**Numerical Summary of Data**

**Central Tendency:** Central tendency refers to the typical or central value around which data points tend to cluster. It gives an idea of the “average” or typical value in a dataset. The most common measures of central tendency are mean, median, and mode.

**Mean:** The arithmetic average of all the values in a dataset. It is calculated by adding up all the values and dividing by their number of values.

**Median:** The middle value in a dataset when they are arranged in ascending or descending order. If there is an even number of values, then it is the average of the two middle values.

**Mode:** The value that appears most frequently in a dataset.

**Dispersion:** Dispersion measures how spread out the values in a dataset are from the central value (i.e., the measure of central tendency). It provides information about how much variability or spread there is among data points.

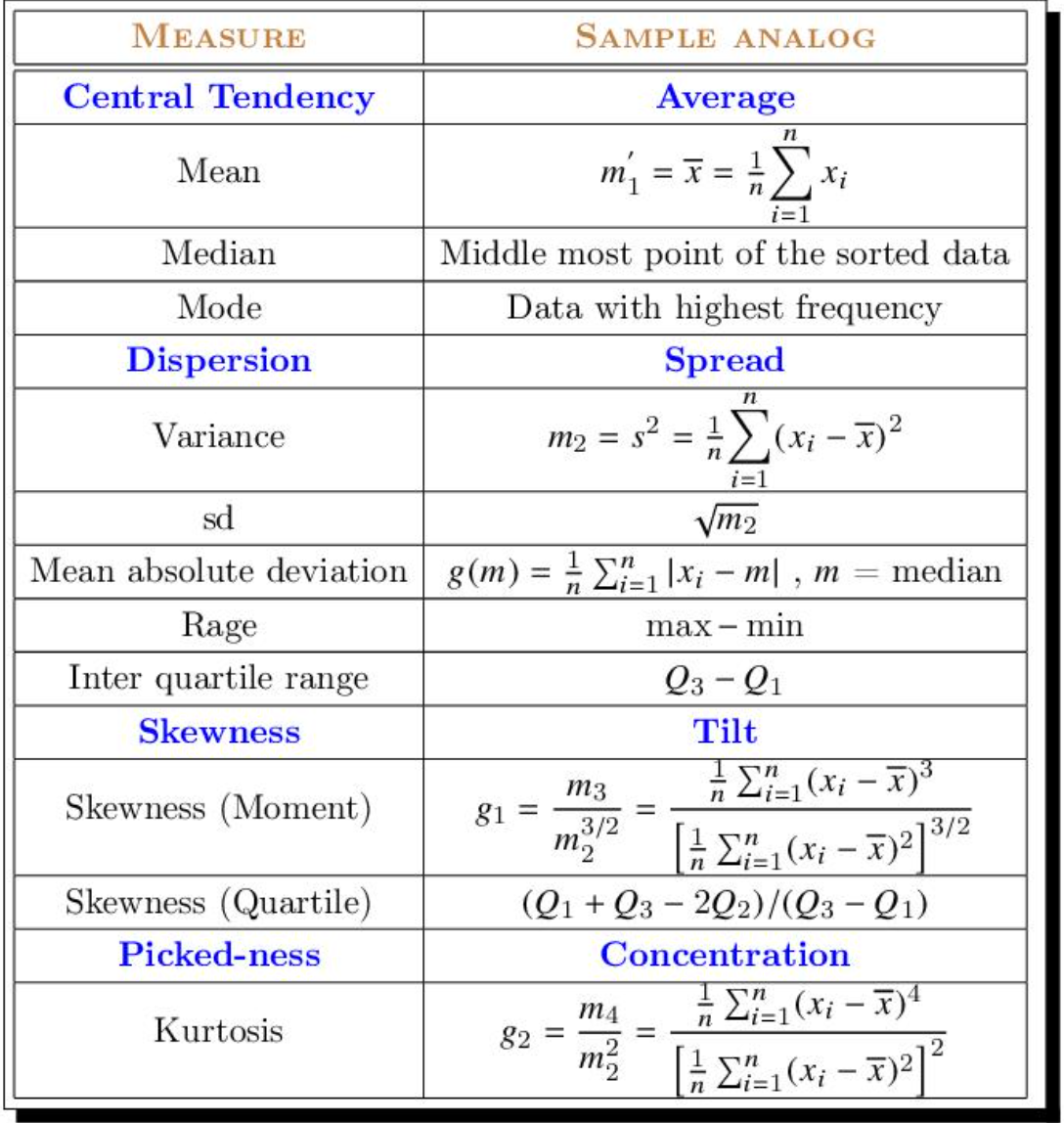
**Range:** The difference between the maximum and minimum values in a dataset.

