2020

Clear-Sighted Statistics: Module 2: Types of Data

Edward Volchok
CUNY Queensborough Community College

Clear-Sighted Statistics: An OER Textbook

Module 2: Types of Data

"In God we trust, all others bring data."[1]

-- W. Edwards Deming

I. Introduction

In this module, we will discuss what data are. (**Remember:** data is a plural noun; the singular of data is datum.) In Module 3: Where Do Data Come From?, we will review how to obtain trustworthy data.

After completing this module, you will understand:

- The difference between a sample and a population.
- The difference between statistics and parameters.
- The difference between facts and value.
- Sampling error and uncertainty.
- The different types of data: Qualitative data, quantitative data, discrete quantitative data, and continuous quantitative data.
- The difference among univariate, bivariate, and multivariate data.
- The four levels of measurement: Nominal, Ordinal, Interval, and Ratio and why this distinction is important.

II. The Difference Between Populations and Samples

Data are individual facts that people using statistical techniques process to acquire usable information. In Module 1, we defined the science of statistics as the science involved with the collection, organization, analysis, interpretation, and presentation of data. Data can come from either a population or a sample. A *population*, or universe, is a collection of all possible individuals, objects, or measurements of interest. Counting elements in a

population is called a census. Article 1, Section 2 of the Constitution of the United States mandates that the federal government conduct a census of the people living in the country every ten years.[2] Data collected from populations are called *parameters*. A *sample* is a portion or part of a population. Data derived from samples are called statistics.

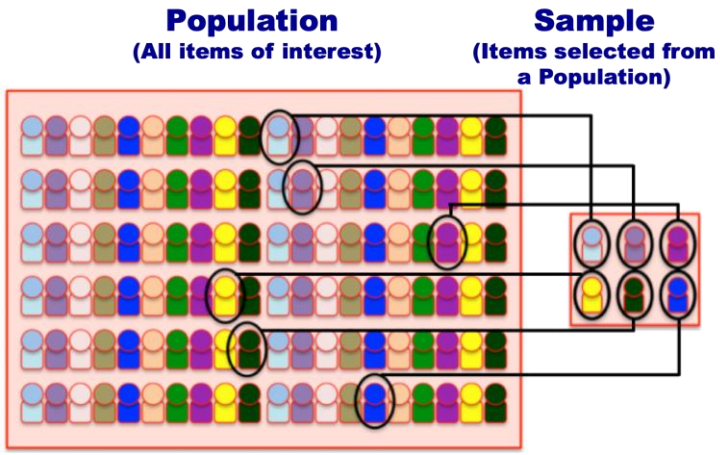

Figure 1: Sample vs. Population

In an ideal world, we would only use parameters. But neither you nor I live in an ideal world. We seldom have the time or money to study every element of a population. Sometimes it is difficult to do so, particularly when attempting to count constantly changing or large populations like salmon swimming between Canada and Greenland. Sometimes we destroy or damage the items we select for our sample. Cars are tested for crash worthiness. The price of a 2019 Ferrari SF90 Stradale is around $600,000. How many of these very expensive automobiles will Ferrari crash while testing their safety? The makers of Ferrari, therefore, crash a very small sample from their production. For these reasons, we use sample statistics to estimate population parameters. We will review the problem of estimating population parameters from sample statistics in Modules 10 and 11 and in the remaining modules covering inferential statistics. When properly conducted, data derived from samples are reliable.

Inferential statistics is the use of sample data and probability theory to estimate population parameters. With inferential statistics, we always expect sampling error. In later modules, we will discover that sampling error is not the result of someone making a mistake. Sampling error is due to the random process of sampling. **Please note:** "Random" in this context means that elements of the population are selected by chance with each element having an equal likelihood of selection. Sampling error simply means that the value of the statistic does not equal the value of the parameter. Sampling error is, in fact, unavoidable. In Module 3: Where Do Data Come From? we will highlight the problem of systematic error, which results from biases introduced by poor research design, human error, fraudulent research practices, and unconscious or deliberate distortions introduced by research respondents.

