CHAPTER 1: INTRODUCTION AND DATA COLLECTION

[This chapter is based on Sections 1.1 and 1.2 from the textbook]

Many people has the idea that statistics has to do with numbers and percentages. Indeed, we often see that people share news such as "80% of the population has had COVID," or "50% of US citizens consider themselves happy," etc[1]. But there is more than percentages and fun facts. A formal definition is presented below.

Definition 1.1.

(i) *Data is a fact or proposition used to draw a conclusion or make a decision*

(ii) *Statistics is the science of collecting, organizing, summarizing and analyzing data to draw conclusions or answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.*

So yes, statistics is about the numbers. But not only these numbers. It is also about the process of collecting information (data), computing summarizing measures and quantifying how certain we are that our conclusions are representative. The goal of statistics is to transform data (or observations) into knowledge and understanding of the world around us.

Let's see some examples.

Example 1.1.

- *Statistics may be used to study if a newly discovered drug can cure cancer, and the data would be the reaction of several patients to the drug*
- *Statistics is used in sports to help the coach decide which player is the best fit for a team, and the data would be the performance of the players in several games*
- *Statistics is used by politicians to help them understand the population's needs, and the data would be the answers that people give to several polls*
- *Statistics is used to predict the length (in days) of the summer weather, and the data would be the length of the summer in previous years and the information that researchers have in terms of variables that affect the length of seasons*
- *Etc. Can you think of more examples?*

We must be very careful when we draw conclusions based on the data we collected because we may ignore relevant information. Let's consider the following example.

Example 1.2. *A study wants to find out if students learn more when the instructor uses slides or whiteboard in the lectures. Similarly to clinical trials, we need two control groups: one being taught with slides and one using whiteboard. At the end of the term, we will make both groups take the same exam to measure their knowledge.*

Suppose that the exam was graded from 0 to 100 points, the group with slides got an average score of 87 points, and the group with whiteboard, 80 points. Can we conclude that using slides is better than whiteboard? Or, could we explain the difference in the score by other factors? Were the students similar? Were the instructors similar? In other words, are there other variables that can affect the result and we did not consider?

In the example above, suppose that the group of students taught with slides (and got a higher score in the exam) is composed by students in third or fourth year of a math major, and the class is an advanced math class (such as linear algebra). The second group is more heterogeneous, and many of the students had not taken any math classes in college before this one. Can we draw any meaningful conclusion to answer our slides vs. whiteboard question?

The situation described above illustrates the concept of lurking variable, that is, variables that affect the results of our study, but we did not consider. In this course we will learn to deal with these situations. You will learn when you can trust a statistical study and when you should be skeptical.

Let's get started with some key definitions.

[1] These numbers are made up, and do not represent the results of any statistical study. They are meant to be simple examples only.

Definition 1.2.

- *The population is the entire group to be studied*
- *A sample is a subset of the population that is being studied*
- *An individual is a person or object that is member of the population*

Let's apply these definitions to the example about slides vs. whiteboard. We have:

- Population = All students in the world
- Sample = The two groups we selected for the study
- Individual = Any student, not necessarily from the groups we study

Definition 1.3.

- *A variable is a characteristic of each individual that we are observing/surveying*
- *A statistic is a numerical summary of a sample*
- *A parameter is a numerical summary of the population*

In our example, a variable is the exam score, a statistic could be the total average score of the students, the number of students who scored over 85, etc. To compute a parameter, we would need to give the test to the entire population, that is, all the students in the world. Since such experiment is practically impossible, we wish to use the results from our sample and generalize them to the population.

Before moving on, let's consider another example.

Example 1.3. *Suppose that we want to study the percentage of students in this class that were born in Virginia, USA.*

(a) What is the population? Give an example of a sample and an individual.

(b) Give an example of a variable, a statistic and a parameter

Solution. In this case,

- The population is the entire class, that is, the 50 students registered in this class.
- A sample is any subset of students. For example, the students that sat in the first two rows in the first class.
- An individual is every single student in the class
- A variable is the answer of each student to our question. In this case, it can be 'yes' or 'no'
- A statistic would be the percentage of students that were born in Virginia among the students that sat in the first two rows in the first class. That is, the percentage of students **from the sample** that were born in Virginia
- A parameter would be the percentage of students that were born in Virginia considering the whole class, that is, out of the 50 students.