**Variance:** A measure of how much the values vary from their mean. It is calculated by taking an average of squared differences between each data point and the mean.

**SD (Standard Deviation):** The square root of variance.

**IQR (Interquartile Range):** The range between the first quartile (25th percentile) and third quartile (75th percentile) of a dataset. It represents the summary spread for the middle 50% of the dataset.

Central Tendency formulae



Probability Distributions

<https://www.youtube.com/watch?v=YXLVjCKVP7U&list=PLTNMv857s9WVzutwxaMb0YZKW7hoveGLS>

### 4. Probabilistic Characterization of data

In statistics, an empirical distribution function (commonly also called an empirical cumulative distribution function, eCDF) is the distribution function associated with the empirical measure of a sample. This cumulative distribution function is a step function that jumps up by 1/n at each of the n data points. Its value at any specified value of the measured variable is the fraction of observations of the measured variable that are less than or equal to the specified value.

ChatGPT Explanation:

Empirical Distribution Function (eCDF)

Definition:

An empirical distribution function (eCDF) is a way to understand how data points are distributed in a sample.

How it works:

Imagine you have a list of numbers (data points) that you've collected.

The eCDF shows the percentage of data points that are less than or equal to any given value.

Step Function:

The eCDF is like a staircase that goes up by a small step each time you encounter a data point.

If you have nnn data points, the eCDF will increase by 1/n (a small step) for each data point.

Value at a Point:

To find the eCDF value at a specific point xxx, you look at what fraction of your data points are less than or equal to xxx.

For example, if you have 10 data points and 3 of them are less than or equal to 5, the eCDF value at 5 is 3/10 or 0.3.

Example

Suppose you have the following data points: [2, 3, 5, 7, 11]

Step 1: Order the data points (already ordered here).

Step 2: Calculate the fraction of data points less than or equal to each value:

