CSC1311

STATISTICS FOR PHYSICAL SCIENCE AND ENGINEERING

November 30, 2021

INTRODUCTION

Why Statistics?

- This is a required course for your degree.
- It is a prerequisite for many other topics.
- Data everywhere, particularly in this big data era.
- Sampling vs censusing.
- Decision Making: Statistics will help you make important decision.

Data collection is a crucially important step in Statistics. We use the collected and observed sample to make statements about a much larger set — the population.
A **population** consists of all units of interest. Any numerical characteristic of a population is a **parameter**. A **sample** consists of observed units collected from the population. It is used to make statements about the population. Any function of a sample is called **statistic**.
Information is gathered in the form of samples, or collections of observations. Samples are collected from populations, which are collections of all individ- uals or individual items of a particular type. At times a population signifies a scientific system. For example, a manufacturer of computer boards may wish to eliminate defects. A sampling process may involve collecting information on 50 computer boards sampled randomly from the process. Here, the population is all

- Costs of surveying the entire population may be too large or prohibitive
  - e.g., Television networks monitor the popularity of their programs;
- Destruction of elements during investigation
  - e.g., Manufacturers estimate the average lifetime of light bulbs; doctors take a blood sample to check for disease.
- Unknown future
- Destruction of elements during investigation
  - e.g., stock index; temperature tomorrow

INTRODUCTION

Decision Making

- How do large retailers (like COSTO, Walmart) fill their store shelves so as to meet the customer demand while minimizing their operating cost?
- How do doctors in the hospital make diagnose? How do political leaders run their campaign?
- How do Investment Banks in Wall Street (or Bay Street) decide which stock (or stocks) to invest?
- How do insurance companies decide the premium for a particular client?
- How do car dealers decide how many car models of each brand to be kept in their locations?

INTRODUCTION

What is Statistics?

It is the science and art

- of collecting, organizing, and representing data in such a way that the characteristics and patterns of the data can be easily captured (Descriptive Statistics);
- also, of estimating attributes and drawing inference from a sample about the entire population (Inferential Statistics).

INTRODUCTION

Examples: descriptive or inferential?

- In 1995, 45% of Canadian households owned a computer and 25% were connected to internet; On average, Canadians spend 1.3 hours per day commuting, and 1.5 hours per day with their children.
  - descriptive
- The accounting department of a firm will select a sample of invoices to check for accuracy of all the invoices of the company.
  - inferential

