7 BASICS OF INFERENTIAL STATISTICS

In real life, the data points about all the possible observations are rarely available for analysis and we have to make do with samples. That is where inferential statistics comes into the picture, which we will discuss in this chapter.

7.1 POPULATION & SAMPLE

Inferential statistics allows us to draw inferences about a given **population**, using only its subset called the **sample**. A sample is typically collected and extrapolated for the entire population.

For example, you would have filled a survey questionnaire or given a rating about a product in your life. The company manufacturing the product tries to gauge the general feedback of all its customers (the population in this case) by only surveying a fixed number of people (sample). If the average rating given by all the people in the sample is 4.9 out of 5 stars, the company can be happy about its product. If on the other hand it is 3 out of 5 stars, the company's management would want to improve the quality of the product.

This is both time efficient and cost efficient for the company. It is important that the sample is a good representative of the population, so that the inferences drawn from the sample generalize well to the population.

7.2 SAMPLING DISTRIBUTION

In our example, what the company is really interested in knowing is what all the customers (population) think about the product. A good metric to gauge this is the average rating given by all the customers, which is the population mean μ. However, as rating from each and every customer can't be obtained, it has to use the average rating from a sample (the sample mean or $\underline{x}$) as its proxy.

Consider a scenario where the company collects ratings from two samples of fifty customers each. The average rating for the first sample being 4.9 out of 5 stars and for the second sample it is only 3 out of 5 stars. Which sample result should the company take into account now?

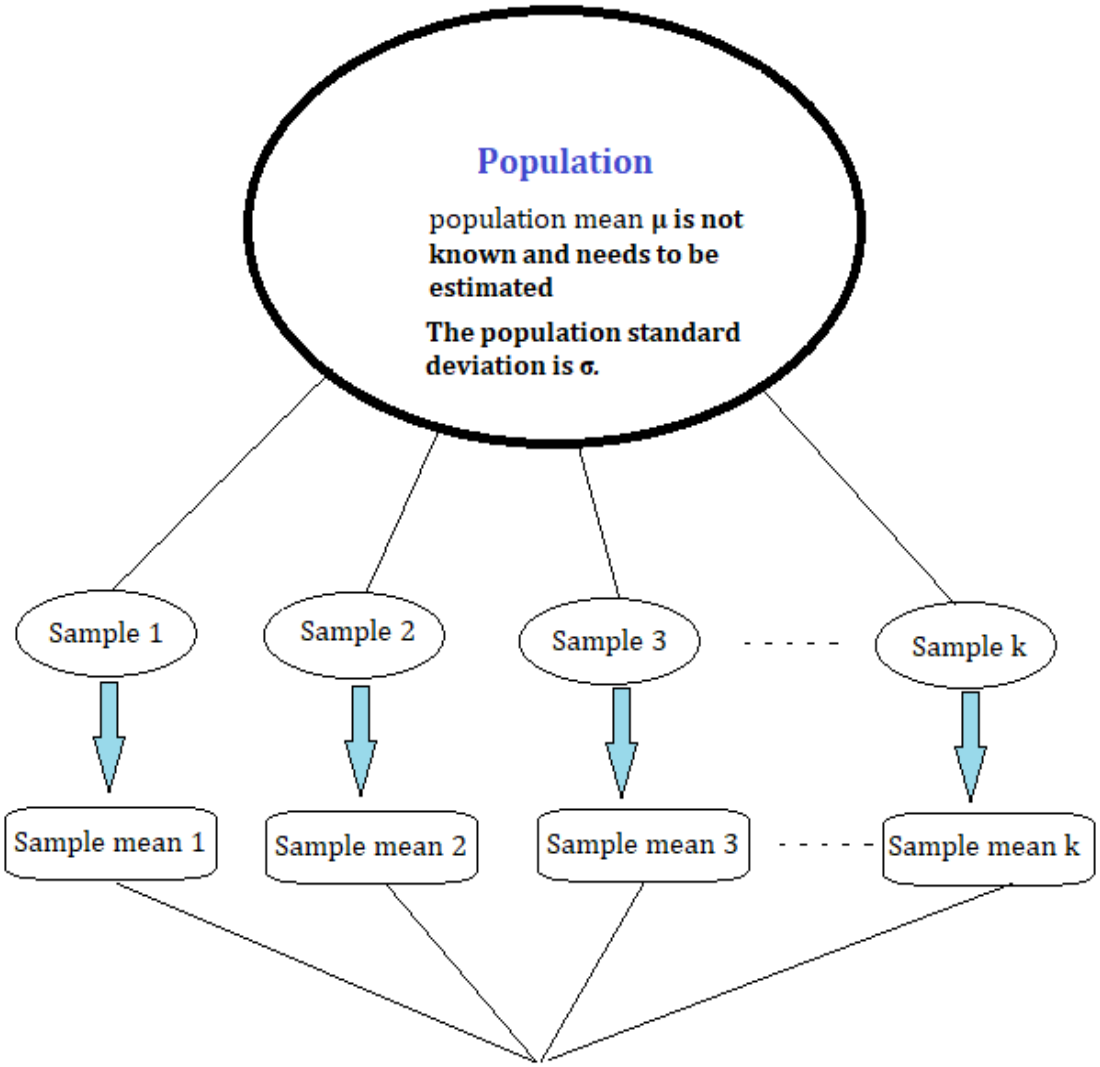
We can see that the average rating from a sample will vary depending on the batch of people who give the rating. In other words, the average rating ($\underline{x}$) **is a random variable itself**, and like any other random variable has a distribution of its own, which is called its **sampling distribution**.

7.3 THE CENTRAL LIMIT THEOREM (CLT)

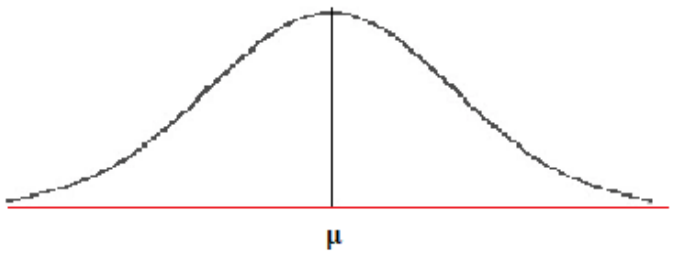
CLT is arguably one of the most important theorems in all of statistics. We have learnt that the sample mean itself is a random variable. CLT states that irrespective of the distribution of the

original population, **for a large enough sample size (n), the sample mean $\underline{x}$ is normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.**

The following schematic diagram summarizes the whole process:



Sample mean ($\bar{x}$) is a random variable itself and for large enough sample size(n), according to CLT, has a normal distribution with its mean as the population mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.



Probability distribution of sample mean $\bar{x}$

7.4 Implications of CLT

CLT has two major implications:

- Using CLT we can now make use of the properties of normal distribution to define confidence intervals (a range of values) for the unknown population mean. We shall delve into more details on this in the EPAT course.

- Knowing that the probability distribution of the sample mean is normal, allows us to design statistical tests for testing hypotheses about the distribution of the population. This is what we will discuss in the next chapter.