At 2: 1 out of 5 data points ≤ 2 → eCDF(2) = 1/5 = 0.2

At 3: 2 out of 5 data points ≤ 3 → eCDF(3) = 2/5 = 0.4

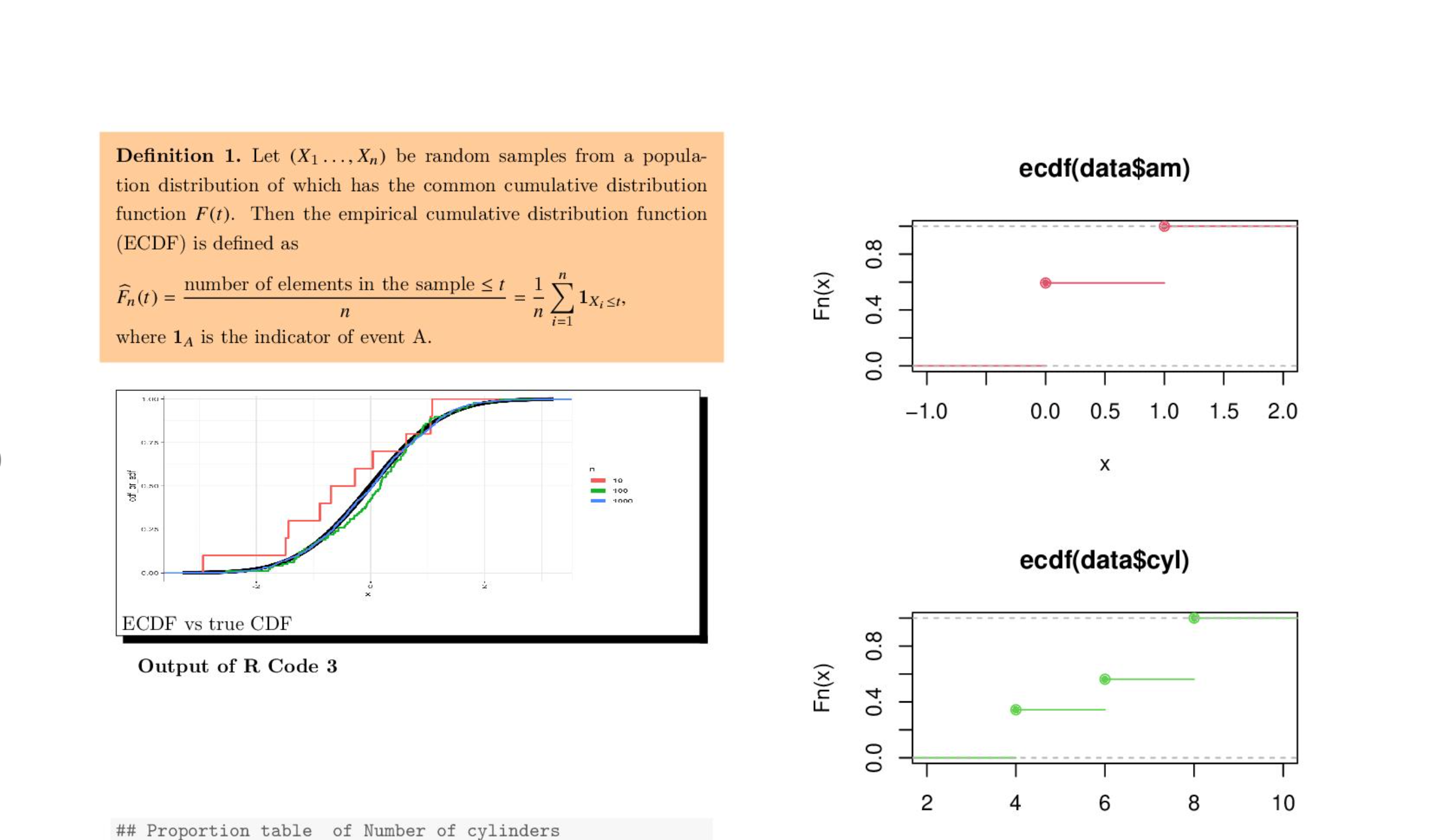
At 5: 3 out of 5 data points ≤ 5 → eCDF(5) = 3/5 = 0.6

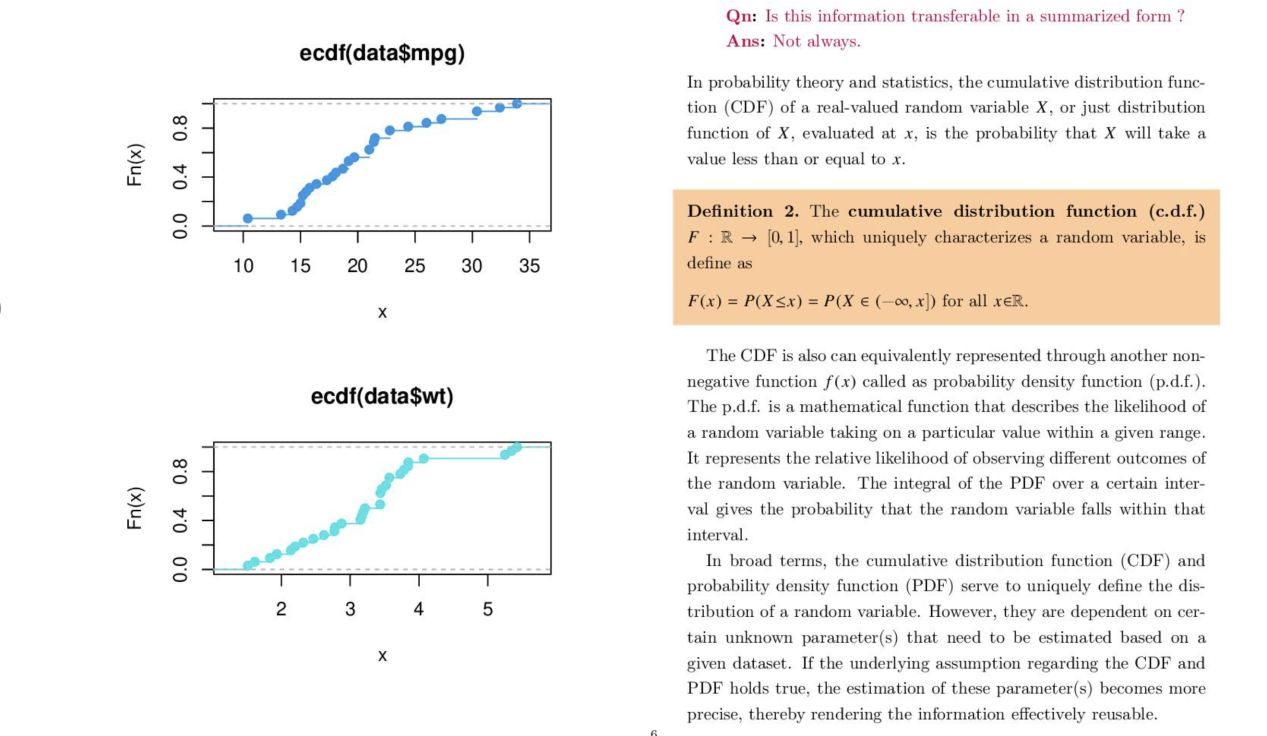
At 7: 4 out of 5 data points ≤ 7 → eCDF(7) = 4/5 = 0.8

