Chapter 1

Introduction to Data Analysis

We will start our discussion of statistics with some of the important words and vocabulary involved in analyzing statistical data. We will look at different ways to measure the center of a data set, as well as measure the spread or variation of a data set. You will also learn some other ways of summarizing data such as using percentiles.

Statistics plays a major role in many aspects of our lives. We use statistics in sports to compare players' abilities. Politicians use statistics to better understand public opinion. Biomedical researchers use statistics to help determine the effectiveness of new drugs.

Statistics is very useful in providing a better understanding of the world around us, when used appropriately. Unfortunately, statistics are often used inappropriately and lead to inaccurate conclusions and decisions.

1.1 Some Basic Concepts

Before we start analyzing data we need to establish some definitions and concepts.

Definition 1.1.1. Statistics is the science of collecting, organizing, summarizing, and analyzing data. Moreover, statistics is about providing a measure of confidence in any decision or conclusion.

From the definition we note that there are four key parts to statistics. Data collection is a critical step in that we want to make an inference about a population based on a sample of data we collect. Organizing and cleaning the data is necessary so that we can move forward with our analysis. Summarizing the data through exploratory data analysis is our first step in better understanding the relations in the data. We make use of descriptive statistics and graphs to learn about the relationships. Finally, we delve into the data and conduct our analysis: build models and interpret the results.

In statistical analysis we work with a sample and make inferences about the population. What is the difference between a sample and the population?

Definition 1.1.2. The entire group of individuals being studied is the **population**. An **individual** is a member of the population. A **sample** is a subset of the population.

1.2 Scientific Method

In order to conduct research it is important to follow certain steps, this is know as the *scientific method*. Broadly speaking, the scientific method is as follows:

Ask a Question: The scientific method starts when you ask a question about something that you observe: How, What, When, Who, Which, Why, or Where? The researcher must formulate the question he or she hopes to answer through the research.

Do Background Research: This is often referred to as literature search. Find articles and books outlining similar work in the field to help you find the best way to do things and insure that you do not repeat mistakes from the past.

Construct a Hypothesis: A hypothesis is an educated guess about how things work. It is an attempt to answer your question with an explanation that can be tested. State both your hypothesis and the resulting prediction you will be testing. Predictions must be easy to measure.

Test Your Hypothesis by Doing an Experiment: Your experiment tests whether your prediction is accurate and thus your hypothesis is supported or not. It is important for your experiment to be a fair test. You conduct a fair test by making sure that you change only one factor at a time while keeping all other conditions the same. You should also repeat your experiments several times to make sure that the first results weren't just an accident.

Analyze Your Data and Draw a Conclusion: Once your experiment is complete, you collect your measurements and analyze them to see if they support your hypothesis or not. You may find that your predictions were not accurate and your. In such cases go back and construct a new hypothesis and prediction based on the what you learned during your experiment. This starts much of the process of the scientific method over again.

Communicate Your Results: You complete your research by communicating your results to others in a paper and/or a poster.

1.3 Data and Measurement Issues

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.

1.3.1 Classifying Variables

Statisticians refer to an entire group that is being studied as a population. Each member of the population is called a unit. Suppose we want to study different types of salmon. In this example, the population is all the salmon in the rivers and streams, and the units are the individual fish.

A researcher studying salmon would be interested in collecting information about different characteristics of salmon. Those characteristics are called **variables**. These variables include the species, length, weight, and age, to name a few. 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. Variables like weight and length are numerical or quantitative.

1.3.2 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 salmon example, it would be impossible to collect data on every salmon. 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.

If you wanted to find an estimate for the average weight of the population of salmon, you would go into the field and locate and weigh a number of salmon. You would then use statistical techniques that we will discuss later to obtain an estimate for the average weight in the population. In statistics, we call the actual average weight of salmon a parameter. Any number that describes the individuals in a sample (length, weight, age) is called a statistic. Each statistic is an estimate of a parameter, whose value may or may not be known.

1.3.3 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., 76) and also as an interval estimate. For example, a statistician would report: "I am fairly confident that the average weight of salmon is between 50 cm and 120." 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.