III. The Difference Between Facts and Values

Statistics is based on empirical knowledge; that is knowledge obtained through the observation of the world. With the use of statistics and the scientific method we can uncover facts. *Facts*, or empirical evidence, are repeated observations or measurements. Facts are objective and verifiable. Facts deal with "what is." Max Weber, the great twentieth century sociologist, drew a distinction between fact and values.[3] Values deal with "what ought to be." Values are subjective beliefs that cannot be verified through the use of statistics and the scientific method. Values can be culturally determined.

Here are two examples of values regarding strawberries:

- I think strawberries are the best berries made by God.
- I think strawberries taste better than cherries.

These values are not facts. I cannot use statistics to prove wrong your preference for cherries over strawberries. That you prefer cherries to strawberries is a statement about your values, not fact.

Compare the value statements shown above to some facts about strawberries:

- The average strawberry has 200 seeds.
- Strawberries are among the first fruits to ripen in the spring.
- According to the U.S. Department of Agriculture, the average American eats 3.4 pounds of strawberries a year.
- According to the U.S. Department of Agriculture, California produces 72 percent of all the strawberries grown in the U.S.

These facts are derived from empirical observations and can be verified using statistics. I should point out that social scientists use statistics to study peoples' values and beliefs. They may ask questions like:

- What proportion of the population likes or dislikes strawberries?
- How does the ranking of the taste of strawberries compare to cherries and other fruits?
- What is the intensity of people's feelings about strawberries?
- What are the cultural meanings associated with strawberries?

While social scientists study values scientifically, they do not, indeed cannot, verify values.

IV. Data or Variables

Data are often called variables or random variables. Why do we call data variables? The answer is simple: Data vary by chance. Consider the following variables:

- The number of minutes it takes you to commute from your home to school or work. (Each commute will vary depending on the time of your commute, the weather, traffic, your selected mode of travel, and chance.)
- The number of hours you sleep at night. (The amount of time you sleep varies day-to-day with your schedule and chance.)
- The number of classmates who attend your statistics class. (The number of students attending varies from class-to-class due to several factors and chance.)

- The number of calories you consume a day. (Your calorie consumption varies with your choice of meals, the size of your portions, how hungry you are, and chance.)

In many of the formulas, we will cover, variables are symbolized with letters. Most often the letter X is used. But other letters—Y, A, B, or C—are sometimes used. Using a Latin letter is shorthand for the fact that the variable is from a sample. English, Spanish, French, German, Italian, among other languages are written in Latin letters. The use of Greek letters means that the data are from a population. Here is a link to the Greek alphabet.

Data are not the same as information. Data are raw, unorganized facts in need of processing. Information, on the other hand, is data that has been organized in a context that makes it useful. The science of statistics uses a variety of techniques to turn data into useful information.

V. Types of Data/Variables

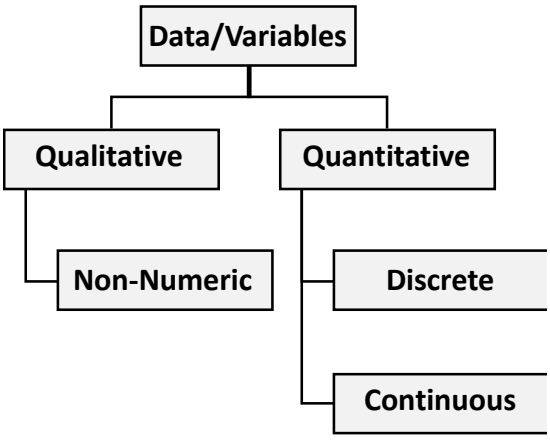

Figure 2: Types of Variable

We can distinguish between two broad categories of variables:

1) Qualitative variables.
2) Quantitative variables.

A. Qualitative Variables

Let's start with qualitative variables, which are also called categorical or attribute variables. Qualitative variables refer to a "quality," not a "quantity." Qualitative variables, therefore, are not numerical because these variables do not include numbers. Qualitative data rely on adjectives and other descriptive words to indicate appearance, color, texture, or other qualities. Some examples of qualitative data are:

- Your hair color.
- Your eye color.
- Your marital status.
- The species of your favorite pet.
- The breed of your neighbor's dog.
- The make of car that you used to learn to drive.
- Your favorite meal, baseball team, or article of clothing.
- The names of your classmates in your statistics class.
- The softness of a fur coat.

