When analyzing data, it's important to understand the distribution of the data. The distribution refers to how the data is spread out or clustered around certain values or ranges. Lets know about data distribution below…..

What is a distribution function?

In statistical terms, a distribution function is a mathematical expression that describes the probability of different possible outcomes for an experiment.

Data Types

At a higher level, we have **Qualitative** and **Quantitative** data. And in **Quantitative** data, we have **Continuous** and **Discrete** data types.

Continuous data is measured and can take any number of values in a given finite or infinite range. It can be represented in decimal format. And the random variable that holds continuous values is called the Continuous random variable.(It's used in **Probability Density Function**)

Examples: A person's height, Time, distance, etc.

Discrete data is counted and can take only a limited number of values. It makes no sense when written in decimal format. And the random variable that holds discrete data is called the Discrete random variable.(It's used in **Probability Mass Function**)

Example: The number of students in a class, number of workers in a company, etc.

Types of distribution functions:

Based on the types of data we deal with, we have two types of distribution functions.

For discrete data, we have discrete distributions; and for continuous data, we have continuous (**density**) distributions.

Discrete distributions	Continuous distributions
Uniform distribution	Normal distribution
Binomial distribution	Standard Normal distribution
Bernoulli distribution	Student's T distribution
Poisson distribution	Chi-squared distribution

Before deep-diving into the types of distributions, it is important to revise the fundamental concepts like Probability Density Function (PDF), Probability Mass Function (PMF), and Cumulative Density Function (CDF).

Probability Density Function (PDF):

The types of probability density function are used to describe distributions like continuous uniform distribution, normal distribution, Student t distribution, etc. The probability density function gives the probability that the value of a random variable will fall between a range of values. A cumulative distribution function and the probability density function are used to describe a continuous distribution.

Probability Density Function Definition

A probability density function can be defined as a function that gives the likelihood of occurrence of a random variable between a given interval of values. To determine this probability, the probability density function has to be integrated between the two specified limits. Suppose the probability that a random variable, X, lies between points a and b has to be determined then the general formula is given as follows:

Probability Density Function

$$F(x) = P(a \leq x \leq b) = \int_a^b f(x)dx \geq 0$$

There are two important properties followed by a probability density function, f(x). These are given as follows:

- f(x) ≥ 0. The probability density function for all real numbers will always be positive.
- The total area under the probability density curve has to be equal to 1.

Cumulative Distribution Function (CDF)

It is another method to describe the distribution of a random variable (either continuous or discrete).

$$F_X(x) = P(X \leq x)$$

$F_X(x)$ = function of X

X = real value variable

P = probability that X will have a value less than or equal to x

Probability Mass Function (PMF)

It is a statistical term that describes the probability distribution of a discrete random variable.

$$p(x) = P(X = x)$$

The probability of x = the probability(X = one specific x)

• Discrete distributions

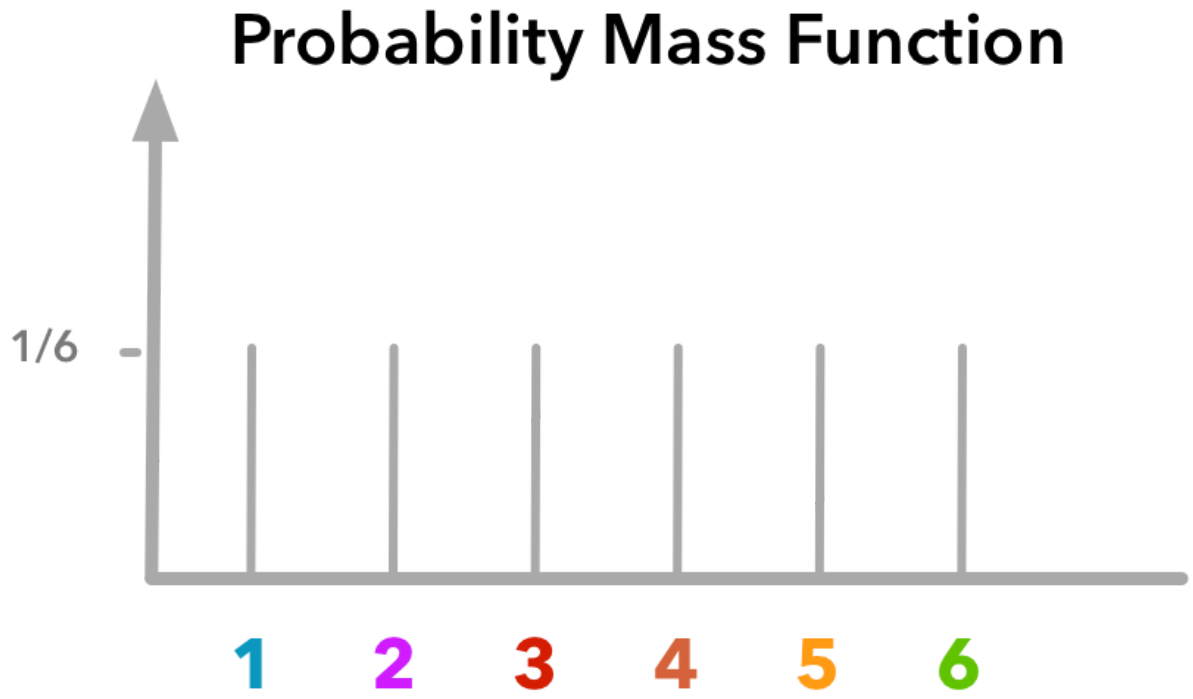
Let us start with the easiest one – Uniform distribution.

Discrete Uniform distribution (U)

It is denoted as X ~ U (a, b). And is read as X is a discrete random variable that follows uniform distribution ranging from a to b.