INTRODUCTION

Population parameters and sample statistics

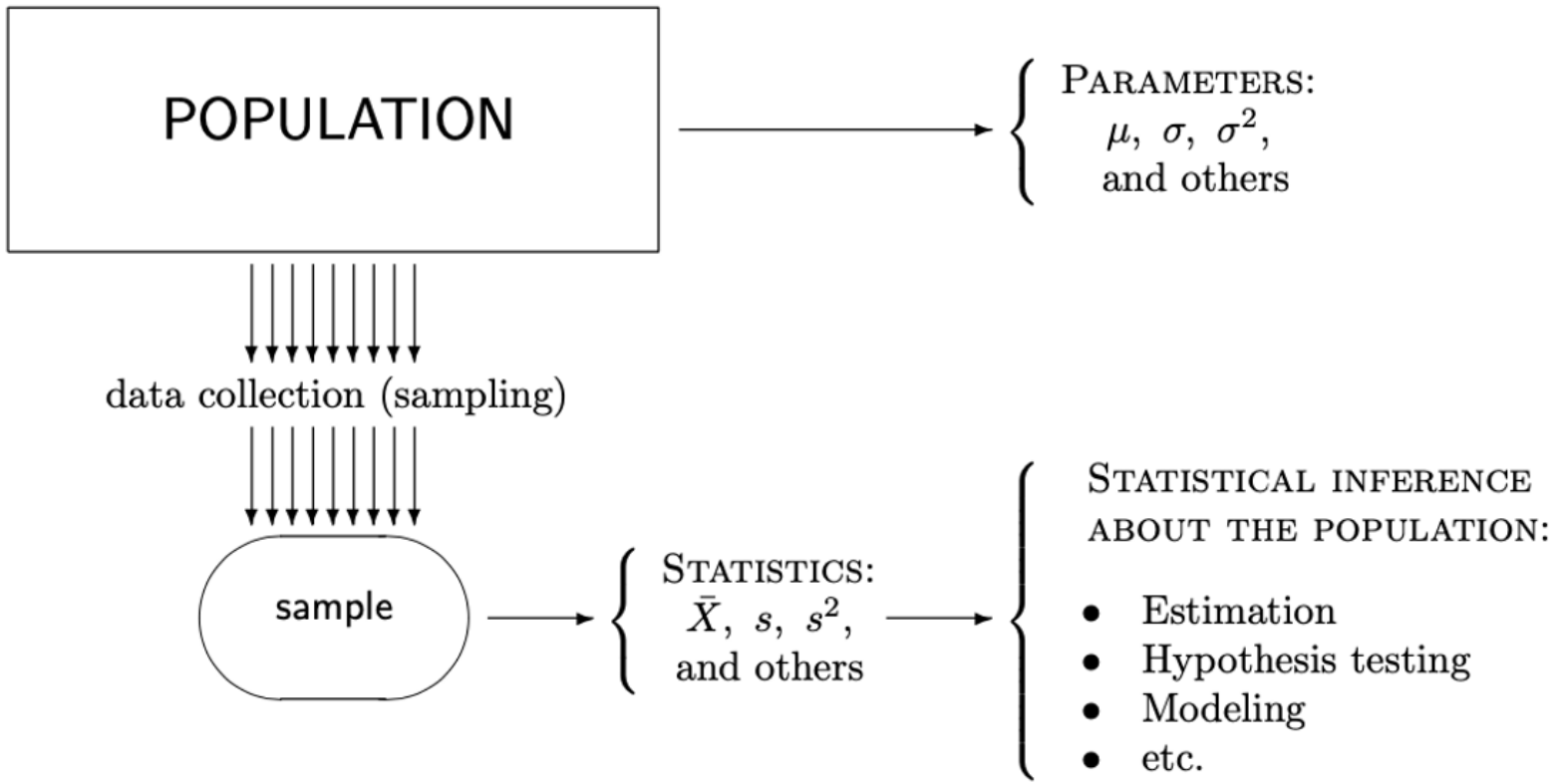

Figure: Population parameters and sample statistics

- Sampling and non-sampling errors refer to any discrepancy between a collected sample and a whole population.
  - **Sampling errors** are caused by the mere fact that only a sample, a portion of a population, is observed. For most of reasonable statistical procedures, sampling errors decrease (and converge to zero) as the sample size increases.
  - **Non-sampling errors** are caused by inappropriate sampling schemes or wrong statistical techniques. Often no wise statistical techniques can rescue a poorly collected sample of data. Look at some examples of wrong sampling practices.

In this course, we shall be using *simple random sampling*, which is one way to avoid non-sampling errors.

Example 1 (Sampling from a wrong population)

To evaluate the work of a Windows help desk, a survey of **social science students** of some university is conducted. This sample poorly represents the whole population of all Windows users. For example, **computer science students** and especially computer professionals may have a totally different opinion about the Windows help desk.

Example 2 (Dependent observations)

Comparing two brands of notebooks, a senior manager asks all employees of her group to state which notebook they like and generalizes the obtained responses to conclude which notebook is better. Again, these employees are not randomly selected from the population of all users of these notebooks. Also, their opinions are likely to be dependent. Working together, these people often communicate, and their points of view affect each other. Dependent observations do not necessarily cause non-sampling errors, if they are handled properly. The fact is, in such cases, we cannot assume independence.

Example 3 (Not equally likely)

A survey among passengers of some airline is conducted in the following way. A sample of random flights is selected from a list, and ten passengers on each of these flights are also randomly chosen. Each sampled passenger is asked to fill a questionnaire. Is this a representative sample? Suppose Mr. X flies only once a year whereas Ms. Y has business trips twice a month. Obviously, Ms. Y has a much higher chance to be sampled than Mr. X. Unequal probabilities have to be taken into account, otherwise a non-sampling error will inevitably occur.

