**Distribution in statistics**

**Random Experiment**

An experiment whose outcome cannot be predicted with certainty in advance and can be repeated any number of times under the same conditions is called a random experiment

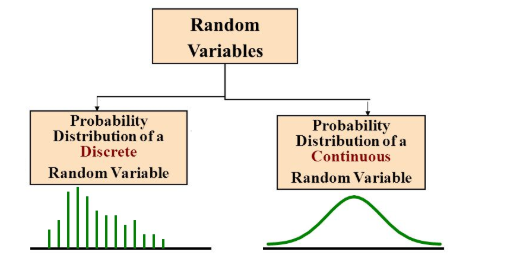
Eg: Tossing a coin, Rolling a die

**Random Variable**

A random variable is a variable whose values are outcomes of a random experiment or random process. Random variables could be either discrete or continuous.

Eg: Suppose in a coin toss experiment we consider getting head as success and tail as failure. Then the random variable X can be defined as:

* X = 1 if the coin lands heads (H).
* X = 0 if the coin lands tails (T).

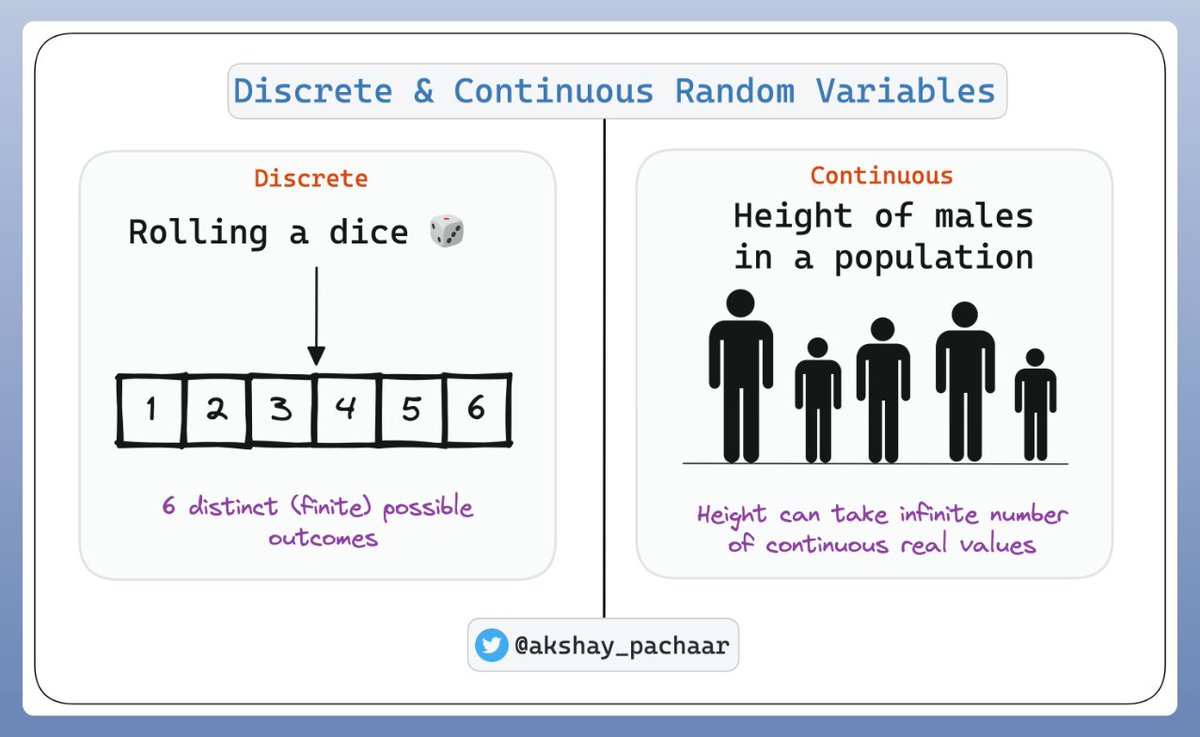


A random variable is said to be discrete if it can take only a finite number of distinct values such as 0, 1, 2, 3, 4, … and so on.

Eg: Number of students in a class, Outcome of tossing a coin, Outcome of rolling a dice

A random variable is said to be continuous if it can take on any value within a continuous range, often represented by an interval.

