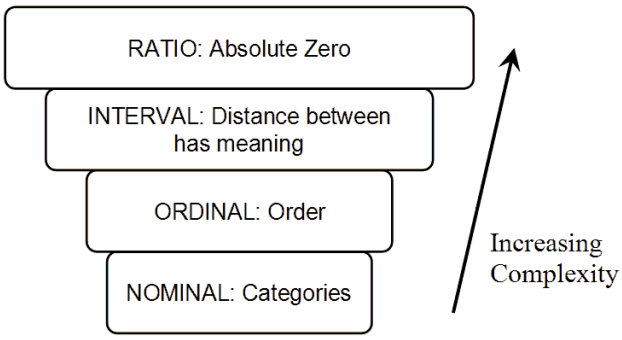
If we collect information about the students' gender, race, political opinions, or the town or sub-division in which they live, we have a nominal measurement.

If we collect data about the students' year in school, we are now ordering that data numerically (9^{th}, 10^{th}, 11^{th}, or 12^{th} grade), and thus, we have an ordinal measurement.

If we gather data for students' SAT math scores, we have an interval measurement. There is no absolute 0, as SAT scores are scaled. The ratio between two scores is also meaningless. A student who scored a 600 did not necessarily do twice as well as a student who scored a 300.

Data collected on a student's age, height, weight, and grades will be measured on the ratio level, so we have a ratio measurement. In each of these cases, there is an absolute zero that has real meaning. Someone who is 18 years old is twice as old as a 9-year-old.

It is also helpful to think of the levels of measurement as building in complexity, from the most basic (nominal) to the most complex (ratio). Each higher level of measurement includes aspects of those before it. The diagram below is a useful way to visualize the different levels of measurement.



Lesson Summary

Data can be measured at different levels, depending on the type of variable and the amount of detail that is collected. A widely used method for categorizing the different types of measurement breaks them down into four groups. Nominal data is measured by classification or categories. Ordinal data uses numerical categories that convey a

meaningful order. Interval measurements show order, and the spaces between the values also have significant meaning. In ratio measurement, the ratio between any two values has meaning, because the data include an absolute zero value.

Point to Consider

- How do we summarize, display, and compare data measured at different levels?

Review Questions

1. In each of the following situations, identify the level(s) at which each of these measurements has been collected.
   a. Lois surveys her classmates about their eating preferences by asking them to rank a list of foods from least favorite to most favorite.
   b. Lois collects similar data, but asks each student what her favorite thing to eat is.
   c. In math class, Noam collects data on the Celsius temperature of his cup of coffee over a period of several minutes.
   d. Noam collects the same data, only this time using degrees Kelvin.
2. Which of the following statements is not true.

   (a) All ordinal measurements are also nominal.

   (b) All interval measurements are also ordinal.

   (c) All ratio measurements are also interval.

   (d) Steven's levels of measurement is the one theory of measurement that all researchers agree on.

3. Look at **Table 1.2** in Section 1. What is the highest level of measurement that could be correctly applied to the variable 'Population Density'?

   (a) Nominal

   (b) Ordinal

   (c) Interval

   (d) Ratio

NOTE: If you are curious about the "*does not apply*" in the last row of **Table 1.2,** read on! There is only one known individual Pinta tortoise, and he lives at the Charles Darwin Research station. He is affectionately known as Lonesome George. He is probably well over 100 years old and will most likely signal the end of the species, as attempts to breed have been unsuccessful.