Chapter 1

Introduction

Shaikh Tanvir Hossain

October 21, 2022

East West University

Outline

1. The idea of Statistical Inference
2. Administrative details

1. The idea of Statistical Inference

2. Administrative details

- Statistics starts from data. Suppose we have collected following data, which shows the income of 10 people in Bangladesh (all incomes are in 1,000 tk and the negative value means the person is in debt).

	avg. income/year
1.	20
2.	60
3.	20
4.	−20
5.	−30
6.	−10
7.	80
8.	10
9.	30
10.	40

Table 1: avg. income/year of 10 individuals

- So 20 means 20, 000 tk.
- In Statistics we always think the data is a *random sample*, Question - What does this mean?
- When you hear the word sample, automatically you should think there is a *"population"* in the background.
- Question - What is a population? In simple words, the population is a bigger *set* from which wanted to learn something, and a sample is just a part of it, or a subset.

- In real life most of the time we don't have access to the population, or get data from the population, so the solution is - *sampling*. In the example above, you can think our population is the set of all people in Bangladesh who can earn money, and we picked a sample of 10 out of it.

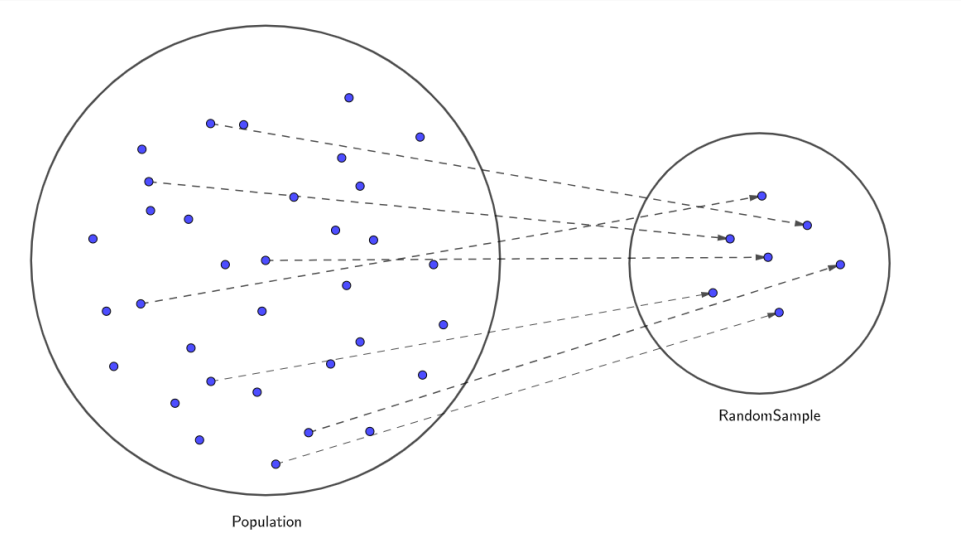

Figure 1: Sample of 7 out a larger population!

The idea of a Inferential Statistics

- Now let's try to understand *why we are doing sampling here or what is our goal here?*. As we have already mentioned almost all of the time in Statistics, our goal is to learn something about the population or *make some conclusion about the population*.
- For example, maybe we are interested to know about the *population average income* of Bangladesh. This is also called *population mean*.
- Now if we don't know the income of all of the people in Bangladesh, it is impossible for us to calculate it. Suppose this uknown number is μ and we do not know μ. However what we could do is, we could calculate the sample average or sample mean, which is

$$\frac{20 + 60+, \ldots, +40}{10} = 20$$

- Now with this sample mean we can conclude the *"maybe population mean is close to 20 thousand tk"*.
- This is the idea of inferential statistics or statistical inference, that is we want to say something or find something about the population, but we don't have access to the population so we *infer* it using the sample*.

*here the word infer means a systematic guess!

