CENG 222 - Chapter 8: Introduction to Statistics

Burak Metehan Tunçel - April 2022

We learned analyzing problems and systems involving uncertainty, to find *probabilities, expectations*, and *other characteristics* for a variety of situations, and to produce forecasts that may lead to important decisions.

What was given to us in all these problems? Ultimately, *we needed to know the distribution and its parameters*, in order to compute probabilities or at least to estimate them by means of Monte Carlo. Often the distribution may not be given, and we learned how to fit the suitable model, say, Binomial, Exponential, or Poisson, given the type of variables we deal with. In any case, parameters of the fitted distribution had to be reported to us explicitly, or they had to follow directly from the problem.

This, however, is rarely the case in practice. Only sometimes the situation may be under our control, where, for example, produced items have predetermined specifications, and therefore, one knows parameters of their distribution.

Much more often *parameters are not known*. To apply the knowledge we learned, *we need to collect data*. A properly collected sample of data can provide rather sufficient information about parameters of the observed system. We will learn how to use this sample

- to visualize data, understand the patterns, and make quick statements about the system's behavior;
- to characterize this behavior in simple terms and quantities;
- to estimate the distribution parameters;
- to assess reliability of our estimates;
- to test statements about parameters and the entire system;
- to understand relations among variables;
- to fit suitable models and use them to make forecasts

1 Population and Sample, Parameters and Statistics

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.

Definition 1

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**.

In real problems, we would like to make statements about the population. To compute probabilities, expectations, and make optimal decisions under uncertainty, we need to know the population *parameters*. However, the only way to know these parameters is to measure the entire population, i.e., to conduct a census.

Instead of a census, we may *collect data in a form of a random sample from a population* (Figure 1). This is our data. We can measure them, perform calculations, and estimate the unknown parameters of the population up to a certain measurable degree of accuracy.

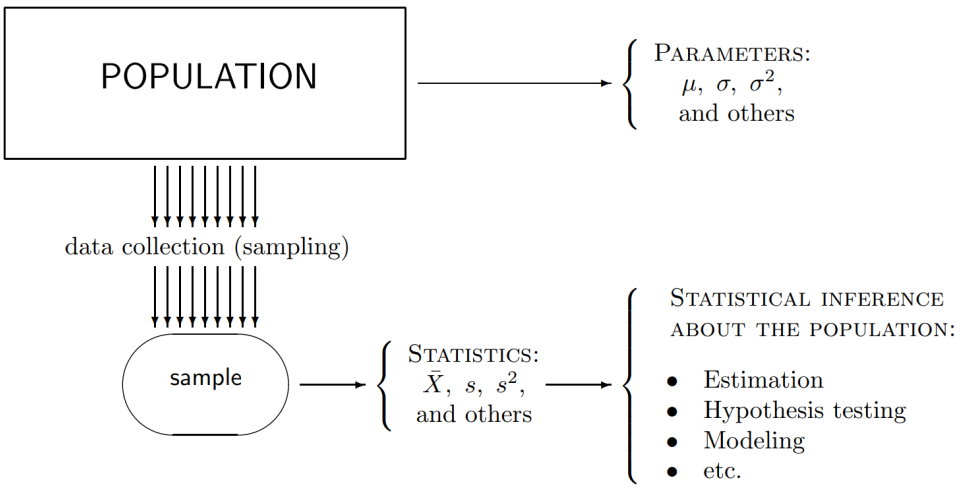

Figure 1: *Population parameters and sample statistics.*

A sample may sometimes give a rather misleading information about the population although this happens with a low probability. *Sampling errors cannot be excluded.*

1.1 Sampling and Non-sampling Errors

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.

Note: *Check the examples 8.1-8.5 in the textbook.*

We will focus on *simple random sampling*, which is one way to avoid non-sampling errors.

Definition 2

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.

Example 1

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.