DISTRIBUTION

By
Dr. Anjula Mehto
Assistant Professor
Thapar Institute of Engineering and Technology,
Patiala, Punjab
Email-anjula.mehto@thapar.edu

Discrete and Continuous Data

- When looking at a set of numbers, they are typically :
  - Discrete (countable)
  - Continuous (measurable)

Discrete Data

- Refers to individual and countable items (discrete variables).
- Involves counting rather than measuring.
- **Examples-**
  - Count number of computers in each department.
  - Count the number of students in a class.

Discrete Data

- **Characteristics-**
  - Discrete variables are finite, numeric, countable, and non-negative integers (5, 10, 15, and so on).
  - It can be easily visualized and demonstrated using simple statistical methods such as bar charts, line charts, or pie charts.
  - It can also be categorical - containing a finite number of data values, such as the gender of a person.

Continuous Data

- It is a type of numerical data that refers to the unspecified number of possible measurements between two realistic points.
- Continuous data is all about accuracy.
- Variables in these data sets often carry decimal points.
- **Examples-**
  - Measuring daily wind speed
  - Measuring temperature of a city
  - Measuring a person's height

Continuous Data

- **Characteristics-**
- Data changes over time and can have different values at different time intervals.
- Data is made up of random variables, which may or may not be whole numbers.
- Data is measured using data analysis methods such as line graphs, skews, and so on.
- Regression analysis is one of the most common types of continuous data analysis.

Statistical Distributions

- Also called as probability distribution.
- Statistical distributions are mathematical functions that describe the behavior and characteristics of random variables.
- *Statistical distribution helps to understand a problem better by assigning a range of possible values to the variables, making them very useful in data science and machine learning.*

Types of Statistical Distributions

- Depending on the type of data, distribution are grouped into two categories:
  - Discrete distributions for discrete data
  - Continuous distributions for continuous data

Discrete Distributions

- A discrete distribution is a probability distribution that describes the probability of occurrence of each possible outcome in a set of discrete values.
- It is characterized by a probability mass function (PMF), which gives the probability of each possible outcome.

Probability Mass Function (PMF)

- Gives the probability of a discrete random variable taking on a specific value.
- Maps each possible outcome of a random variable to its probability.
- The PMF is defined as:
- $P(X=x)$
  - X is the discrete random variable
  - x is the value of the random variable,