These data are qualitative. They are descriptive. They are not directly measurable.

While qualitative variables are not numerical, we can count qualitative variables. This gives qualitative data a quantitative dimension. M&Ms candy, for example comes in six colors: Blue, Brown Green, Orange, Red, and Yellow. We can count the number of each color in the bag. The results of these counts are called *frequencies*. We can also calculate the proportion of each candy color to the total number of pieces in the bag. The results are called the *relative frequencies*. **Please note:** The colors are qualitative variables. But the frequencies and relative frequencies are numerical and, therefore, quantitative data.

Table 1: Qualitative Data: M&Ms Candy Colors

Color	Frequency	Relative Frequency
Blue	5	8.77%
Brown	6	10.53%
Green	7	12.28%
Orange	16	28.07%
Red	13	22.81%
Yellow	10	17.54%
Total	57	100.00%

Please note: Qualitative data can contain numbers. These numbers, however, are not meaningful. Your student ID number, social security number, zip code, cell phone number, or credit card number are qualitative data despite the fact that they contain numbers. They are, in effect, only identification codes. We do not treat them as numbers. The arithmetic average of your classmates' zip codes or social security numbers, for example, would not provide any meaningful information.

Dichotomous or Binary data

Qualitative variables are sometimes *dichotomous* or *binary* variables. That means that they can only be placed in one of two groups. A variable is dichotomous if you use a yes or no question to classify the data. Here are a few examples:

A) Which of the following American presidents are alive?

- Abraham Lincoln: (No, died on April 15, 1865)
- Richard Nixon: (No, died on April 22, 1994)
- Ronald Reagan: (No, died on June 5, 2004)
- Jimmy Carter: Alive (Yes, as of April 26, 2020)
- George H. W. Bush: (No, died on November 30, 2018)
- Barack Obama: Alive (Yes, as of April 26, 2020)
- Donald J. Trump: Alive (Yes, as of April 26, 2020)

B) Which of the following successful business executives earned a college degree:

- Sophia Amoruso: (No)
- Jeff Bezos: (Yes)
- Richard Branson: (No)
- Warren Buffet: (Yes)
- Ursula Burns: (Yes)
- Barry Diller: (No)
- Stacy Ferreira: (No)
- Henry Ford: (No)
- Bill Gates: (No)
- Steve Jobs: (No)
- Elon Musk: (Yes)
- Sheryl Sandberg: (Yes)

- Meg Whitman: (Yes)
- Mark Zuckerberg: (Yes)

Other dichotomous variables include anything that has only two outcomes; the outcome of a coin toss (heads or tails), passing your statistics class (yes or no), completing an assignment on time (yes or no), voted in the last election (yes or no), or being a veteran of the military (yes or no).

Gender has traditionally been considered a dichotomous variable; that is, either male or female. But the binary nature of gender has never been universally accepted. All we have to do to confirm this assessment is to look at the work of the ancient Greek philosopher Plato (c. 428/427 BCE – c.348/347 BCE). In the *Symposium*, Plato presents a dialogue about love that takes place at a drinking party. Among the attendants are Socrates, who is Plato's teacher, Agathon the author of dramatic tragedies, Eryximachus, a physician, and Aristophanes, the great comic playwright. At this party, the guests give extemporaneous speeches on love. Aristophanes begins his speech by telling a myth about the origin of humans (*Symposium*, 189d – 190).[4] According to the myth, originally people had three genders, and we looked very different. We were shaped like spheres and had four arms, four legs, two heads, and two sets of genitals. Some of us were male, some female, and some a mixture of male and female. We were powerful and confident. So powerful, in fact, the gods began to fear us. Soon the god Zeus and his son Apollo hatched a plan to make humanity less threatening. The Gods did not want to eliminate humanity as they did the giants, so they decided to split us into two parts with each part having one head, two arms, two legs, and one set of genitals. Thereafter, we longed for our missing part and found pleasure when we embraced someone who felt like our missing half.

