Probability Distribution Function and Cumulative Distribution Function:

Random Variable:

Random variable is a variable, which takes any outcome from a random experiment.

Example:
1. Rolling a Dice:
X= {1,2,3,4,5,6}
X is a random variable

2. Toss a coin:
Y={Head, Tail}
Y is a random variable

Random variable is of two types.
Discrete Random Variable
Continuous Random Variable

Discrete Random Variable:

A random variable, which can take one value from a finite set of values is called Discrete Random Variable.

Example:

Rolling a Dice - X={1,2,3,4,5,6} – It can take one of the value from a discrete set.

Continuous Random Variable:

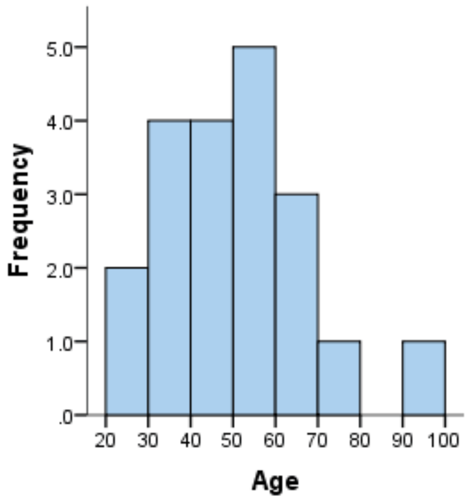
A random variable, which can take any real value is called Continuous Random Variable.

Example:
Human Height – Y – It can take any real value between 120 cm to 190 cm.

Histogram:

To construct a histogram from a continuous variable we need to split the data into intervals, called **bins**. In the below example, **age** has been split into bins, with each bin representing a 10-year period starting at 20 years.

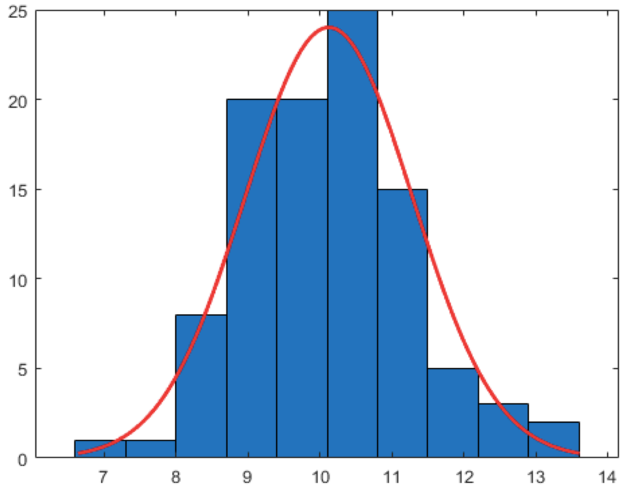
Bin	Frequency	Scores Included in Bin
20-30	2	26,27
30-40	4	33,37,32,35
40-50	4	43,42,48,45
50-60	5	51,54,53,58,52
60-70	3	61,68,62
70-80	1	75
80-90	0	-
90-100	1	92



In a histogram, it is the area of the bar that indicates the frequency of occurrences for each bin.

PDF: Probability Density Function:

PDF is a smoothed form of Histogram.



Red Curve: PDF

By using Kernal Density Estimation, we can smooth the histogram.

For a continuous random variable (which takes value over a finite or infinite), we can model the distribution using its PDF.

Example: the petal length of a particular flower could take any value between a reasonable interval.

Let the petal length of all the flowers is between 10 mm and 30 mm.

It makes no sense to ask the probability that the petal length of the flower is 22 mm.

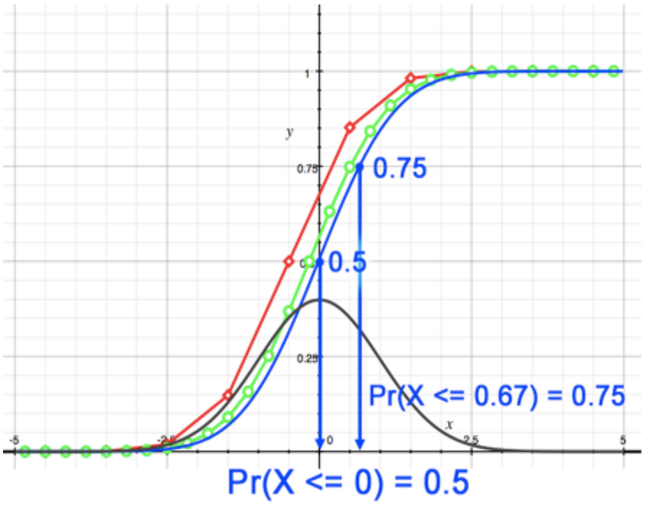
We can calculate the probability of petal length of the flower between 22 mm to 24 mm. The probability is given by the area under the curve between 22 to 24 mm in the PDF.

However, we can define PDF as the probability of random variable falling within a particular range of values, as opposed to taking on any one value. The function that gives this probability density is referred to as the probability density function or PDF.

CDF: Cumulative Distribution function:

A CDF is defined over any interval where the PDF is defined.
Suppose a PDF is defined over the interval [a,b].
Let $a<c<b$
Then the CDF over the interval [a,c] is obtained by accumulating the value of PDF for all values in the interval [a,c]. Typically, this is obtained by integrating the PDF from a to c.