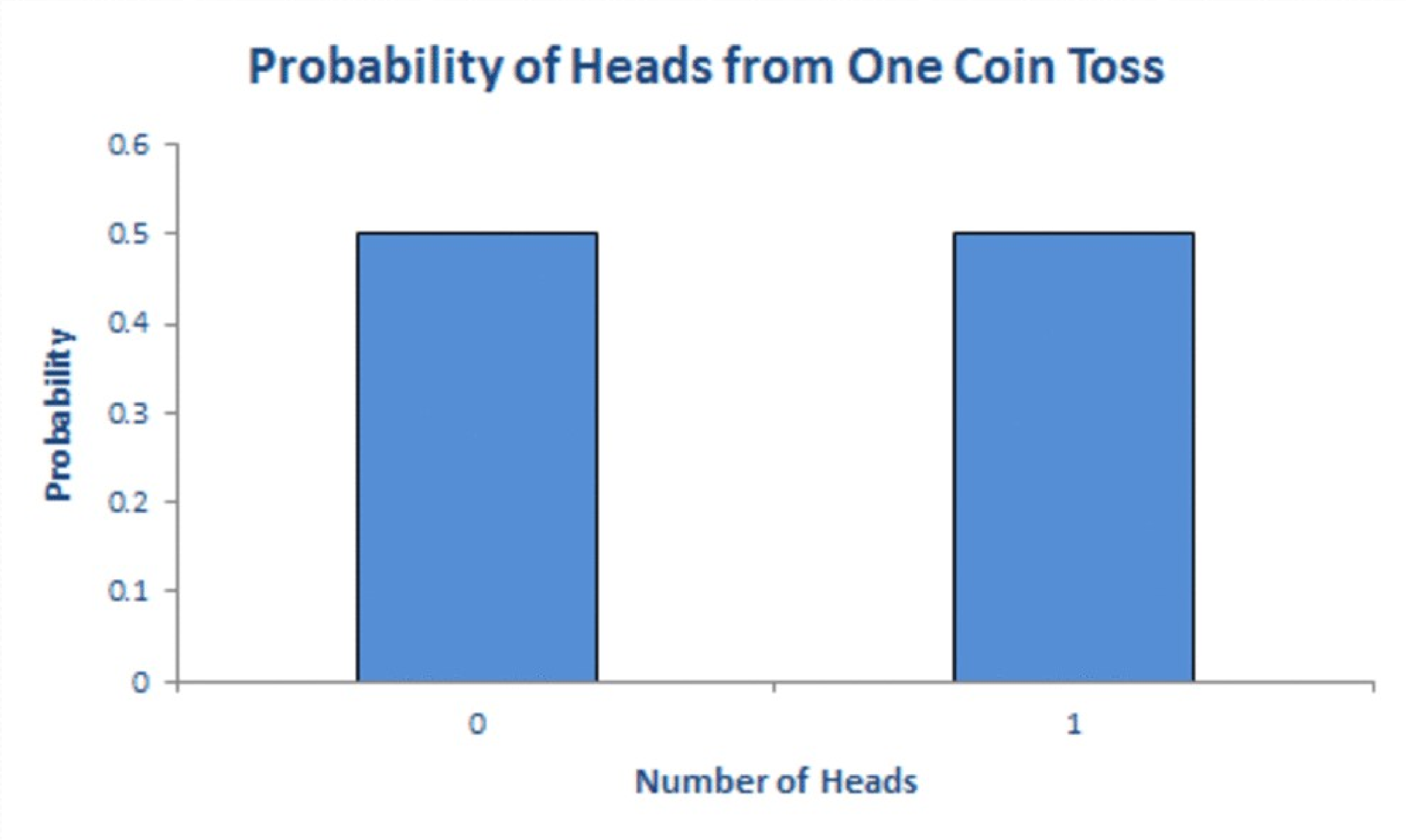
Types of Discrete Distributions

- Bernoulli distribution
- Binomial distribution
- Poisson distribution

Bernoulli Distribution

- Single Trial with Two Possible Outcomes.
- Any event with a single trial and only two possible outcomes follow a Bernoulli distribution.
- **Example-**
  - Flipping a coin.
  - Choosing between True and False in a quiz.

Bernoulli Distribution



Bernoulli Distribution

- The PMF of Bernoulli distribution=

$$\mathbf{p^x (1 - p)^{1 - x}, x \in \{0, 1\}}$$

'p' probability of success

(1-p) or 'q' probability of failure

Bernoulli Distribution

- The expected value or Mean of Bernoulli distribution:

$$E(x) = p$$

- Variance of Bernoulli distribution:

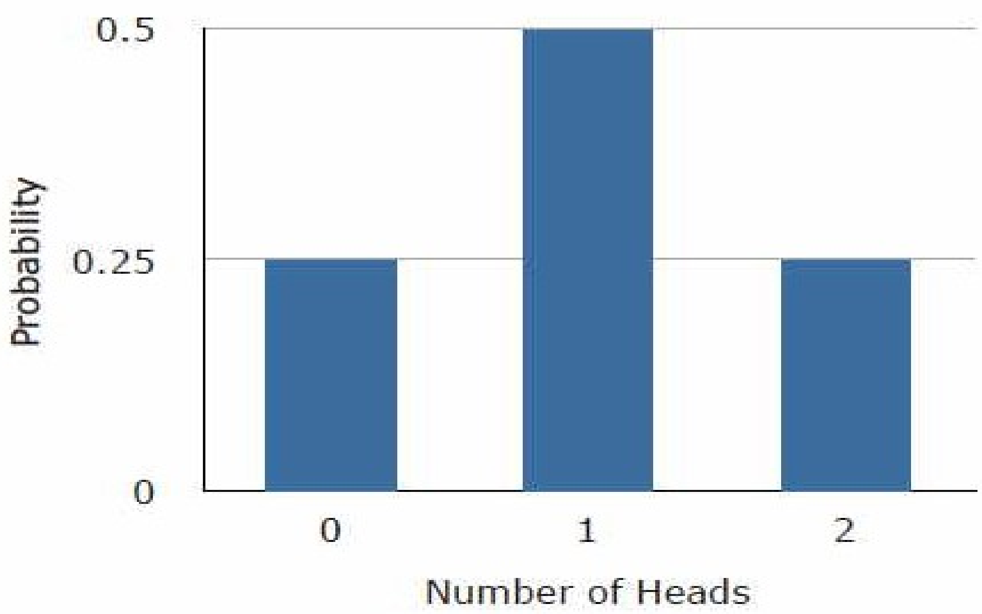
$$Var(x) = p(1-p)$$
$$= pq$$

Binomial Distribution

- **A sequence of Bernoulli events.**
- It can be thought of as the sum of outcomes of an event following a Bernoulli distribution.
- Therefore, it is used in binary outcome events, and the probability of success and failure is the same in all successive trials.
- **Example -**
- Flipping a coin multiple times to count the number of heads and tails