Today, many people are demanding that gender not be considered a binary variable. Facebook, for example, offers users 51 options to define their gender identity.[5] As Dan Levin writes in the June 30, 2019 issue of the *New York Times*, "These days, many teenagers view gender identity as existing on a spectrum. A person's pronouns, too are not limited to binary male and female, with the gender-neutral they/them gaining wider usage and acceptance."[6] In a June 28, 2019 *New York Times* article, Levin reports that in a recent survey on gender identity, "Respondents used a total of 116 different words and phrases to describe their sexual and gender identities, and their relationships."[7]

Data can also be *cross-sectional* or *longitudinal*. Cross-sectional data describes various segments of a population using sample data taken at a specific time. Longitudinal data tracks changes in the data over time.

B. Quantitative Variables

Quantitative variables are numerical. The information is reported with numbers. Here are examples of quantitative variables:

- The balance in your checking account.
- The amount of money you owe on your credit card.
- The number of parking spots in a parking lot.
- The number of seconds a Ferrari SF90 Stradale takes to complete a quarter-mile race from a standing start.
- The number of World Series championships won by your favorite Major League baseball team.
- The length of the longest bridge in the world.
- The temperature outside at dawn this morning.
- The number of hairs remaining on a bald man's head.
- The number of rooms in your home.
- The number of college students at your school who earn a degree within six years.
- The paid attendance at the July 4, 2019 New York Mets baseball game.
- Your weight, height, or blood pressure.

Quantitative variables can be either *discrete* or *continuous*. They are considered discrete if the value of the variables does not include fractional numbers. The number of puppies in a litter is an example of a discrete variable. A litter may have seven or eight puppies or any other whole number. A whole number is an integer or a number without fractions. No litter has seven-and-a-half puppies. Continuous variables, on the other hand, can take on any number in a range of numbers; which is to say, continuous variables include fractional numbers. The weight of an object is a continuous variable because it can be reported in kilograms, grams, milligrams, micrograms, and so on. The value of our data is limited only by the precision of our measuring device.

Selecting from this list of quantitative variables, the following variables are discrete:

- The balance in your checking account. (The discrete units are pennies, not fractional pennies.)
- The amount of money owed on your credit card (The discrete units are pennies. If fractional pennies are used, this variable would be continuous.)
- The number of parking spots in a parking lot. (There are no fractional parking spaces.)
- Number of World Series championships won by a Major League baseball team. (A team can win zero, one, two, three or more world series. No team can win a quarter of a world series. As of 2018, seven teams never won a world series. The New York Yankees won the greatest number of world series, 27.)
- Number of hairs remaining on a bald man's head. (No man, balding or otherwise, has a fractional hair growing on his head.)
- The number of college students at your school who earn a degree within six years. (There are no fractional students.)
- The paid attendance at the July 4, 2019 New York Mets baseball game. (A person either paid for a ticket or did not, ticket buyers are a discrete variable.)

From the list shown above, the following variables are continuous:

- Number of seconds a Ferrari SF90 Stradale takes to complete a quarter-mile race from a standing start (We measure time in fractional seconds.)

- The temperature outside at dawn this morning. (The Weather Bureau may have reported the temperature as 65º F, but temperature is measured on a continuous scale. The actual temperature, when measured with a precise device could have been 65.375º F.)
- The length of the longest bridge in the world. (The Danyang-Kunshan Grand Bridge in China is the world's longest bridge. It is said to be 164,800 meters long. Given the fact that bridges expand and contract with temperature, the length of this bridge is not exactly 164,800 meters.)
- Your weight. (You may have weighed yourself this morning, but your weight varies during the day. (Your scale may say you weighed 150 pounds, but your weight might actually be 150.375 pounds.)

V. Univariate, Bivariate, and Multivariate Data

Univariate data contain a single variable. The prefix of this word, *uni*, means one. Examples of univariate data include: The weight of the athletes on the football team, the annual income of college graduates, the number of states where the recreational use of marijuana is legal, or the total annual revenue of PepsiCo.

Bivariate data contain two variables. The prefix, *bi*, means two. We use bivariate data to determine whether there is a relationship between the two variables. For example, you weight and the average number of calories you consume. We will deal with bivariate data in Module 19: Linear Correlation and Regression.

Multivariate data contain three or more variables. For example, your weight, your average daily calorie consumption, and your average daily number of minutes of aerobic exercise. The prefix, *multi*, means many. There are many different types of multivariate analysis. These techniques are typically not covered in an undergraduate introductory statistics class.