Eg: The height of people in a population, Temperature at a location



Video link: [Random variables](https://youtu.be/lHCpYeFvTs0?si=LguLd-rNBmpXxtUI)

**Statistical distribution**

A statistical distribution / probability distribution is a way to describe how data is spread or distributed in a dataset. It tells us the likelihood of different values or outcomes occurring. A distribution helps us to understand a variable by giving us an idea of the values that the variable is most likely to obtain.

Probability Distribution is a statistical function which links or lists all the possible outcomes a random variable can take, in any random process, with its corresponding probability of occurrence.

The probability distribution is divided into two: - discrete and continuous distribution

1. **Discrete distributions**

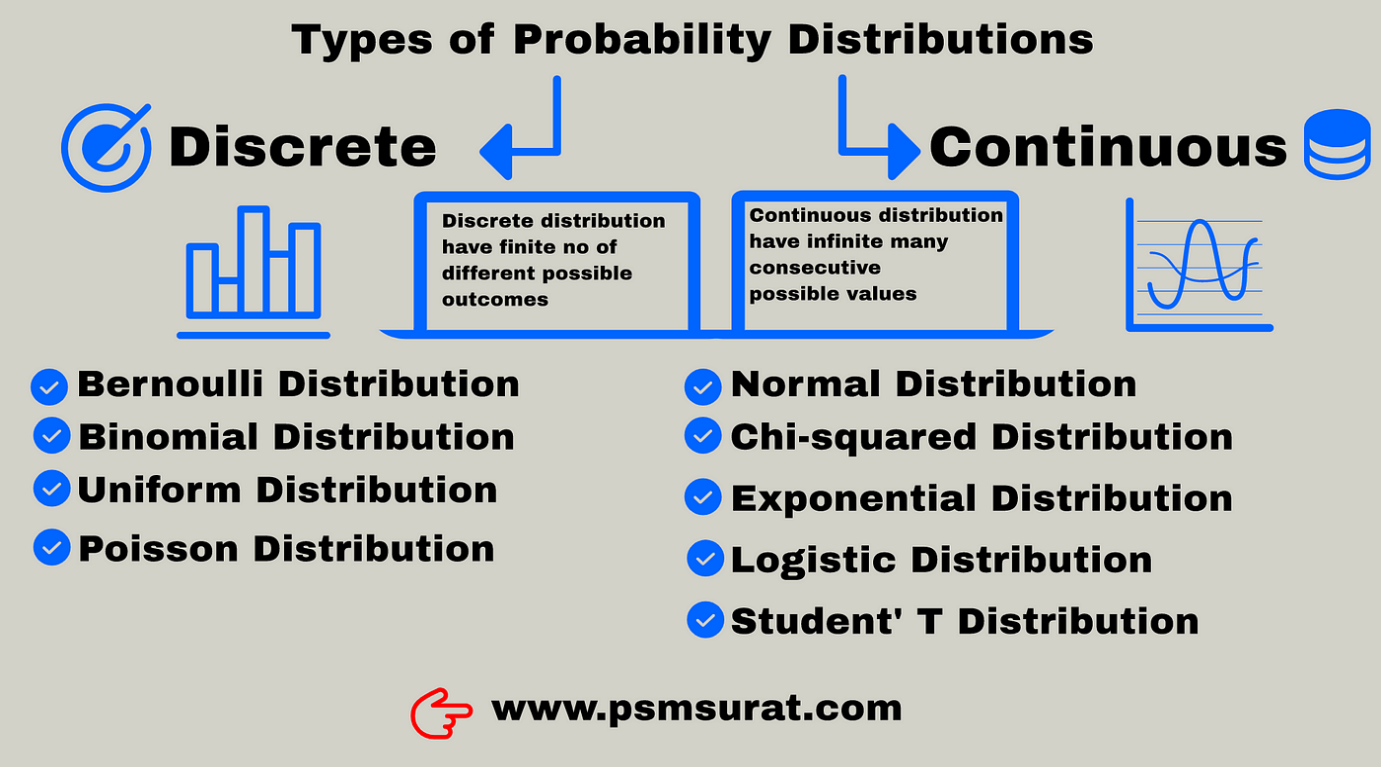
Discrete distributions model the probabilities of random variables that can have discrete values as outcomes. For example, the possible values for the random variable X that represents the number of heads that can occur when a coin is tossed are the set {0, 1} and not any value from 0 to 1 like 0.1 or 0.9.