Uniform distribution is when all the possible events are equally likely. For example, consider an experiment of rolling a dice. We have six possible events X = {1, 2, 3, 4, 5, 6} each having a probability of P(X) = 1/6.

The PMF graph of the above experiment is:



The formula for PMF, CDF of Uniform distribution function are:

Support	$k \in \{a, a+1, \ldots, b-1, b\}$
PMF	$\frac{1}{n}$
CDF	$\frac{\lfloor k \rfloor - a + 1}{n}$

The Mean and Variance of Uniform distribution are:

Mean = (a+b)/2

Variance = (n^2 - 1)/12

Binomial distribution (B):

Binomial distribution is a discrete probability distribution of the number of successes in ‘n’ independent experiments sequence. This type of distribution is used when there are only two possible outcomes for each trial, such as Success/Failure, Pass/Fail/, Win/Lose, etc.

It is denoted as X ~ B(n, p). And is read as X is a discrete random variable that follows Binomial distribution with parameters n, p.
Where n is the no. of trials, and **p** is the success probability for each trial.

Generally, the outcome success is denoted as 1, and the probability associated with it is p.
And Failure is denoted as 0, and the probability associated with it is q = 1-p.

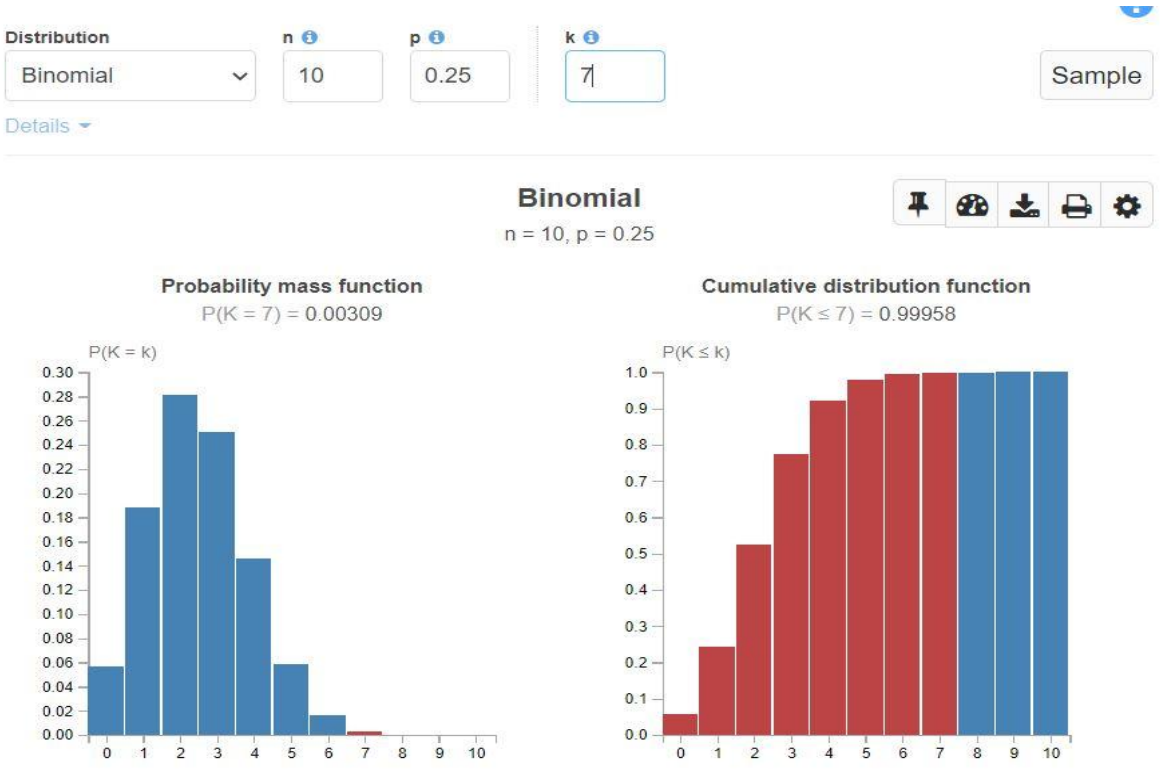
The formula for PMF, CDF of Binomial distribution are:

PMF	$\binom{n}{k} p^k q^{n-k}$
CDF	$I_q(n-k, 1+k)$

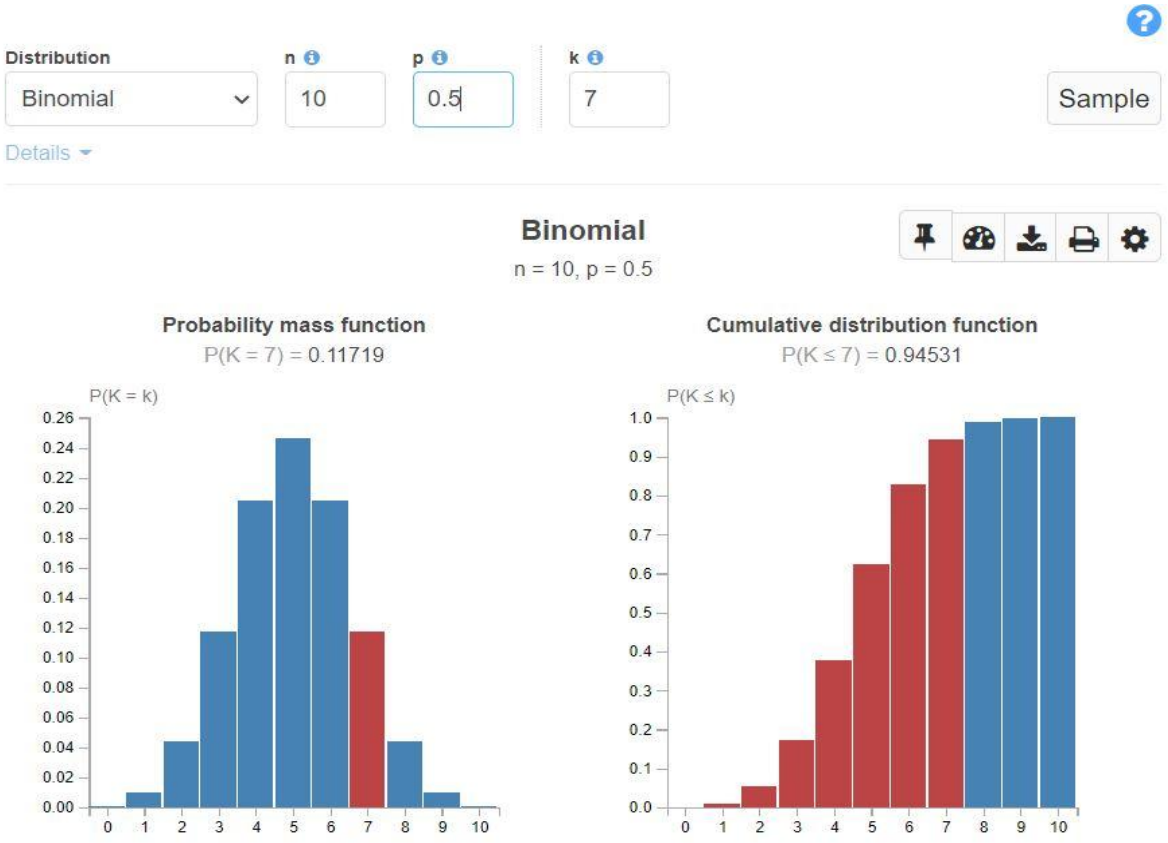
K in the above formula is the number of successes.

The mean and variance of a binomial distribution are given as:
Mean = np
Variance = npq

Now consider, we ran a Binomial experiment 10 times, and the probability of success = 0.25. Below are how PMF, CDF look.



PMF, CDF for a Binomial experiment with Probability of success = Probability of failure look like below.



Bernoulli distribution (Bern):

It is denoted as X ~ Bern(p). And is read as X is a discrete random variable that follows Bernoulli distribution with parameter p.

Where p is the probability of success.

Bernoulli can be represented as a Binomial experiment with a single trial.
X ~ Bern(p) —-> X ~ B(1, p)

The formula for PMF, CDF of Bernoulli distribution is:

PMF	$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$ $p^k(1-p)^{1-k}$
CDF	$\begin{cases} 0 & \text{if } k < 0 \\ 1 - p & \text{if } 0 \leq k < 1 \\ 1 & \text{if } k \geq 1 \end{cases}$

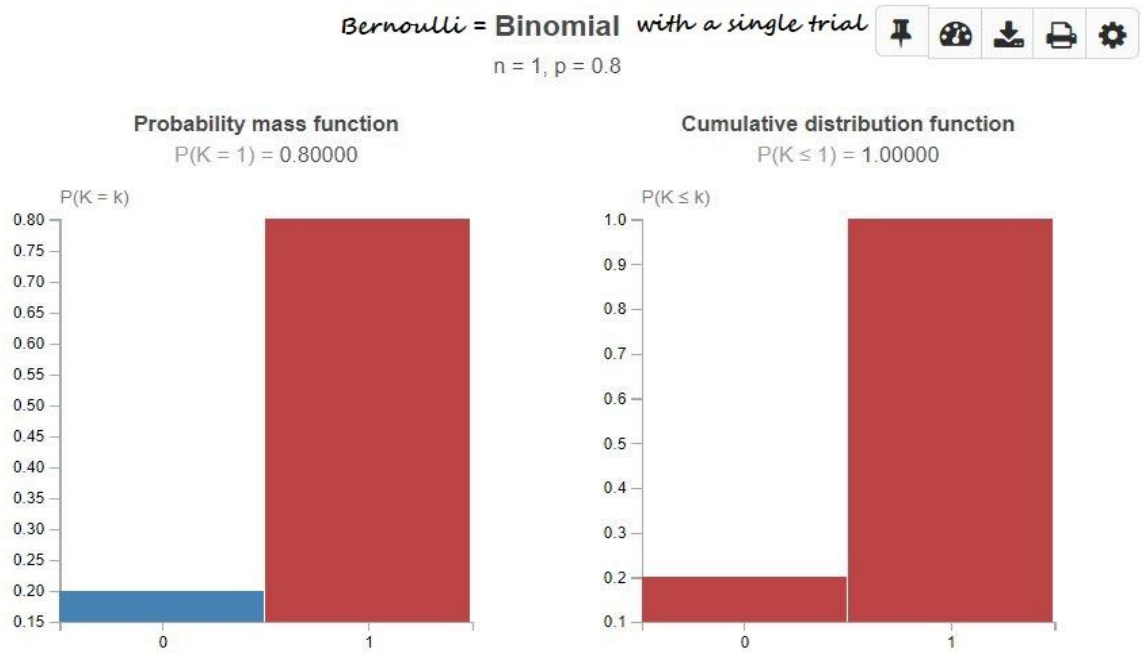
The Mean and Variance of Bernoulli distribution are given as:

Mean = p

Variance = p(1-p) = pq

Example: Consider an example of tossing a fair. The two possible outcomes are Heads, Tails. The probability (p) associated with each of them is 1/2.

If we take an unfair coin, the probability associated with each of them need not be 1/2. Heads can have a probability of p = 0.8, then the probability of tail q = 1-p = 1-0.8 = 0.2



Bernoulli's event suggests which outcome can be expected for a single trial. Whereas, a Binomial event suggests the no. of times a specific outcome can be expected.