VI. The Four Levels of Measurement

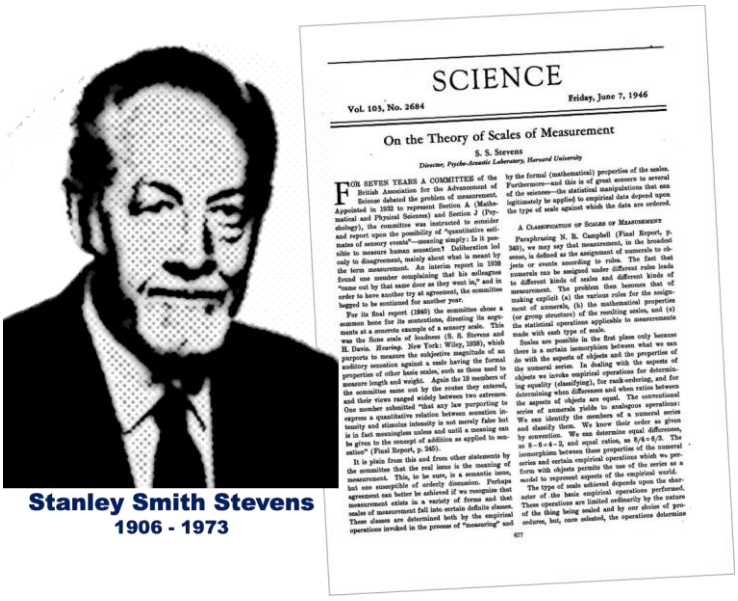

Figure 3: Stanley Smith Stevens, "On the Theory of Scales of Measurement," Science, June 7, 1946.

In 1946, Stanley Smith Stevens, an American psychologist and founder of Harvard University's Psycho-Acoustic Laboratory, published a very important article in the peer-reviewed journal *Science*. This article, "On the Theory of Scales of Measurement," introduced the concept of the four levels of measurement.[8] Knowing the level of measurement of your data is important because it determines what type of statistical analysis you can conduct.

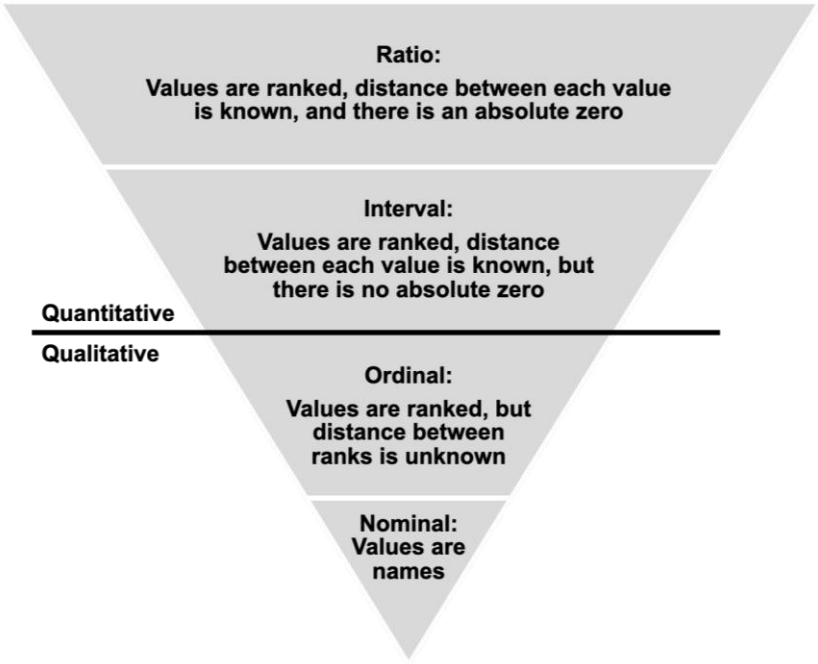

Figure 4: The 4 Levels of Measurement

According to Stevens, there are four measurement levels are **N**ominal, **O**rdinal, **I**nterval, and **R**atio. The mnemonic I use to remember the four levels of measurement is NOIR, the French word for black.