- Is this a good conclusion? The answer depends on many things, but two are very crucial,
  - ✓ How "good" is our sample? A good sample means, the sample is a good *representative* of the population.
  - ✓ How big is our sample?
- We will talk about the benefit of having large sample later, first let's think about a situation when we do NOT have a good sample.
- Suppose somehow we got a sample of all rich people in Bangladesh, is this a good sample of the population? The answer is NO? why? Ans - not a good representative.
- So we need to pick a sample such that in terms of variability it is similar to the population, we can do such sampling if *everyone in the population has the same probability (or chance)* of getting selected in our sample.
- This sampling is known as *Simple Random Sampling*[†] and we will call the sample taken in this way a *random sample* of the population.

[†] there are other types of random sampling, but we do not need them here!

Figure 2: How to make the sample random?

- Question - How can we get a random sample? Ans - There are many ways we can do this, here is a simple way, we write all names in the population in pieces of paper, then put them in a rolling basket, roll it and then start picking. One important point is, every-time we pick one name, we have to put it back to the basket and roll again before we take the next one (this is called sampling with replacement! question - *what if we don't do this?*).
- So we will pick 10 names in this way so that everyone in the population has an equal chance of getting selected, by this we will avoid only rich or only poor people. After that we will write down their avg. income/year, then we will have a good sample. The data will look like what we have in Table 1.
- We will call each of the data point - *a random data point*, and the total sample together is called - *a random sample*.

- In real life definitely there are better ways to do this (e.g., using generated or pseudo random numbers).
- If you have understood till now then you alrready understood the word "random". Actually the word *"random"* is a synonym for *"uncertain"*. So randomness means there is an uncertainty[‡]
- Ques - what is the uncertainty here?
  - ✓ Note when we pick a random sample from the population, in principle any 10 people might appear in our sample.
  - ✓ So before taking the sample, we don't know which 10 individuals will come in our sample (or which 10 income levels we will see).
  - ✓ But after the sampling is done, we have the sample at our hand and this uncertainty is gone!
- So when we think about a data set in statistics we always think about a random data set or a random sample. So Statistics deals with random data (keep this always in your mind)
- Since the sample is random, *the sample mean 20 whcih is a number here is also random*. Depending upon the sample it could be a different number.

[‡]Uncertainty appears when we do not have all information! if one knows everything there is no uncertainty to him!

- So here is the story of what we did
  - ✓ There is a population (maybe the set of all people who can earn in Bangladesh)
  - ✓ We are interested in a population characteristics, e.g., *population average income*. This is our *target object*.
  - ✓ But this is unknown to us, let's denote this with μ. Here μ is just a symbol for this unknown quantity. So μ is the population average income.
  - ✓ So we gathered a random sample of 10 individuals and calculated a sample mean $20,000$ tk. This is an *estimate* of the unknown quantity μ.
  - ✓ Finally we can conclude that "maybe our estimate is a good estimate, this means that maybe the sample mean $20,000$ is close to the unknown population mean μ"
- This is what we cal *Statistical Inference* - using a sample to say somethig about the population. And this course is about Statistical Inference.
- Before we start Statistics, we need to review Probability theory.
- Probability theory is a systematic study of *uncertainty*. You have already seen probability theory in ECO-104. We will have a short and quick recap before we start.

1. The idea of Statistical Inference

2. Administrative details

Some adminisrative stuffs

- Welcome to the course! This is ECO204 - Section 2, so if you are looking for a different class, please leave, otherwise you can enjoy!
- You will get the detailed course outline via email. Please check your emails regularly, this is very very important!
- The classes are on Sundays and Tuedays (10.10 - 11.40), Room - 112
- There will be a *google classroom*, again you will recieve the invitation *via email*. If you don't know how to use google classroom, please ask your friends who can, or let me know.

- About the computer work. I will try to book the computer lab room and we will solve some problems using Microsoft Excel or R.
- All important dates are in the academic calendar https://www.ewubd.edu/academic-calendar-details/fall-2022-undergraduate-programs, so please check it carefully.
- All exam dates are in the exam calendar https://www.ewubd.edu/academic-calendar-details/fall-2022-exam-schedule
- If you have any questions or concerns please do not hesitate to ask me, the fastest way to get your answer is via email, so please write your question clearly in a formal way and also write your name, course, section so that I can recognize you. The email address is tanvir.hossain@ewubd.edu
- So now let's start!
- In the next chapter we will make a quick recap of Probablity Theory.