Poisson Distribution (Po):

Poisson Distribution is a discrete probability distribution function that expresses the probability of a given number of events occurring in a fixed time interval when its rate is known but its exact timing cannot be predicted accurately enough to measure it directly. This type of distribution is useful for modeling random occurrences such as customer arrivals at stores, phone calls received by call centers etc, where the average rate of occurrence is known but the exact timing cannot be measured.

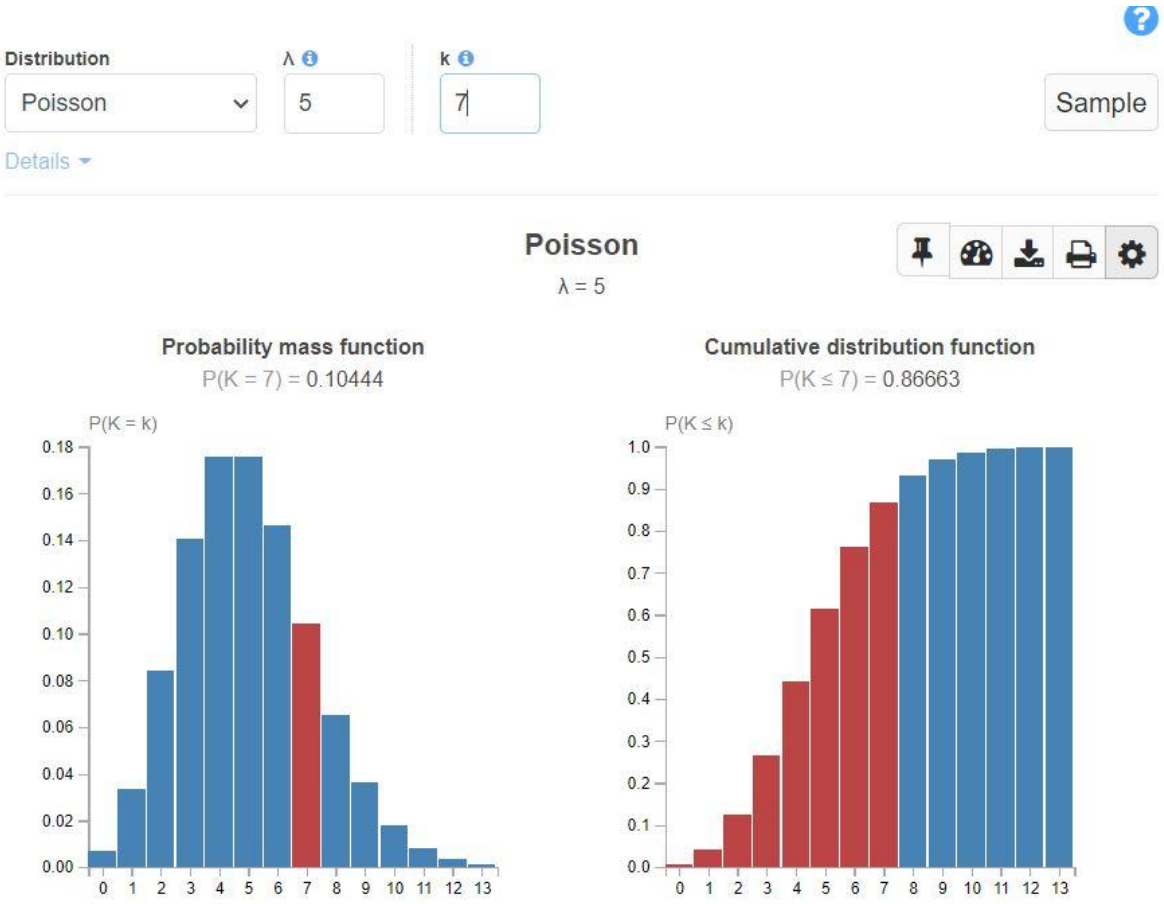
The formula for PMF, CDF of poisson distribution are:

PMF	$\frac{\lambda^k e^{-\lambda}}{k!}$
CDF	$e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$

The Mean and Variance of Poisson distribution are given as:

Mean = Variance = λ

A Poisson distribution with λ = 5 look like below



• Continuous Distributions

Normal or Gaussian Distribution (N)

Normal Distributions are one of the most commonly used data distributions. This distribution measures data points in a bell-shaped curve, with an equal number of data points to the left and right of the mean value. Normal Distributions can be used to predict future outcomes based on past trends.
Examples: Heights of people, exam scores of students, IQ Scores, etc follows Normal distribution.