Binomial Distribution

- Example- If you flipped a coin twice
- [{H,H}, {H,T}, {T,H}, {T,T}]
- {H,H} = ½ * ½ = ¼, {T,T} = ½ * ½ = ¼
- {H,T} or {T,H} = ½ * ½ + ½ * ½ = ½



Binomial Distribution

- A binomial distribution is represented by :

B (n, p)

'n' is the number of trials,

'p' is the probability of success in a single trial

- The probability of success (x) for these n trials or PMF:

$$P\left(X = x\right) = \frac{n!}{x!\left(n-x\right)!}\, p^{x} \left(1 - p\right)^{\left(n-x\right)}$$

x= 0,1,2.............n

Binomial Distribution

- Expected value or Mean of a binomial distribution can be represented as :

$$E(x) = np$$

- Similarly, variance is represented as:

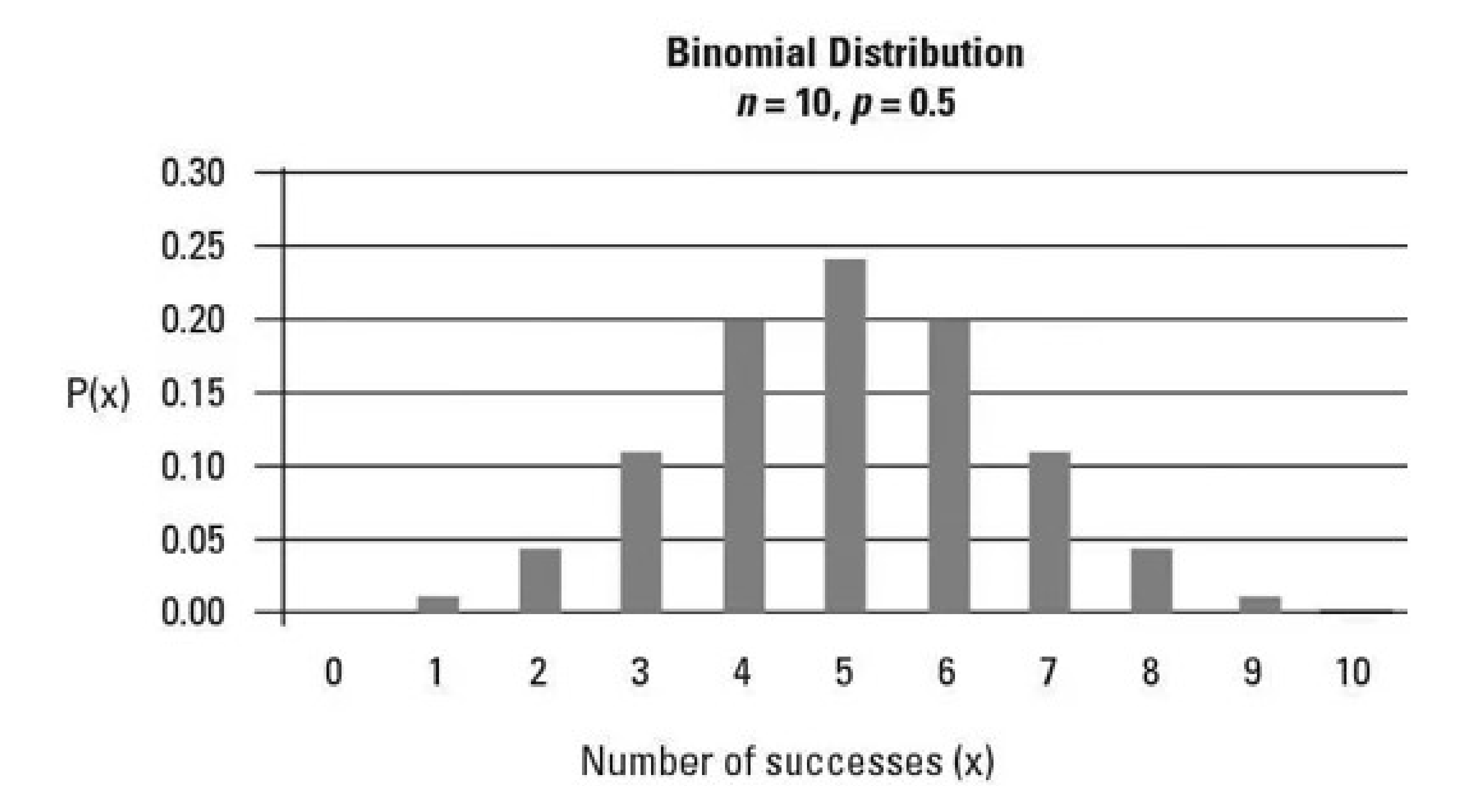
$$Var(x) = np(1-p)$$
$$= npq$$

Binomial Distribution

- For example, suppose that a candy company produces both milk chocolate and dark chocolate candy bars. The total products contain half milk chocolate bars and half dark chocolate bars.
- Say choose **ten candy bars** at random and **choosing milk chocolate is defined as a success.**
- **$n=10$, $p=1/2=0.5$**

Binomial Distribution

- The probability distribution of the number of successes during these 10 trials with p = 0.5



Numerical

- Suppose a basketball player makes a free throw with a probability of 0.7. If the player attempts 10 free throws, what is the probability that they make exactly 7 of them?

Numerical

- **Solution-** Binomial probability problem

$$P\left(X = x\right) = \frac{n!}{x!(n-x)!} p^x \left(1 - p\right)^{(n-x)}$$

Put the values; n=10, p=0.7, x=7 in the above formula:

P(7) = 0.266

Poisson Distribution

- The probability that an event May or May not occur.
- It gives the probability of an event happening a certain number of times (x) within a given interval of time or space.

Poisson Distribution

- **Examples-**
- The number of phone calls received by a call center during one hour of operation
- Text messages per hour
- Website visitors per month

Poisson Distribution

- **Characteristics:**
  - The events are independent of each other.
  - An event can occur any number of times (within the defined period).
  - Two events can't take place simultaneously.

Poisson Distribution