Examples: Bernoulli, Binomial, Negative Binomial, Hypergeometric, etc.,

1. **Continuous distributions**

Continuous distributions model the probabilities of random variables that can have any possible outcome. For example, the possible values for the random variable X that represents weights of citizens in a town which can have any value like 34.5, 47.7, etc.,

Examples: Normal, Student’s T, Chi-square, Exponential, etc.,

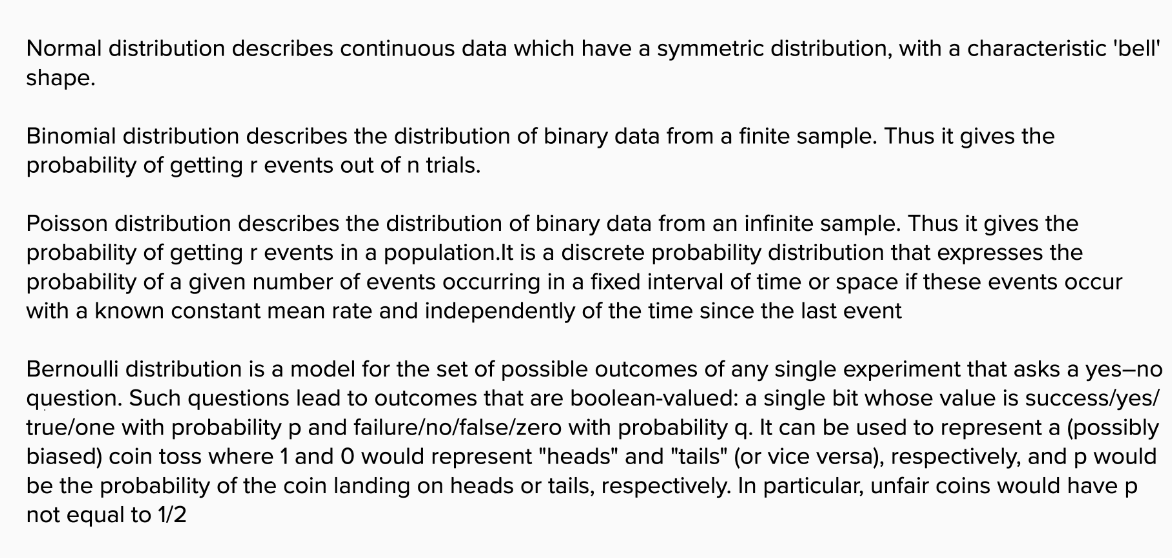


Video link: [Introduction to probability and distribution](https://youtu.be/mOD7DiJwxf8?si=AuBPyWapNeqHU6vI)

**Types of distributions**

There are different types of probability distributions used in statistics to describe how data or random variables are distributed. The most frequently used distributions in our project are:

* Normal distribution
* Binomial distribution
* Poisson distribution
* Bernoulli distribution

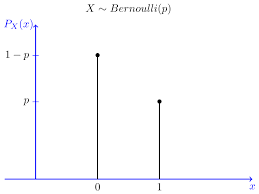


Video link: [Types of distributions](https://youtu.be/Xg7ng3-Pm-8?si=vljw7aoblnW6jwWj)

**Bernoulli Distribution**

Bernoulli Distribution is a type of discrete probability distribution where every experiment conducted asks a question that can be answered only in yes or no. If in a Bernoulli trial the random variable takes on the value of 1, it means that this is a success. The probability of success is given by p. Similarly, if the value of the random variable is 0, it indicates failure. The probability of failure is q or 1 – p.

For example, suppose there is an experiment where you flip a coin that is fair. If the outcome of the flip is heads, then you will win. This means that the probability of getting heads is p = 1/2. If X is the random variable following a Bernoulli Distribution, we get P (X = 1) = p = 1/2.