A) Nominal Level

Nominal is the lowest level of measurement. Such data are qualitative or categorical data and deal with naming things. Nominal data can be classified into categories, but these categories cannot be arranged in a meaningful order. While alphabetical order is very important for a dictionary or phone book, it is not considered important when classifying data by level of measurement. As with any qualitative data, you can add a quantitative dimension by counting frequencies or calculating relative frequencies. Here are some examples of nominal data:

- The color of your favorite shirt.
- The color of the socks you wore yesterday.
- The color of your hair.
- Your gender whether you use a binary classification or one of Facebook's 51 gender identities.
- The numbers on athletes' uniforms.
- Your religious affiliation.
- Your birthplace.
- Your political party, if any.

B) Ordinal Scale

Like nominal data, ordinal data are qualitative or categorical but they have an additional dimension. As the name "ordinal" suggests, we have a meaningful order or ranks. The difference between the ranks, however, cannot be determined or they are meaningless. Here are some examples of ordinal measurements:

- First place, second place, third place, and so forth in a beauty contest.
- Tee shirt sizes: Extra small, small, medium, large, and extra-large.
- The hotness of chili peppers: Hot, hotter, and hottest.

- Letter grades: A (Excellent), B (Good), C (Satisfactory), D (Passing), and F (Failing).
- Responses to a Likert question: Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree.

With each example, we cannot measure the distance between the ranks. Regarding the beauty contest, for example, we do not know the number of votes that determined who came in first, second, or third.

C) Interval Scale

With the interval scale, data are quantitative. The interval scale has a big advantage over the ordinal scale; we can now measure the interval between the ranks. But the interval scale lacks a natural, or true, zero.

The classic example of interval scale data is temperature measured on a Fahrenheit or Celsius scale. Today, for example, the high temperature is 87° F. Yesterday the high was 84º F. The interval between the high temperatures for these two days is 3° F. Tomorrow the predicted high temperature is 90° F. Last November the high temperature was 45° F. Because there is no real zero on the Fahrenheit scale, we cannot say that tomorrow will be twice as warm as the November day when the high temperature was 45° F. All we can say is that the difference in temperature is 45° F.

Other examples of interval scale data include:

- Calendar dates. (There is no day zero.)
- Calendar years. (There is no year zero.)
- Shoe size. (There is no size zero; a size 12, therefore, is not twice as big as a size 6.)
- Longitude on the map. (0^0, the prime meridian that goes through Greenwich England, is arbitrary.)
- Tee shirts with the following sizes: 36, 38, 40, 42, 44, and so forth
- SAT scores: Total scores on the SAT exam range from a low of 400 to a high of 1,600. Because there is no zero in this scale, we cannot say that

a student who achieved a score of 1,600 did twice as well as someone who scored 800.

D) Ratio Scale

The ratio scale is the most sophisticated of the four scales. It has everything the interval scale has plus a real, non-arbitrary zero. We can calculate ratios or proportions. Here are some examples:

- The number of minutes you have spent so far reading this module
- The amount of money in your pocket. (If today you have $20 and yesterday you had $10, the amount of money you have has doubled or is up 100 percent.)
- The number of credits toward your degree you have completed to date. (If you completed 30 credits, you are halfway towards an Associate's degree and one-quarter of the way towards a Bachelor's degree.)
- Initiation fees for President Trump's Mar-a-Lago Club doubled since he became president. As of June 2019, fees are $200,000, up from $100,000[9].

E) Level of Measurement Summary

Table 3: Levels of Measurement Summary

	Qualitative Data		Quantitative Data	
Key Feature	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Names	X	X	X	X
Rank Order		X	X	X
Intervals			X	X
Real Zero				X
Examples	Gender, color, zip code, make of car	Military rank, order of finish (1st, 2nd, 3rd)	Temperature, shoe size, SAT scores	Income, weight, height, distance, time

Knowing the level of measurement is important because it determines the type of statistical analyses that can be conducted.