Example 4 (Presidential Election of 1936)

A popular weekly magazine The Literary Digest correctly predicted the winners of 1920, 1924, 1928, and 1932 U.S. Presidential Elections. However, it failed to do so in 1936! Based on a survey of ten million people, it predicted an overwhelming victory of Governor Alfred Landon. Instead, Franklin Delano Roosevelt received 98.49% of the electoral vote, won 46 out of 48 states, and was re-elected. So, what went wrong in that survey? At least two main issues with their sampling practice caused this prediction error. First, the sample was based on the population of subscribers of The Literary Digest that was dominated by Republicans. Second, responses were voluntary, and 77% of mailed questionnaires were not returned, introducing further bias. These are classical examples of non-sampling errors.

- **Simple random sampling** is a sampling design where units are collected from the entire population independently of each other, all being equally likely to be sampled.
- Observations collected by means of a simple random sampling design are iid (independent, identically distributed) random variables.
- Simple random sampling implies that any particular sample of a specified sample size has the same chance of being selected as any other sample of the same size. The term **sample size** simply means the number of elements in the sample.

Example 5

To evaluate its customers' satisfaction, a bank makes a list of all the accounts. A Monte Carlo method is used to choose a random number between 1 and N, where N is the total number of bank accounts. Say, we generate a Uniform(0,N) variable X and sample an account number $\lceil X \rceil$ from the list. Similarly, we choose the second account, uniformly distributed among the remaining N - 1 accounts, etc., until we get a sample of the desired size n. This is a simple random sample.

- **Simple random sampling** is a sampling design where units are collected from the entire population independently of each other, all being equally likely to be sampled.
- Observations collected by means of a simple random sampling design are iid (independent, identically distributed) random variables.
- Simple random sampling implies that any particular sample of a specified sample size has the same chance of being selected as any other sample of the same size. The term **sample size** simply means the number of elements in the sample.

Example 6

An example is a data collection of a study conducted at the Bayero University, Kano and a State University on the development of a relationship between the roots of trees and the action of a fungus. Minerals are transferred from the fungus to the trees and sugars from the trees to the fungus. Two samples of 10 northern red oak seedlings were planted in a greenhouse, one containing seedlings treated with **nitrogen** and the other containing seedlings with **no nitrogen**. All other environmental conditions were held constant. All seedlings contained the fungus Pisolithus tinctorus. The stem weights in grams were recorded after the end of 140 days. The data are given in Table 1.1.

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

Table: Data Set for Example 6

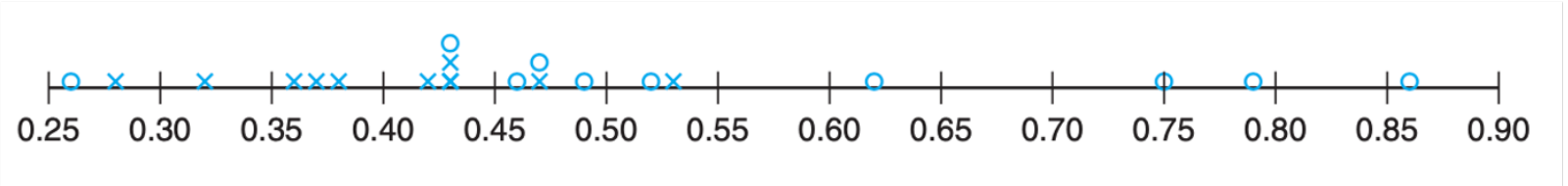

Figure: A dot plot of stem weight data

In this example there are two samples from two separate populations. The purpose of the experiment is to determine if the use of nitrogen has an influence on the growth of the roots.

The method of Experimental Design

1. Define the experimental goal or a working hypothesis
2. Design an experiment
3. Collect and represent data
4. Estimate the values/relations
5. Draw inferences
6. Predict and prepare policy analysis

The measures that help characterize the nature of the data set fall into the category of **descriptive statistics**.