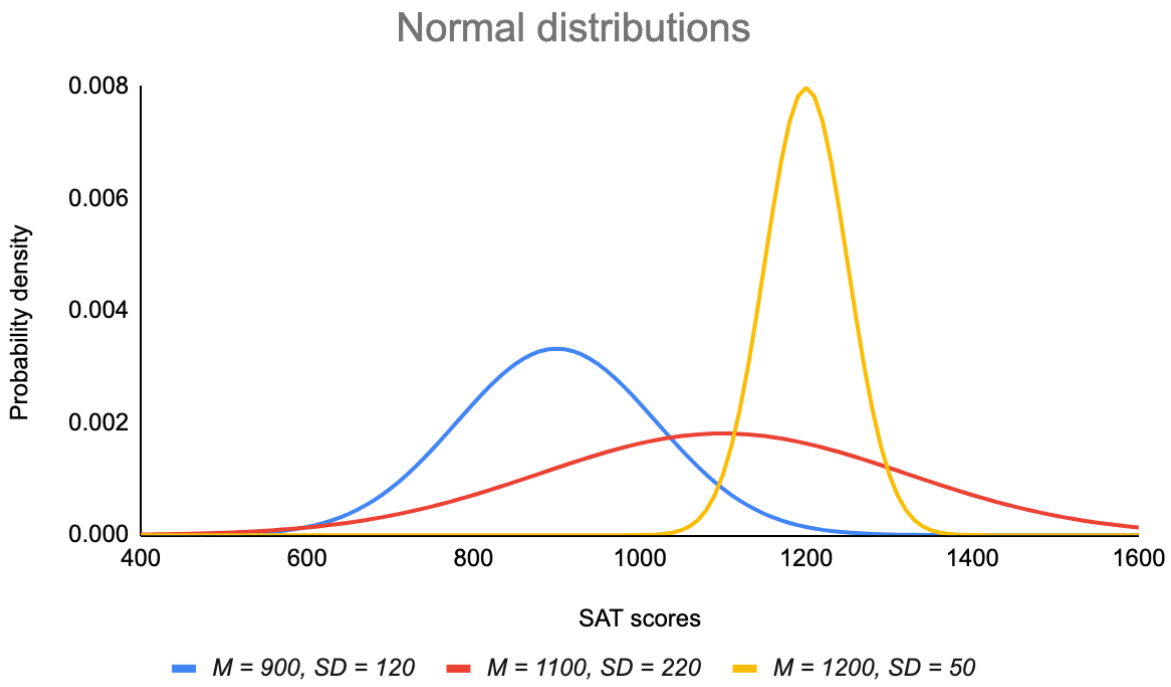
It is denoted as X ~ N (μ, σ2). And is read as X is a continuous random variable that follows a Normal distribution with parameters μ, σ2.

Where μ is the mean, and σ2 is the variance. Mean, Variance together talks about shape statistics.

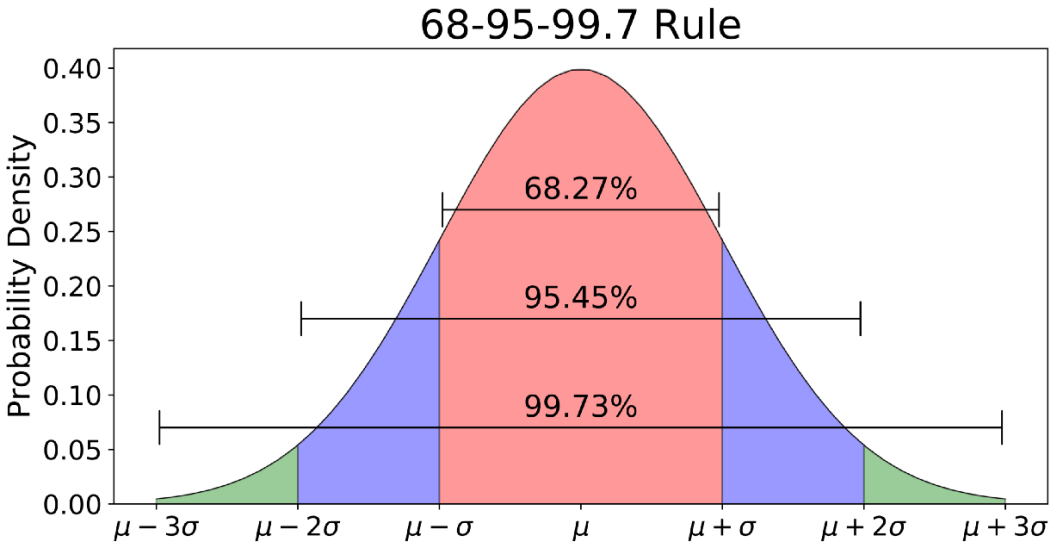
Properties of Normal distribution:

- The random variable takes values from -∞ to +∞
- The probability associated with any single value is Zero.
- looks like a bell curve and is symmetric about x=μ. 50% of data lies on the left-hand side and 50% of the data lies on the right-hand side.
- The area under the curve (AUC) = 1
- All the measures of central tendency coincide i.e., mean = median = mode

A normal distribution with different means, standard deviations look like below:



Normal distribution follows the 68-95-99.7 rule. This rule is also known as the empirical rule. According to it, 68% of data lies in the first standard deviation range, 95% of data lies in the second standard deviation range, and 99.7% of data lies in the third standard deviation range.