At 11: 5 out of 5 data points ≤ 11 → eCDF(11) = 5/5 = 1.0

Summary

The eCDF helps you see the distribution of your data by showing how many data points fall below or at each value. It builds a "staircase" graph that steps up at each data point, giving you a visual representation of the cumulative distribution.





### Explanation

#### Cumulative Distribution Function (CDF)

* **Definition**:
  + The CDF of a real-valued random variable X is a function that gives the probability that XXX will take a value less than or equal to xxx.
  + Mathematically, the CDF F(x)is defined as:
  + This means for any value x, F(x) tells us how likely it is that X is less than or equal to x.

#### Properties of the CDF

**Uniqueness**:

The CDF uniquely characterizes a random variable. This means that the CDF contains all the information about the distribution of the random variable.

It can be defined for all

#### Probability Density Function (PDF)

**Relation to CDF**:

The PDF is another way to represent the distribution of a random variable, but it focuses on the likelihood of the variable taking on a specific value.

The PDF f(x)is a non-negative function that describes the relative likelihood of different outcomes.

The CDF can be obtained by integrating the PDF: F(x)=∫−∞xf(t) dt

**Properties**:

The integral of the PDF over a certain interval gives the probability that the random variable falls within that interval.

#### Estimation of Parameters

**Dependence on Parameters**:

Both the CDF and PDF depend on certain unknown parameters that describe the distribution (e.g., mean, variance).

These parameters need to be estimated from data.

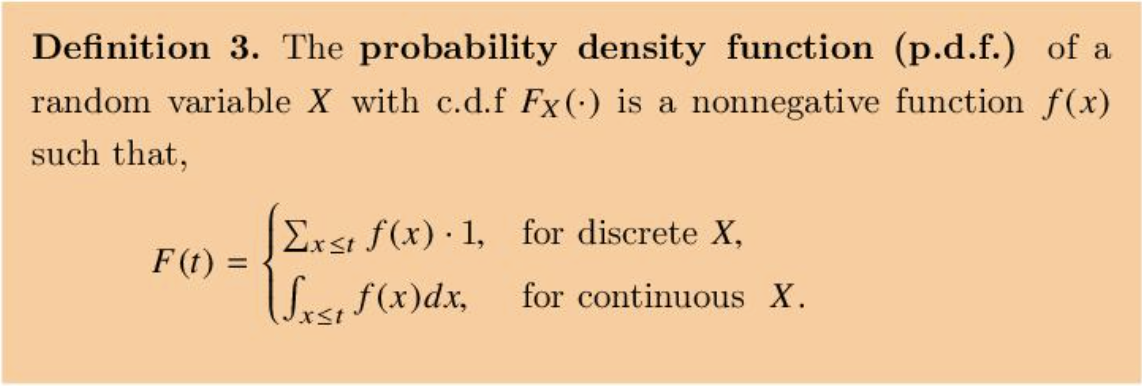
**Transferability**:

If the assumptions about the CDF and PDF hold true, the estimation of these parameters becomes more precise.

