# Lecture 27: Introduction to Statistics

Satyajit Thakor IIT Mandi

#### Statistical inference

- ► Till now we studied <u>probability theory</u>, i.e., random variables through their distributions.
- ► In practice, often an experimenter has no knowledge of distribution.
- A task of the experimenter is to find out as much information as possible about the distribution.
- This is done through experimentation and the collection of a data set relating to the random variable.
- Statistical inference is the science of deducing properties of an underlying probability distribution from such a data set.

### Population and sample

- A population consists of all possible observations available from a particular probability distribution.
- Example: let X be the weight of a milk container produced at a particular dairy. Then, weight of each container is the population.
- A sample is a particular subset of the population that an experimenter measures and it is used to investigate the unknown distribution.
- Example: a sample or data set for "the weight of a milk container" is obtained by weighing the contents of n containers.
- ▶ A <u>random sample</u> is one in which the elements of the sample are chosen <u>at random</u> from the population, and this procedure is often used to ensure that the sample is representative of the population.
- ightharpoonup Example: if we choose the first n containers produced then this does not provide random sample.

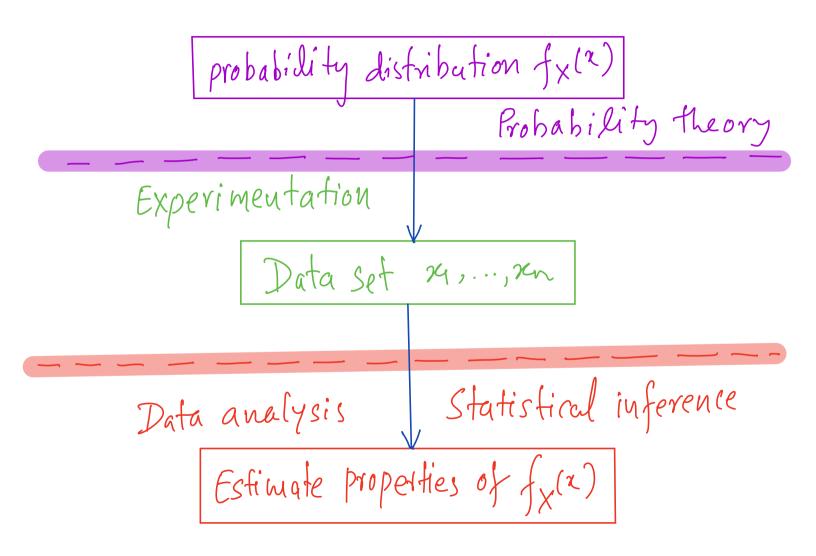
## Population and sample

- The PDF  $f_X(x)$  provides complete information about the probabilistic properties of the random variable X and is unknown to the experimenter.
- The experimenter proceeds by obtaining a sample of observations of the random variable X, which may be written

$$x_1, x_2, \dots, x_n$$
 sample / data set

An appropriate analysis of the data gives the experimenter some information about  $f_X(x)$ .

# Probability theory & statistical inference



## Sample mean

- ➤ The sample mean of a data set is simply the arithmetic average of the data observations.
- ▶ That is, if a data set consists of the n observations  $x_1, \ldots, x_n$ , then the sample mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The sample mean  $\bar{x}$  can be thought of as being an estimate of the expectation of the unknown underlying probability distribution of the observations in the data set.

## Sample variance

The sample variance of a set of data observations  $x_1, \ldots, x_n$  is defined as

$$s^{2} = \frac{\sum_{i=1}^{n} (x_{i} - \bar{x})^{2}}{n-1}.$$

- $\triangleright$  The sample standard deviation is s.
- Why the denominator of the formula for  $s^2$  is chosen as n-1 and not n? -We will see this in the next lecture.

#### Parameter and statistics

- A parameter is a property of a population or a probability distribution.
- For example, the PDF of a population of r.v. X is  $f_X(x)$  and  $\mu_X$  is a parameter.
- ► A statistic is a property of a sample from the population.
- For example, suppose that a sample of size n is collected of observations from a particular probability distribution  $f_X(x)$ . The data values recorded,  $x_1, \ldots, x_n$ , are the observed values of a set of n random variables  $X_1, \ldots, X_n$ , and each has the probability distribution  $f_X(x)$ .

### Parameter and statistics

- In general, a statistic is any function  $g(X_1, \ldots, X_n)$  of these random variables.
- ► For example, the sample mean

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

and the sample variance

$$S^{2} = \frac{\sum_{i=1}^{n} (X_{i} - \bar{X})^{2}}{n-1}$$
 are statistics.

For a given data set  $x_1, \ldots, x_n$  these statistics take the observed values

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ and } s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}.$$