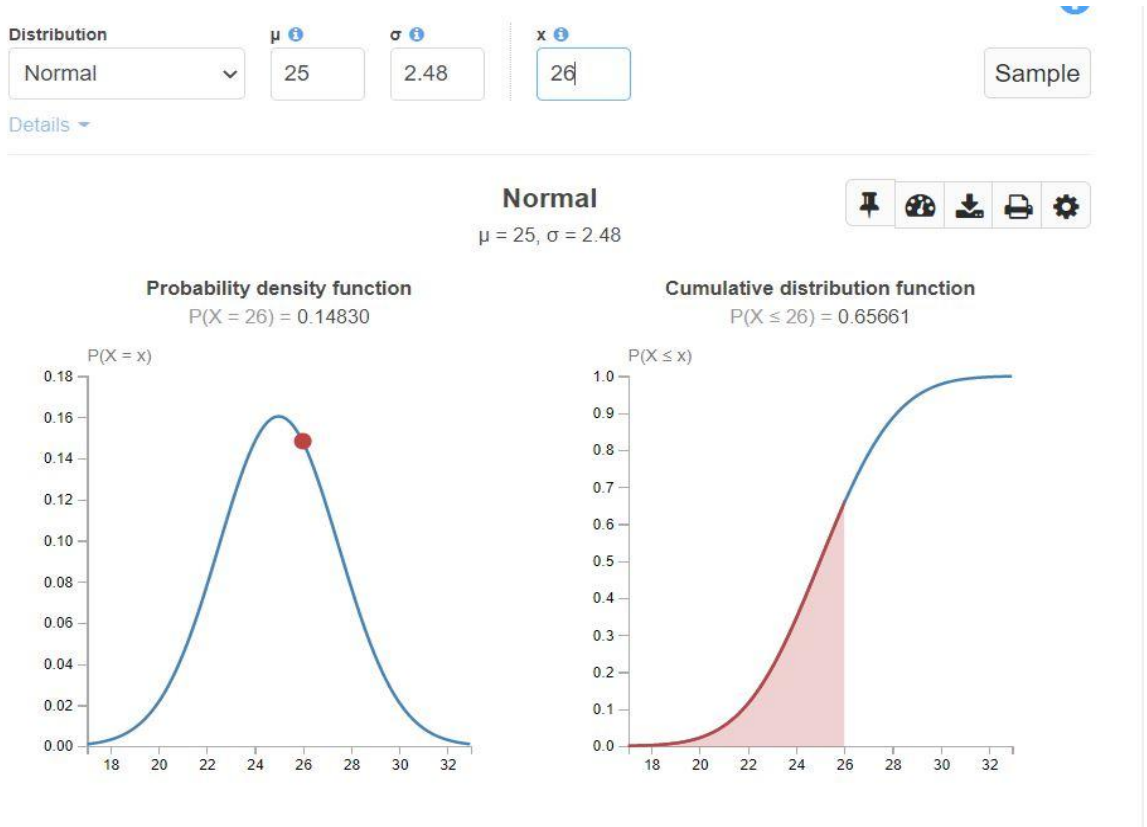


The formula for PDF, CDF of the normal distribution are:

PDF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
CDF	$\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$

The Mean and Variance of a Normal distribution are given as:
Mean = μ
Variance = σ2

Let's assume we have a height distribution with mean = 25, standard deviation = 2.48. Below is how the graph looks like.



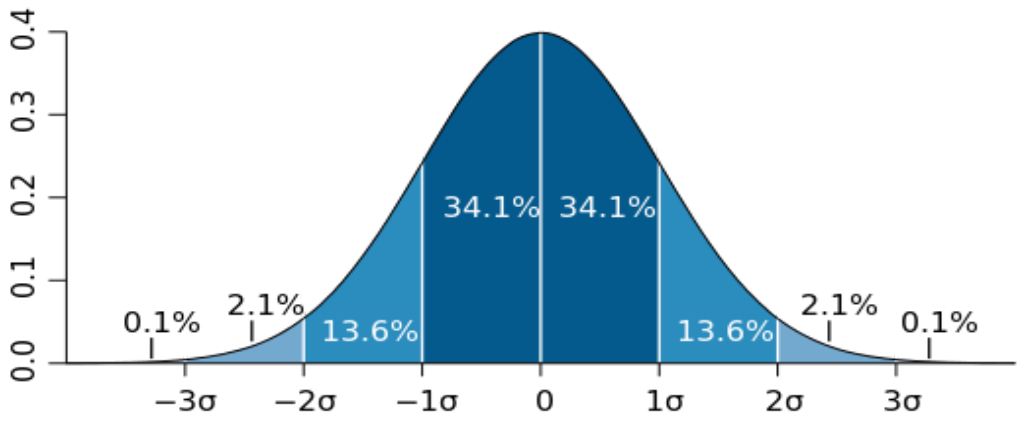
Standard Normal Distribution or SND:

It is a transformation of Normal distribution in such a way that Mean = 0, and standard deviation 1.

Transformation is a way in which we alter every element of distribution to get a new distribution with similar characteristics.

It is denoted as Z ~ N(0, 1). And is read as X is a continuous random variable that follows Normal distribution with mean 0 and variance 1.

All the properties of a Normal distribution will be satisfied by a Standard Normal distribution.



And in addition, there exists a table that summarizes the most commonly used values of a CDF of s Standard Normal Distribution. This table is known as a Z-score table.
The formula for standardisation is Z = (X-μ)/σ

The formula for PDF, CDF of Standard Normal distribution are given as:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}, \qquad \text{for all } z \in \mathbb{R}.$$

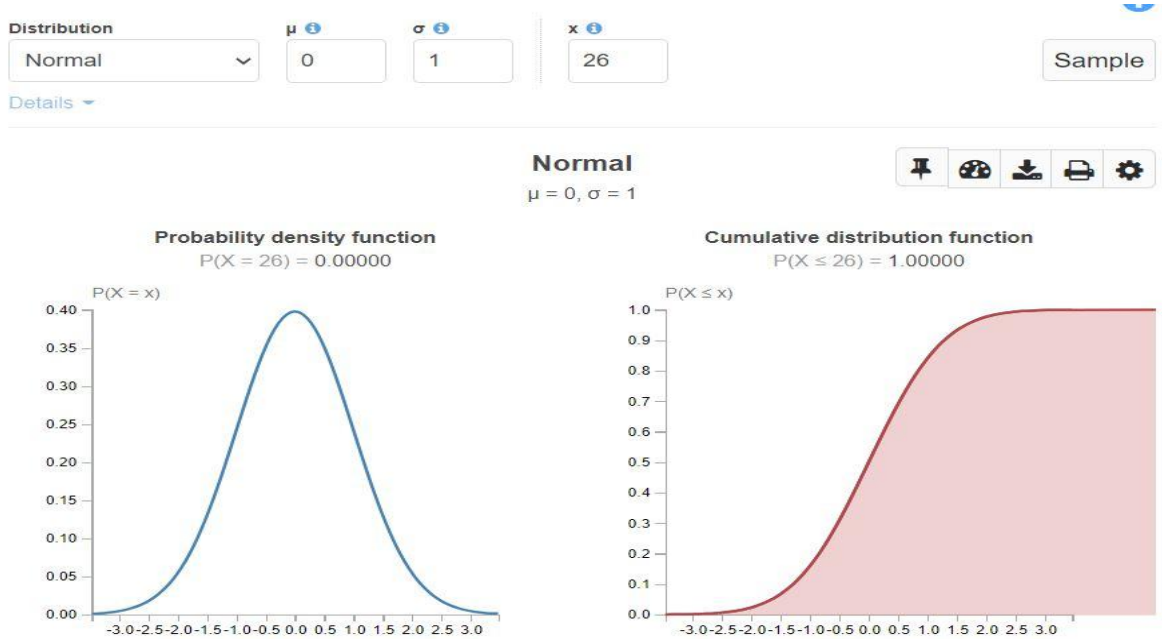
$$\Phi(x) = P(Z \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{u^2}{2}\right\} du.$$