****

A discrete random variable is said to follow the Bernoulli distribution with parameter p if its pdf is given by:

f (x, p) = px (1 - p)1 - x, x ϵ {0, 1}, 0<p<1, p+q=1

= 0, elsewhere

**Mean** = p

**Variance** = pq

video link: [Bernoulli distribution](https://youtu.be/nl9WiZMZnYs?si=00bPbEJF_T8RA86i)

**Binomial Distribution**

The binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent Bernoulli trials, where each trial has only two possible outcomes: success or failure.

Key characteristics of the binomial distribution:

1. Two Outcomes: Each trial can result in one of two possible outcomes, typically labelled as "success" and "failure."
2. Fixed Number of Trials: The distribution is defined over a fixed number of trials, denoted as "n."
3. Independent Trials: Each trial is assumed to be independent of the others, meaning that the outcome of one trial does not affect the outcome of the next trial.
4. Constant Probability of Success: The probability of success (usually denoted as "p") remains constant from trial to trial. The probability of failure is then equal to 1 - p.
5. Discrete Values: The binomial distribution is a discrete distribution, meaning that the random variable, which represents the number of successes, can only take on integer values from 0 to n.

A discrete random variable is said to follow the binomial distribution with parameter n and p if its pdf is given by:

|  |
| --- |
| P (x: n, p) = nCx px (1-p) n-x  Or  P (x: n, p) = nCx px (q)n-x |

Where,

n = the number of experiments

x = 0, 1, 2, 3, 4, …n

p = Probability of Success in a single experiment

q = Probability of Failure in a single experiment = 1 – p

**Mean** = np

**Variance** =npq

**Poisson Distribution**

A Poisson distribution is a discrete [probability distribution](https://www.scribbr.com/statistics/probability-distributions/). It gives the probability of an event happening a certain number of times (*k*) within a given interval of time or space. The Poisson distribution has only one [parameter](https://www.scribbr.com/statistics/parameter-vs-statistic/), λ (lambda), which is the [mean](https://www.scribbr.com/statistics/central-tendency/#mean) number of events.

You can use a Poisson distribution to predict or explain the number of events occurring within a given interval of time or space. “Events” could be anything from disease cases to customer purchases to meteor strikes. The interval can be any specific amount of time or space, such as 10 days or 5 square inches. For example, a Poisson distribution could be used to explain or predict text messages per hour, influenza cases per year etc.

The Poisson distribution has only one [parameter](https://www.scribbr.com/statistics/parameter-vs-statistic/), called λ.

* The [mean](https://www.scribbr.com/statistics/central-tendency/#mean) of a Poisson distribution is λ.
* The [variance](https://www.scribbr.com/statistics/variance/) of a Poisson distribution is also λ.

The probability mass function of the Poisson distribution is:

P(X = k) = \dfrac{e^{-\lambda} \lambda^k}{k!}

Where:

* X is a random variable following a Poisson distribution
* k is the number of times an event occurs, k=0,1,2….
* \lambda is the average number of times an event occurs, >0

**Mean** = \lambda

**Variance** = \lambda

Video link: [poisson distribution](https://youtu.be/jmqZG6roVqU?si=OO0rg78Wfm5oUkGT)

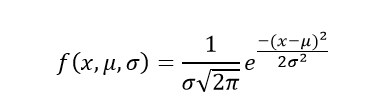
**Normal Distribution**

All kinds of variables in natural and social sciences are normally or approximately normally distributed. Height, birth weight, reading ability, job satisfaction, or SAT scores are just a few examples of such variables.

Normal distributions have the following key characteristics:

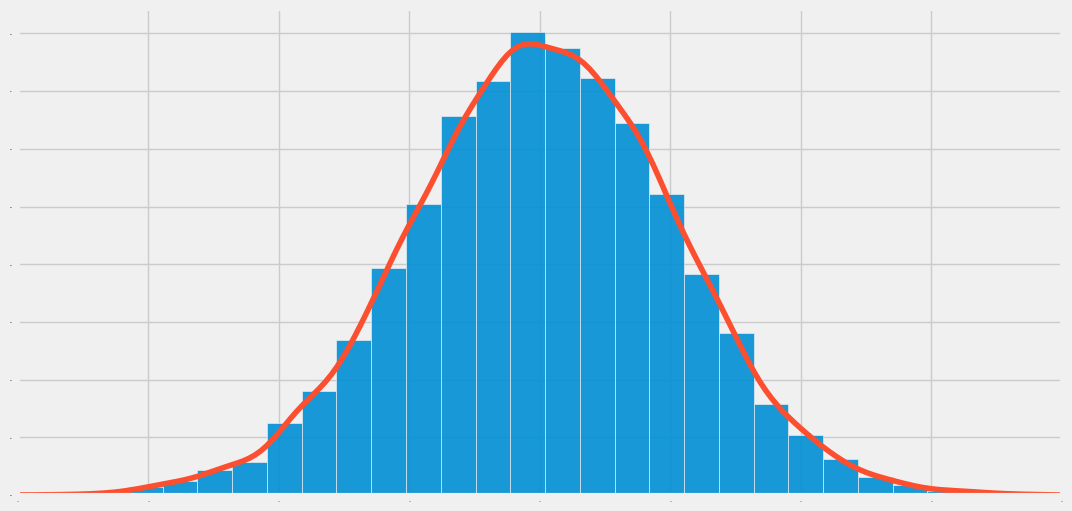
* They’re all [unimodal](https://statisticsbyjim.com/basics/unimodal-distribution/) (has only one peak), symmetric bell curves. The Gaussian distribution cannot model skewed distributions.
* The mean, median, and [mode](https://statisticsbyjim.com/glossary/mode/) are all equal.
* Half of the population is less than the mean and half is greater than the mean.
* The Empirical Rule allows you to determine the proportion of values that fall within certain distances from the mean. The empirical rule predicts that in normal distributions, 68% of observations fall within the first standard deviation (µ ± σ), 95% within the first two standard deviations (µ ± 2σ), and 99.7% within the first three standard deviations (µ ± 3σ) of the mean.

The probability density function of normal or gaussian distribution is given by;



Where,

* x is the variable, -∞< x < ∞
* μ is the mean
* σ is the standard deviation

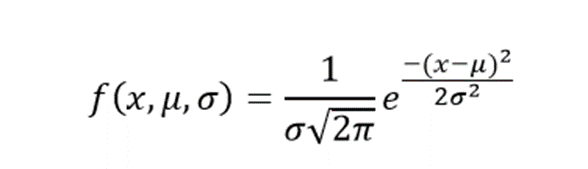


Real life examples of normal distribution are: -

* Height of people:

The height of people is an example of normal distribution. Most of the people in a specific population are of average height. The number of people taller and shorter than the average height people is almost equal, and a very small number of people are either extremely tall or extremely short.

Suppose you are measuring the heights of adult males in a certain population. You collect data from a large sample and plot a histogram of the heights. The histogram might look like a bell-shaped curve, which is indicative of a normal distribution.

* **Mean (μ)**: Let's say the mean height in your sample is 175 centimeters.
* **Standard Deviation (σ)**: The standard deviation measures how spread out the heights are. In this case, let's assume the standard deviation is 7 centimeters. This means that most heights will be within 7 centimeters of the mean.
* **Probability Density Function (PDF)**: The mathematical equation that describes the normal distribution is given by

Here, f(x) represents the probability density at a given height x.

With this information, you can make various statements about the heights of individuals in this population:

1. Approximately 68% of individuals will have heights between 168 and 182 centimeters because this range covers one standard deviation (σ) above and below the mean (175 ± 7).
2. Approximately 95% of individuals will have heights between 161 and 189 centimeters because this range covers two standard deviations above and below the mean (175 ± 2 \* 7).
3. Almost all (about 99.7%) of individuals will have heights between 154 and 196 centimeters because this range covers three standard deviations above and below the mean (175 ± 3 \* 7).

In summary, the normal distribution is a statistical model that helps us understand and make predictions about data with a bell-shaped curve. It is widely used in various fields such as statistics, finance, and natural sciences due to its simplicity and applicability to real-world phenomena.

* Exam scores of students:

Exam scores for a large group of students often approximate a normal distribution.

* Blood pressure:

Blood pressure measurements in a population often follow a normal distribution. Healthcare professionals use this distribution to identify individuals with high or low blood pressure.

Video link : [normal distribution](https://youtu.be/RKdB1d5-OE0?si=7Jn3tWcoV78MafWU)