Variables and types

Variable

- A variable is a characteristic/attribute of a population/sample that is interest in a particular investigation and the value of the variable can "vary" from one entity to another.
  - A person's gender is a variable, which could have the value of "Male" for one person and "Female" for another.
  - The rank of faculty members in Business Administration is a variable, which could have the value of "Full Professor" for one person, "Associate Professor" for another, and "'Assistant Professor"' for yet another.
  - Temperatures in this classroom is a variable which could have the value of "20" or "100".
  - Annual salary of NBA players (or hockey players in Canada), which could have the value of "10M" or "5M".

Types of Data: Qualitative vs. Quantitative Variables

- Qualitative (a.k.a. categorical)
  - Qualitative variables take on values that are names or labels.
    - Examples: gender, country names, color
- Quantitative (a.k.a, numeric)
  - Quantitative variables are numeric.
    - Examples: number of bedrooms in houses, number of minutes to the end of this class, distance between two cities

Types of Data: Discrete vs. Continuous Variables

- Quantitative variables can be further classified as discrete or continuous
  - **Continuous Variable:** a variable can take on any value within a range;
    - Examples: the number of minutes to the end of this class, distance between two cities, pressure in a tire, weight of a mutton cut, height of students in a class
  - **Discrete variable:** a variable can take only certain value (finite or countably infinite) within a range.
    - Examples: number of bedrooms in houses

Level of measurement

Suppose now we represent all data /variable using numbers, then

- Level of measurement (a.k.a. scales of measure) are the different ways numbers can be used.
  - There are four levels of measurements
    1. Nominal
    2. Ordinal
    3. Interval
    4. Ratio
  - However, representing variables as numbers does not give you the license to perform the regular logical/arithmetic operations all the time (such as comparison, addition, subtraction, multiplication, and division etc.); or to infer anything about the magnitude or quantitative difference between the numbers.

Nominal level

- A variable is at the nominal level if none of the five operations (namely comparison, addition, subtraction, multiplication, and division) is meaningful.
- At the nominal level of measurement, numbers are assigned to a set of mutually exclusive and exhaustive categories for the purpose of naming, labeling, or classifying the observations, but no arithmetic operation is meaningful;
- where
  - **Mutually exclusive:** any individual object is included in ONLY ONE CATEGORY
  - **Exhaustive:** any individual object MUST APPEAR in one of the categories

Nominal level

Example

- Examples
  - Barcodes
  - Social insurance numbers (SIN)
  - Student IDs
  - Phones numbers
- The fact that the barcode for one product is higher than that for another, or that your SIN is higher than mine tells us nothing.
- In surveys we often use arbitrary numbers to code variables such as religion, ethnicity, major in college or gender.

Ordinal level

- A variable is at the ordinal level if only comparison is meaningful.
  - Examples
    - Rank of faculty members
    - QS ranking of World Universities
    - US News Ranking of
  - The differences between data values cannot be determined or are meaningless.
    - For instance, first class is better than economy and that business is in between. Just how much better first class is compared to business, and business compared to economy varies from airline to airline, and even from flight to flight.

Interval

- A variable is at the interval level if only division is meaningless.
- Reason: Interval data have meaningful intervals between measurements, but there is no true starting point (zero).
  - Temperatures in Celsius and Fahrenheit are interval data
  - Certainly order is important and intervals are meaningful.
  - However, a $20^{\circ}C$ dashboard is not twice as hot as the $10^{\circ}C$ outside.

Ratio

Data

- A variable is at the ratio level if all logical/arithmetic operations are meaningful.
- Reason: Ratios between measurements as well as intervals are meaningful because there is a non-arbitrary zero point.
- Examples
  - Income
  - Distance
  - Height

A general method for identifying the level of measurement

Data

- Ask yourself the following three questions:
  - Is order meaningful?
    - No! then the data is nominal
  - Is difference meaningful?
    - No! then the data is ordinal
  - Is zero meaningful
    - No! then the data is interval
    - Yes! then the data is ratio