This makes the information effectively reusable in a summarized form. However, if the assumptions do not hold, the information may not be accurately transferable.

### Summary

In summary, the CDF and PDF are two fundamental concepts in probability and statistics that describe the distribution of a random variable. The CDF gives the probability that the variable takes on a value less than or equal to a specific value, while the PDF describes the relative likelihood of the variable taking on specific values. Both functions are characterized by certain parameters that need to be estimated from data. The reliability and transferability of the summarized information depend on the accuracy of these estimates and the validity of the underlying assumptions.



**ESTIMATION**

**BUDDHANANDA BANERJEE**

Let x=(x1,x2,⋯ ,xn)be the observed/realized values of a set of i.i.d. random variables X=(X1,X2,⋯ ,Xn) where Xi∼iidfθ for some θ∈Θ. Here a family of distributions is denoted by F={f(x∣θ)∣θ∈Θ} or {F(x∣θ)∣θ∈Θ}

**Explanation**: This is describing a set of data points x=(x1,x2,…,xn) that are observed from a set of independent and identically distributed (i.i.d.) random variables X=(X1,X2,…,Xn) each following a probability distribution fθ with a parameter θ. The family of all such distributions is represented by F.

**Parametric Estimation**: In a parametric inference problem it is assumed that the family of the distribution is known but the particular value of the parameter is unknown. We estimate the value of the parameter θ as a function of the observations x. The ultimate goal is to approximate the p.d.f. fθ​ or Fθ​ through the estimation of θ itself. Parametric estimation has two aspects, namely,

**Point estimation** (a) Definition of an estimator (b) Good properties of an estimator (c) Methods of estimation (MME and MLE)

**Interval estimation** (a) Definition of confidence interval (b) Construction of confidence interval

**Explanation**: In parametric estimation, we know the form of the distribution but not the exact parameter values. We aim to estimate the parameter θ using the observed data. Parametric estimation involves:

**Point Estimation**: Finding a single best guess of θ\thetaθ.

* Defining what an estimator is.
* Identifying good properties of estimators (e.g., unbiasedness, efficiency).
* Using methods like the Method of Moments (MME) and Maximum Likelihood Estimation (MLE).

**Interval Estimation**: Finding a range of values within which θ\thetaθ likely lies.

* Defining confidence intervals.
* Constructing these intervals.

**Definition 1. Statistic**: A statistic is a function of random variables and it is free from any unknown parameter. Being a (measurable) function, T(X) say, of random variables it is also a random variable.

**Explanation**: A statistic is a summary of the data that does not depend on unknown parameters. For example, the sample mean or sample variance.

**Definition 2. Estimator**: If the statistic T(X) is used to estimate a parametric function g(θ) then T(X) is said to be an estimator of g(θ). And a realized value of it for X=x, i.e. T(x) is known as an estimate of θ. We often abuse the notation as g(θ^)=T(x) and g(θ^)=T(X) which are understood from the context.

**Explanation**: An estimator is a statistic used to estimate a parameter. When we apply the estimator to our actual data, the result is called an estimate.

### 1. ****Properties****

**Definition**

**3. Unbiased estimator**: An estimator T(X) is said to be an unbiased estimator of a parametric function g(θ) if E(T(X)−g(θ))=0∀θ∈Θ.

**Explanation**: An estimator is unbiased if its expected value equals the true parameter value. This means, on average, the estimator hits the target.

**Remark 1**: It does not require T(x)=g(θ)T(x) = g(\theta)T(x)=g(θ) to hold or it may hold with probability zero.

**Explanation**: This remark clarifies that an unbiased estimator does not have to exactly equal the true parameter every time; it just needs to be correct on average.

**Definition 4. Asymptotically unbiased estimator**: Denoting Tn=T(X1,X2,⋯ ,Xn) an estimator Tn​ is said to be asymptotically unbiased of g(θ) if lim⁡n→∞Bg(θ)(Tn)=lim⁡n→∞E(Tn−g(θ))=0

**Explanation**: An estimator is asymptotically unbiased if it becomes unbiased as the sample size grows to infinity. This means any bias in the estimator diminishes as we collect more data.