1.3.4 Determining Errors That May Have Occurred

What are some possible errors that could be involved in the study of salmon?

The researcher could have picked an area to search for salmon where a large number tend to congregate (near a food or water source, perhaps). If this sample were used to estimate the number of salmon 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.

Exercises

For the following, identify the population, the unit of measure, and the type of variable (categorical or quantitative):

1. A quality control worker weighs every $10th$ candy bar to make sure it is close to the published weight.

2. A girl sorts her socks into piles by color.

3. 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 and records each persons change in blood sugar level after taking the drug for a month.

For the following, indicate for each of the following characteristics of an individual whether the variable is categorical or quantitative:

4. Length of an arm from elbow to shoulder (in inches)
5. The number of DVDs the person owns.
6. A boy's feeling about his own height (too tall, too short, about right)

1.4 Levels of Measurement

Learning Objects:

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

Some researchers and social scientists use a more detailed distinction of measurement, 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.

Definition 1.4.1. Nominal measurement: A nominal measurement is one in which the values of the variable are names.

Definition 1.4.2. 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 , etc.

Examples of Nominal and Ordinal Measurements The names of the different species of salmon are an example of a nominal measurement. If we measured the different species of salmon 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.

1.4.1 Interval measurement

With interval measurement, there is significance to the distance between any two values.

1.4.2 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.

Examples of Interval and Ratio Measurement We can use examples of temperature for these.

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 10°C to 11°C as it is when the temperature goes from 20°C to 21°C. In addition, there is meaning to the values between the ordinal numbers. That is, a half of a degree has meaning.

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 20°C, for example, is not twice as hot as 10°C. On the other hand, for salmon, the idea of a species having a population of 0 individuals is possible. As a result, the estimates of the populations are measured on a ratio level, and a species with a population of 3,300 really is three times as large as one with a population of 1,100.

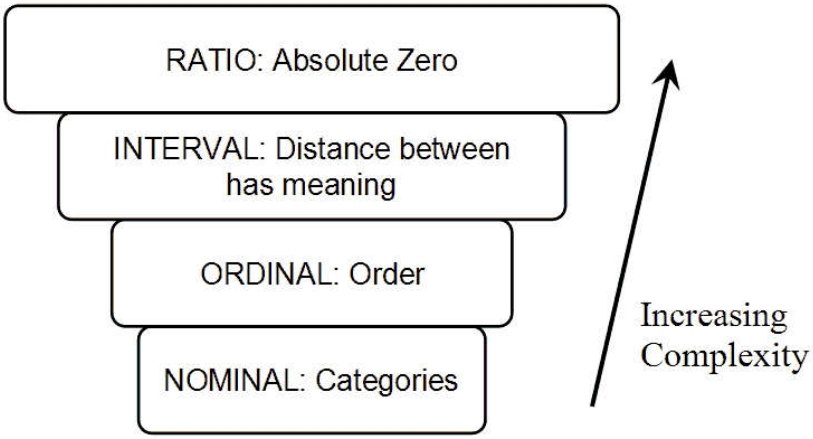
1.4.3 Determining Levels of Measurement

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, or 10 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 students 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.



Exercises

Identify the level at which each of these measurements has been collected.

1. Lois surveys her classmates about their eating preferences by asking them to rank a list of foods from least favorite to most favorite.
2. In a math class, Jack collects data on the Celsius temperature of his cup of coffee over a period of several minutes.
3. Sam collects the same data, only this time using degrees Kelvin.

Explain whether or not the following statements are true.

4. All ordinal measurements are also nominal.
5. All interval measurements are also ordinal.
6. All ratio measurements are also interval.

1.5 Measures of Central Tendency

Learning Objectives:

- Calculate the mode, median, and mean for a set of data, and understand the differences between each measure of center.
- Identify the symbols and know the formulas for sample and population means.
- Determine the values in a data set that are outliers.
- Identify the values to be removed from a data set for a trimmed mean.