□

In the last example, computing parameters was easy because the population was small (50 students). However, most of the time it is impossible to ask every single individual. In such case, we want to use a representative sample to draw conclusions about the population. Depending on what we do with the information of the sample, we may use statistics in two different ways. We define them below.

Definition 1.4.

- *Descriptive statistics consist of organizing and summarizing the data from a sample*
- *Inferential statistics uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.*

In the example of slides vs. whiteboard, a descriptive statistic would be that the average score in the exam was 83.5 (assuming that both groups are of the same size). Here, we don't draw any conclusions about the rest of the population. We simply describe our sample using summary measures.

Inferential statistic would be saying that the average score in the test is between 80 and 90 with 95% of confidence. Observe that we are giving a range of possible outcomes, and we are also stating how sure we are of this range.

So, how do we do statistics?

Principle 1.1. *The process of statistics has the following steps:*

1. **Identify the research objective:** *What is the question we are trying to answer? What is the population involved?*
2. **Collect the data needed to answer the question(s):** *Since surveying the entire population is often difficult, we must select a representative sample. This step is vital because if the data is collected incorrectly, the conclusions we draw are meaningless.*
3. **Describe the data:** *Obtain an overview of the data to determine the statistical methods we will use. We may use charts, tables, histograms, etc*
4. **Perform inference:** *Use the information of the sample to draw conclusions of the population and report a level of reliability of the results*

Let's see an example.

Example 1.4. *A Gallup Poll asked the following question to 1087 adults: "Do you have the occasion to use alcoholic beverages such as liquor, wine or beer, or are you a total abstainer?*

(a) Identify the population, sample, variable, statistic and parameter

(b) Identify each step of the process of statistics

Solution.

(a) In this case, we are not given any information about the goal of the study, so we need to keep it as general as we can. Then,

- The population is composed of all adults in the world
- The sample is the 1087 adults that participated in the poll
- A variable is the answer that the 1087 adults gave to the answer, that is, 'drink alcohol' or 'abstainer'
- A statistic is the percentage of adults that drink alcohol among the 1087 that participated in the poll (sample)
- A parameter is the percentage of adults that drink alcohol considering all adults in the world (population)

(b) The process of statistics is:

1. **Research objective:** Determine the percentage of adults that drink alcohol
2. **Collecting data:** Selecting the adults that have access to the poll
3. **Describe the data:** A table with the results
4. **Perform inference:** Based on the data, what conclusions can we draw?

□

Depending on the statistical study, the variables may be very different. In the examples we have seen so far we have scores (numbers), yes/no answers, and drink/abstainer. So how can we classify them?

Definition 1.5.

- *Qualitative (or categorical) variables allow for classification of individuals based on some attribute or characteristic.*
- *Quantitative variables provide numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.*

In our examples so far, the scores in the exam is the only quantitative variable, and the yes/no answers and drink/abstainer answers are qualitative variables.

Example 1.5. *Determine whether the following variables are qualitative of quantitative.*

(a) Gender

(b) Temperature

(c) Number of pets

(d) Zip code

Solution. We have:

(a) Gender is qualitative

(b) Temperature is quantitative

(c) Number of pets is quantitative

(d) Zip code is a number, but it is also qualitative. Why? Zip code represents a location and adding or subtracting zip codes does not have any meaning.

□

Among quantitative variables, we also have two classifications.

Definition 1.6.

- *A discrete variable is a quantitative variable that can take a finite our countable number of possible values*
- *A continuous variable is a quantitative variable that has an infinite number of possible values that are not countable, that is, they correspond to a continuous range of values.*

For example, the number of pets you have, the number of siblings you have, the number of students in this class, are **discrete** variables. The temperature in the room, height of a person, age of a person are **continuous** variables. Let's see more examples.

Example 1.6. *Determine whether the following variables are qualitative or quantitative. For quantitative variables, determine if they are discrete or continuous.*

(a) Distance between where you live and this classroom

(b) Brand of the refrigerator in a home

(c) Number of a football player's jersey

(d) Assessed value of a house

(e) Goals scored by a soccer player in a season

(f) Length of a song in minutes

Solution. We have:

(a) Distance between where you live and this classroom: Quantitative, continuous

(b) Brand of the refrigerator in a home: Qualitative

(c) Number of a football player's jersey: Qualitative

(d) Assessed value of a house: Quantitative, continuous

(e) Goals scored by a soccer player in a season: Quantitative, discrete

(f) Length of a song in minutes: Quantitative, continuous

□