The Mean and Variance of Standard Normal distribution are:

Mean = 0

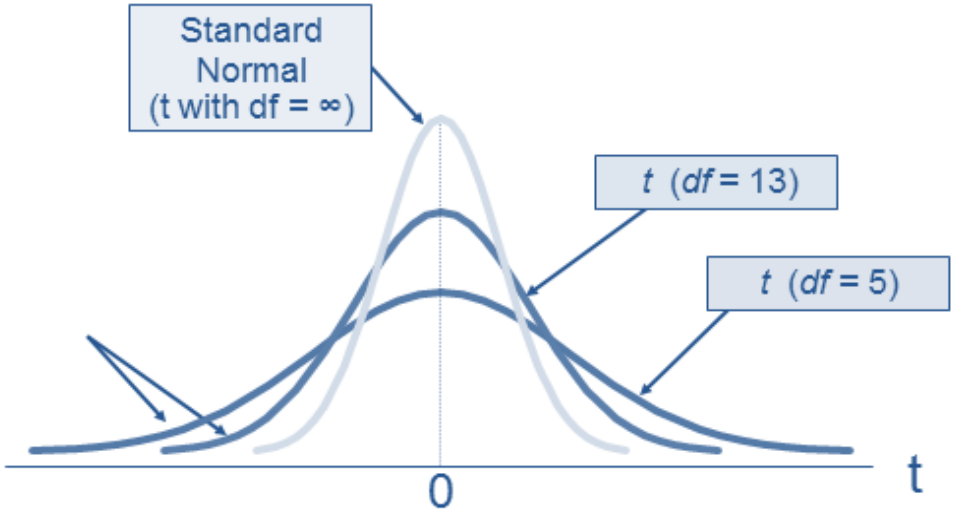
Variance = 1

Normal distribution with mean 0 and variance 1 (SND) looks like below:



Student's T distribution or t-distribution (t)

It is denoted as X ~ t(k). And is read as X is a continuous random variable that follows Student's T distribution with parameter k.
where k is the degrees of freedom. If the sample size is n, then k = n-1.
A student (t) distribution is used when the sample size is very small and a normal distribution cannot be used. As the degrees of freedom (k) increase, t distribution tends to become Standard Normal distribution.



The formula for PDF, CDF of t-distribution are:

PDF	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2}+x\Gamma\left(\frac{\nu+1}{2}\right)\times \frac{{}_2F_1\left(\frac{1}{2},\frac{\nu+1}{2};\frac{3}{2};-\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}$

(v in above formulae is degrees of freedom)
t-distribution can be used in Hypothesis testing (to test if there is any significant difference between two sample means), calculating confidence intervals with population standard deviation is unknown.

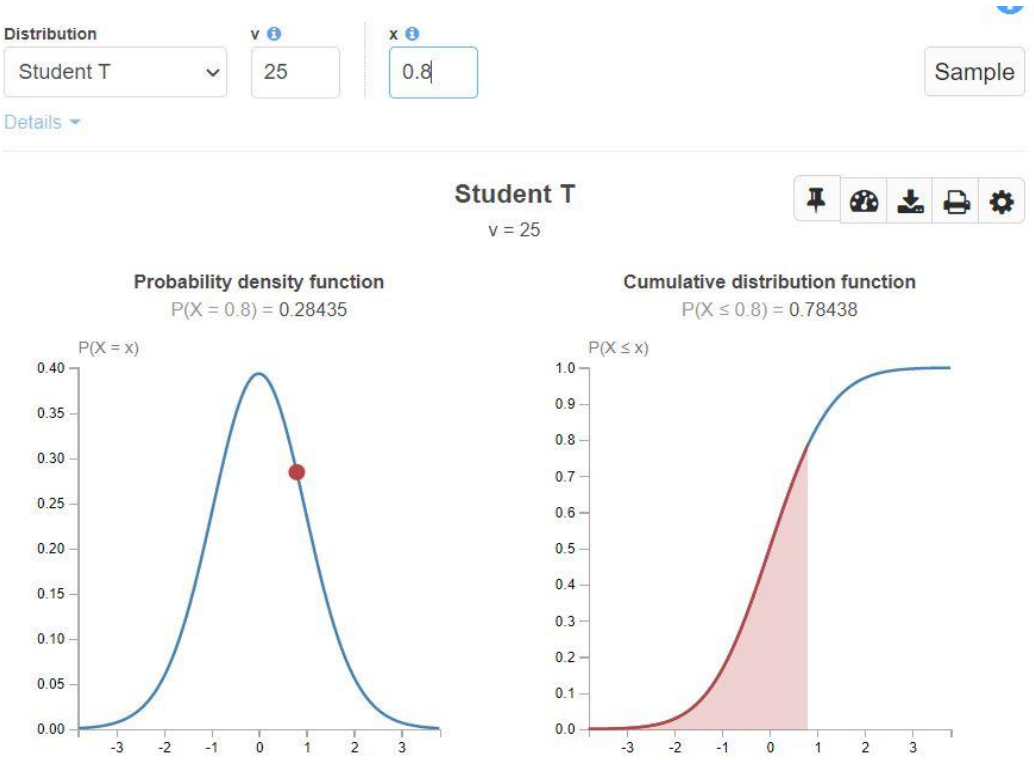
Like standard normal distribution, t-distribution also has a table of its own. This table is known as the t-table.