Three commonly used measures of center are the mode, the median, and the mean. For example let's look at data collected representing the number of cars each family has in a neighborhood. Say the data are:

$$2, 1, 3, 2, 2, 1, 1, 4, 2, 3, 2, 1, 3, 2, 2, 2.$$

1.5.1 Mode

The mode is defined as the most frequently occurring number in a data set. The mode is most useful in situations that involve categorical (qualitative) data that are measured at the nominal level. For the cars data set, 2 is the mode, as it is the most frequently occurring number of cars in the sample.

If there were seven 3-car households and seven 2-car households, we would say the data set has two modes. In other words, the data would be bimodal. When a data set is described as being bimodal, it is clustered about two different modes. Technically, if there were more than two, they would all be the mode. However, the more of them there are, the more trivial the mode becomes. In these cases, we would most likely search for a different statistic to describe the center of such data.

If there is an equal number of each data value, the mode is not useful in helping us understand the data, and thus, we say the data set has no mode.

1.5.2 Mean

Another measure of central tendency is the arithmetic average, or mean. This value is calculated by adding all the data values and dividing the sum by the total number of data points. The mean is the numerical balancing point of the data set.

Statisticians use the symbol $\overline{x}$ read as "x bar" to represent the sample mean and x_i is the symbol for a single measurement.

Symbolically, the formula for the sample mean is as follows:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where $x_i$is the value of the i-th observation and n is the sample size. The mean of the population is denoted by the Greek letter, μ. $\overline{x}$, the sample mean is a statistic, and μ the population mean is a parameter.

1.5.3 Median

The median is simply the middle number in an ordered set of data. Suppose a student took five statistics quizzes and received the following grades:

$$80, 94, 75, 96, 90.$$

To find the median, you first put the data in order. The median will be the data point that is in the middle. Placing the data in order from least to greatest yields: 75, 80, 90, 94, 96. The middle number in this case is the third grade, or 90, so the median of this data is 90.

When there is an even number of numbers, no one of the data points will be in the middle. In this case, we take the average (mean) of the two middle numbers. In general, sort the data

$$x_1 \leq x_2 \leq \cdots x_n.$$

If n is odd

$$m = x_{(n+1)/2}.$$

If n is even

$$m = \frac{x_{(n/2)} + x_{(n+2)/2}}{2}'$$

1.5.4 Mean vs. Median

Both the mean and the median are important and widely used measures of center. Consider the following example: Suppose you got an 85 and a 93 on your first two statistics quizzes, but then you had a really bad day and got a 14 on your next quiz!

The mean of your three grades would be 64. Which is a better measure of your performance? As you can see, the middle number in the set is an 85. That middle does not change if the lowest grade is an 84, or if the lowest grade is a 14. However, when you add the three numbers to find the mean, the sum will be much smaller if the lowest grade is a 14.

1.5.5 Outliers and Resistance

The mean and the median are so different in this example because there is one grade that is extremely different from the rest of the data. In statistics, we call such extreme values outliers. The mean is affected by the presence of an outlier; however, the median is not. A statistic that is not affected by outliers is called resistant. We say that the median is a resistant measure of center, and the mean is not resistant. In a sense, the median is able to resist the pull of a far away value, but the mean is drawn to such values. It cannot resist the influence of outlier values. As a result, when we have a data set that contains an outlier, it is often better to use the median to describe the center, rather than the mean.

1.5.6 Determining the Most Appropriate Measure of Center

In 2005, the CEO of Yahoo, Terry Semel, was paid almost $231,000,000. This is certainly not typical of what the average worker at Yahoo could expect to make. Instead of using the mean salary to describe how Yahoo pays its employees, it would be more appropriate to use the median salary of all the employees.

You will often see medians used to describe the typical value of houses in a given area, as the presence of a very few extremely large and expensive homes could make the mean appear misleadingly large.

1.6 Summary Statistics, Summarizing Univariate Distributions

Learning Objectives:

- Find the minimum and maximum values.
- Calculate other types of means such as trimmed means and weighted means.
- Find percentiles.
- Find quartiles.

1.6.1 More Measures of Center

The mean, median and mode are only a few possible measures of center. While they are the most commonly used measures of center, it is important to be familiar with some other measures of center that are sometimes used as well.