Cumulative Distribution Function of a random variable X is defined as
F(x) = P (X <= x)



Black Curve: PDF
Blue line: CDF

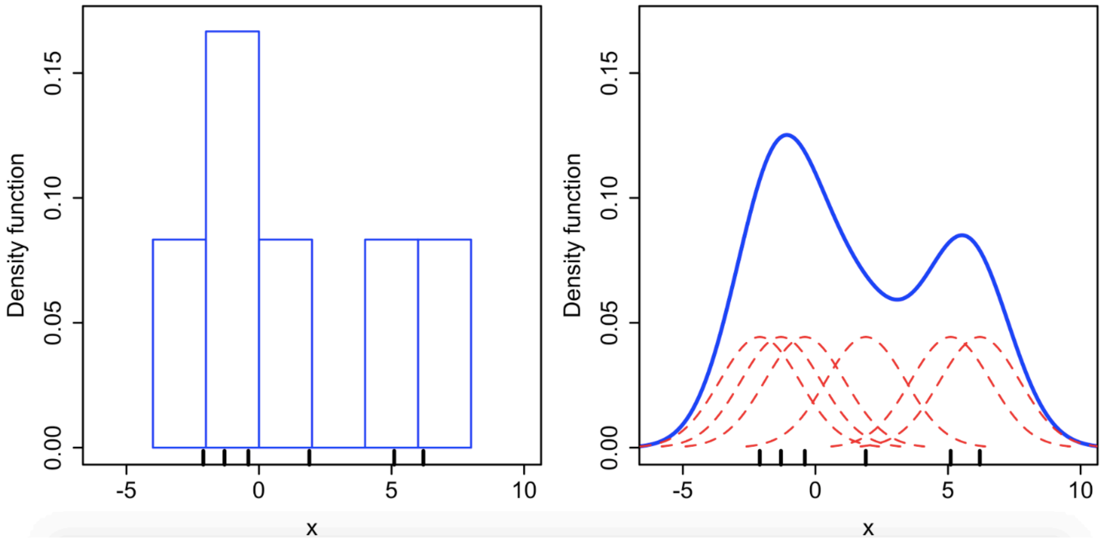
Note:

PDF is not probability but probability density, while **CDF** is really probability.
CDF: Integration of PDF
PDF: Derivative of CDF

Kernal Density Estimation:

Kernal Density Estimation is a fundamental data smoothing problem.
We can smoothen the histograms to draw a PDF by using Kernal Density Estimation.

Example:



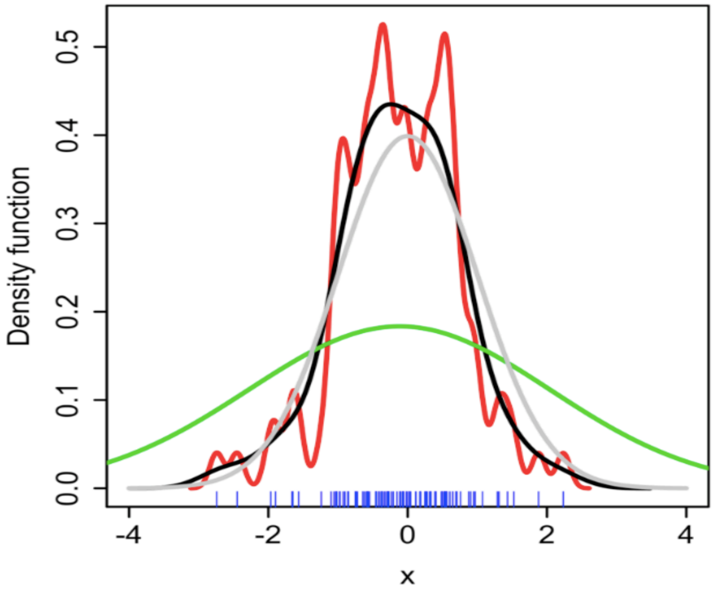
In the above example, we have 6 data points.
Lets take each of the point and draw a bell curve or Gaussian Kernal (red dashes) centered around each point.

Height of the PDF at every point at x= Sum of heights of each of the kernels present at x.

We will get all the heights of the PDF for all the points at x. This is called as Kernal Density Estimation using Gaussian Kernals.

Mean of the Gaussian Kernal is the height of the Kernal.
Variance of the Gaussian Kernal is called as Bandwidth.

Bandwidth selection:



Grey Curve: Actual PDF – True Density (Standard Normal)
Green Curve: KDE with Bandwidth - 2
Black Curve: KDE with Bandwidth – 0.337
Red Curve: KDE with Bandwidth – 0.05

If we will choose our Kernals as very wide (High Bandwidth), our PDF will look like the Green Curve. The green curve is oversmoothed.

If we will choose Medium Bandwidth, we can get a smoother curve like Black Curve.

If we will choose our Kernals as very narrow (Medium Bandwidth, our PDF will look like the Red curve. The Red Curve is undersmoothed.

References:

https://statistics.laerd.com/statistical-guides/understanding-histograms.php

https://www.mathworks.com/help/stats/histfit.html

https://www.scratchapixel.com/lessons/mathematics-physics-for-computer-graphics/monte-carlo-methods-mathematical-foundations/pdf-and-cdf

https://en.wikipedia.org/wiki/Kernel_density_estimation