1) With nominal data, we can:

- Calculate frequencies (Module 4)
- Calculate the mode (Module 5)

- Perform Chi-Square Tests (Module 17)

2) With ordinal data, we can:

- Calculate frequencies (Module 4)
- Calculate the mode (Module 5)
- Perform Chi-Square Tests (Module 17)
- Calculate percentiles (Module 5)
- Calculate the range (Module 5)
- Calculate the Interquartile Range (Module 5)

3) With interval data, we can:

- Calculate frequencies (Module 4)
- Calculate the mode (Module 5)
- Calculate the median (Module 5)
- Calculate percentiles (Module 5)
- Calculate the range (Module 5)
- Calculate the Interquartile Range (Module 5)
- Calculate the mean (Module 5)
- Calculate variance (Module 5)
- Calculate standard deviation (Module 5)
- Calculate the coefficient of variance (Module 5)
- Conduct parametric null hypothesis tests (Modules 13-16)
- Conduct correlation and regression analyses (Module 19)

4) With ratio data, we can:

- Calculate frequencies (Module 4)
- Calculate the mode (Module 5)
- Calculate the median (Module 5)
- Calculate percentiles (Module 5)
- Calculate the range (Module 5)
- Calculate the Interquartile Range (Module 5)
- Calculate the mean (Module 5)
- Calculate variance (Module 5)
- Calculate standard deviation (Module 5)
- Calculate the coefficient of variance (Module 5)
- Conduct parametric null hypothesis tests (Modules 13-16)
- Calculate ratios, proportions, and percentages (Appendix 1)

VII. Exercises

Answers to these questions can be found by carefully reading this module.

1. **Define Data**

2. **What is the difference between parameters and a statistics?**

3. **What is the difference between facts and values?**

4. **What alphabets are used to symbolize data from populations and data from samples?**

5. **What are the key features of Qualitative data?**

6. **What are the key features of Quantitative data?**

7. **What is the difference between continuous and discrete data?**

8. **Distinguish among univariate, bivariate, and multivariate data**

9. **What are the characteristics of Nominal data?**

10. **What are the characteristics of Ordinal data?**

11. **What are the characteristics of Interval data?**

12. **What are the characteristics of Ratio data?**

13. **Your sister receives a 650 on her math SAT exam. What level of measurement is this?**

14. **Jose ranks second in his graduating class. What level of measurement is this?**

15. **While ill with COVID-19, Maria lost 10 lbs. What level of measurement is her weight?**

16. **Sam was born in Philadelphia. What level of measurement is birthplace?**

* * *

* * *

---

[1] Thomas H. Davenport and Jeanne G. Harris. *Competing on Analytics: The New Science of Winning*. (Boston, MA: Harvard Business School Press, 2007), p. 30.

[2] *The Constitution of the United States of America*. (New York: ACLU, 2019), pp. 2-3.

[3] Max Weber, *From Max Weber: Essays in Sociology*, Hans Gerth and C. Wright Mills, editors. (New York: Oxford University Press, 1958), p. 357.

[4] Edith Hamilton and Huntington Cairns (eds.). *The Collected Dialogues of Plato Including the Letters. The Symposium* translated by Michael Joyce. (Princeton, NJ: Princeton University Press, 1971), pp. 526-574.

[5] https://www.thedailybeast.com/what-each-of-facebooks-51-new-gender-options-means

[6] Dan Levin. "The Fluidity of Gender, Language and the 'Human Experience.'" *The New York Times*, June 30, 2019, p. 22.

[7] Dan Levin. "The Human Experience is Infinite." *The New York Times*, June 28, 2019, https://www.nytimes.com/interactive/2019/06/28/us/pride-identity.html?rref=collection%2Fbyline%2Fdan-levin&action=click&contentCollection=undefined&region=stream&module=stream_unit&version=latest&contentPlacement=1&pgtype=collection

[8] Stanley Smith Stevens. "On the Theory of Scales of Measurement." *Science, New Series*, Vol. 103, No. 2684. June 7, 1946) pp. 677-680. http://psychology.okstate.edu/faculty/jgrice/psyc3214/Stevens_FourScales_1946.pdf

[9] https://www.cnbc.com/2017/01/25/mar-a-lago-membership-fee-doubles-to-200000.html. https://www.washingtonpost.com/opinions/2019/05/15/trumps-businesses-are-faltering-thats-good-news/?utm_term=.ed9dd3b40543