Trimmed Mean

Recall that the mean is not resistant to the effects of outliers. Many students ask their teacher to "drop the lowest grade." The argument is that everyone has a bad day, and one extreme grade that is not typical of the rest of their work should not have such a strong influence on their mean grade. The problem is that this can work both ways; it could also be true that a student who is performing poorly most of the time could have a really good day (or even get lucky) and get one extremely high grade. We wouldnt blame this student for not asking the teacher to drop the highest grade! Attempting to more accurately describe a data set by removing the extreme values is referred to as trimming the data. To be fair, though, a valid trimmed statistic must remove both the extreme maximum and minimum values. So, while some students might disapprove, to calculate a trimmed mean you remove the maximum and minimum values and divide by the number of values that remain.

Consider the following quiz grades: 75, 80, 90, 94, 96. A trimmed mean would remove the largest and smallest values, 75 and 96, and divide by 3.

$n\%$ Trimmed Mean

Instead of removing just the minimum and maximums in a larger data set, a statistician may choose to remove a certain percentage of the extreme values. This is called an trimmed mean. To perform this calculation, remove the specified percent of the number of values from the data from each end. For example, in a data set that contains 100 numbers, to calculate a 10% trimmed mean, remove 10% of the data from each end. In this simplified example, the ten smallest and the ten largest values would be discarded, and the sum of the remaining numbers would be divided by 80.

Weighted Mean

The weighted mean is a method of calculating the mean where instead of each data point contributing equally to the mean, some data points contribute more than others. This could be because they appear more often or because a decision was made to increase their importance (give them more weight). The most common type of weight to use is the frequency, which is the number of times each number is observed in the data. The calculation would look like this:

The symbolic representation of this is as follows:

$$\overline{x} = \frac{x_1 w_1 + x_2 w_2 + \cdots + x_n w_n}{n} = \frac{\sum_{i=1}^{n} x_i w_i}{n}$$

where $x_i$is the value of the i-th observation, w_i is the number of times that data point occurs (or the weight for that point), and n is the sample size.

We may be interested in other sections of the data besides the center or middle. We could be interested in some lower percentage of the data or some higher portion of the data. The following topics will explain how to look at certain portions or percentages of a data set.

1.6.2 Percentiles and Quartiles

A percentile is a statistic that identifies the percentage of the data that is less than the given value. The most commonly used percentile is the median. Because it is in the numeric middle of the data, half of the data is below the median, and half is above. Therefore, we could also call the median the 50-th percentile. A 40-th percentile would be a value in which 40% of the numbers are less than that observation.

To check a childs physical development, pediatricians use height and weight charts that help them to know how the child compares to children of the same age. A child whose height is in the percentile is taller than 70% of children of the same age.

Two very commonly used percentiles are the 25-th and 75-th percentiles. These are also known as the first quartile or lower quartile ($Q1$) and the third quartile or upper quartile ($Q3$). The median is a middle quartile and is sometimes referred to as $Q2$.

1.7 Measures of Spread and Dispersion

Learning Objectives:

- Calculate the range and interquartile range.
- Calculate the standard deviation for a population and a sample, and understand its meaning.
- Distinguish between the variance and the standard deviation.
- Calculate and apply Chebyshevs Theorem to any set of data.

Another important feature that can help us understand more about a data set is the manner in which the data are distributed, or spread. Variation and dispersion are words that are also commonly used to describe this feature. There are several commonly used statistical measures of spread that we will investigate in this lesson.

1.7.1 Range

One measure of spread is the range. The range is simply the difference between the largest value (maximum) and the smallest value (minimum) in the data. Take the data below:

$$75, 80, 90, 94, 96$$

The range of this data set is $96 - 75 = 21$. This is telling us the distance between the maximum and minimum values in the data set.

The range is useful because it requires very little calculation, and therefore, gives a quick and easy snapshot of how the data are spread. However, it is limited, because it only involves two values in the data set, and it is not resistant to outliers.

1.7.2 Interquartile Range

The interquartile range is the difference between the $Q1$ and $Q2$, and it is abbreviated IQR. The IQR gives information about how the middle 50% of the data are spread. Fifty percent of the data values are always between $Q1$ and $Q2$.