The mean and variance of Student's T distribution are:

Mean = 0

variance = k/(k-2)

A student's t-distribution with degrees of freedom = 25 looks like below:



Chi-Square distribution

It is denoted as X~χ2(k). And is read as X is a continuous random variable that follows Chi-Square distribution with k degrees of freedom.

It is used in Hypothesis testing, computing confidence intervals, and for the goodness of fit.

It is a transformation of t-distribution. Finding the t-distribution to the power of 2 gives Chi-Square distribution and finding the square root of Chi-Square of distribution gives us t-distribution.

Chi-Square distribution has a chi-square table.

The formula for PDF, CDF of Chi-square distribution are:

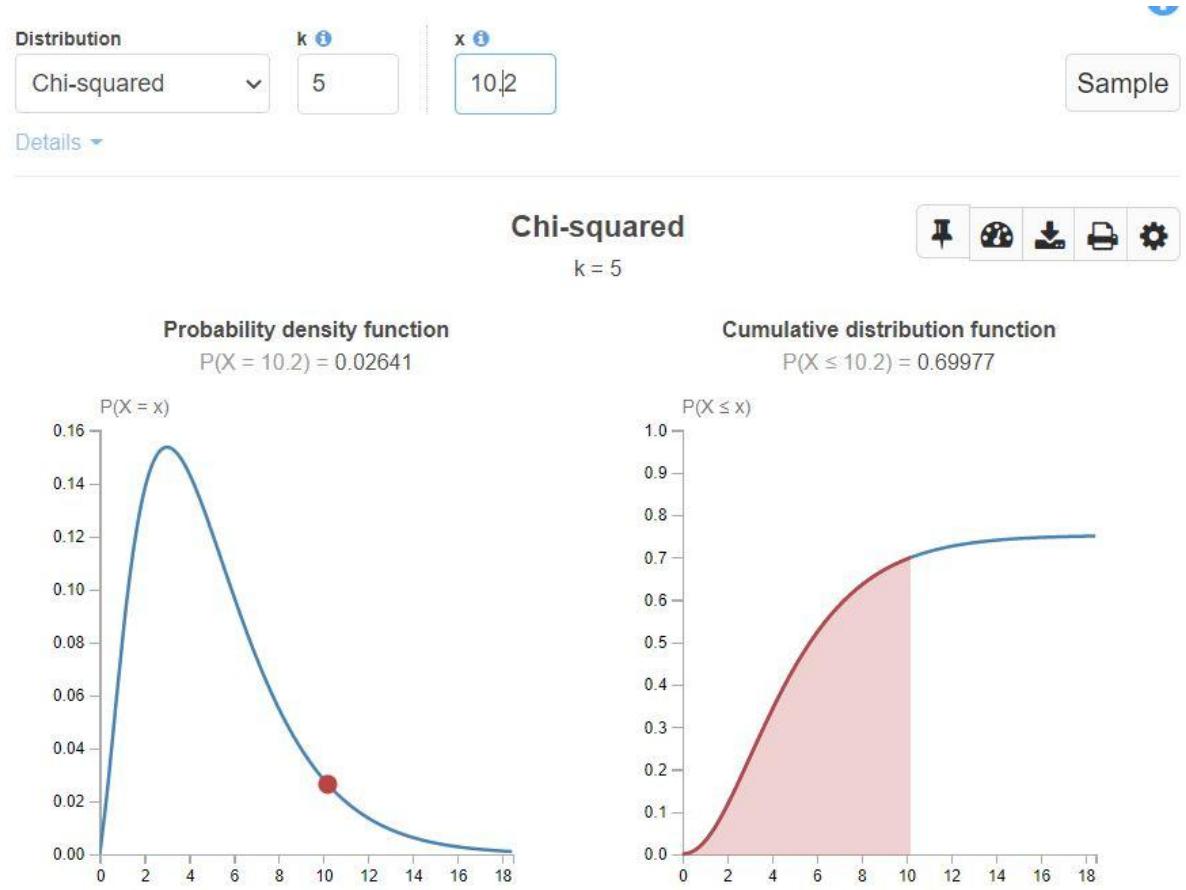
PDF	$\frac{1}{2^{k/2}\Gamma(k/2)}\, x^{k/2-1} e^{-x/2}$
CDF	$\frac{1}{\Gamma(k/2)}\, \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$

The mean and variance of Chi-square distribution are:

Mean = k

Variance = 2k

A Chi-square distribution with degrees of freedom = 5 looks like below.



● summary

1. There are two main types of data distribution in statistics: continuous and discrete.
2. Probability Mass Function (PMF): The PMF is a discrete function that assigns probabilities to specific values of a discrete random variable. It shows the likelihood of each value occurring and is often represented as a table or a graph.
3. Density Function: The density function is a version of the distribution function that allows for continuous data. It represents the likelihood of a random variable taking on different values and is often depicted as a smooth curve.
4. Cumulative Distribution Function (CDF): The CDF gives the probability that a random variable is less than or equal to a specific value. It accumulates the probabilities of all values up to the given point and ranges from 0 to 1.

These functions are used to understand and describe the characteristics of data distributions, helping us analyze and interpret data patterns and make informed decisions.