- The probability mass function (PMF) of the Poisson distribution is:

$$P(X = x) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

- X= random variable following a Poisson distribution
- x=number of times an event occurs
- P(X=x) = probability that an event will occur *x* times
- e = Euler's constant (approximately 2.718)
- λ= is the average number of times an event occurs

Poisson Distribution

- Expected value or Mean of a Poisson distribution can be represented as :

$$E(x) = \lambda$$

- Similarly, variance is represented as:

$$Var(x) = \lambda$$

Numerical

- Suppose that the average rate of calls received by the call center during one hour is 10. Then, calculate the probability of receiving 8 or fewer calls during one hour?

Numerical

- Solution- Poisson Distribution
- $\lambda = 10$
- where λ is the mean or average rate of calls received by the call center during one hour
- $X = x \leq 8$
- where 'X' is the random variable representing the number of calls received by the call center during one hour.

Numerical

$$P(X = x) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

- X=x ≤ 8

  P(X ≤ 8) = Σ P(X = x), for x = 0 to 8

  P(X = 0) = (10^0 * e^(-10)) / 0! ≈ 0.0000454
  P(X = 1) = (10^1 * e^(-10)) / 1! ≈ 0.000454
  P(X = 2) = (10^2 * e^(-10)) / 2! ≈ 0.00227
  P(X = 3) = (10^3 * e^(-10)) / 3! ≈ 0.00757
  P(X = 4) = (10^4 * e^(-10)) / 4! ≈ 0.0189

Numerical

$P(X = 5) = (10^5 * e^{(-10)}) / 5! \approx 0.0378$

$P(X = 6) = (10^6 * e^{(-10)}) / 6! \approx 0.0631$

$P(X = 7) = (10^7 * e^{(-10)}) / 7! \approx 0.0901$

$P(X = 8) = (10^8 * e^{(-10)}) / 8! \approx 0.1126$

$\mathbf{P(X \leq 8) \approx 0.332}$

Numerical

1. Suppose that a manufacturing company produces light bulbs at a rate of 3 defective bulbs per hour. What is the probability that exactly 2 defective bulbs are produced in a 30-minute interval?
2. Suppose a factory produces electronic components, and 5% of the components are defective. If a sample of 200 components is randomly selected, what is the probability that there are fewer than 10 defective components in the sample?

Numerical

- Solution (1)- Poisson Distribution
- $\lambda = (3/60) * 30 = 1.5$

  where λ is the rate parameter for the Poisson distribution
- X=x=2
- Put the values in the formula:

$$P(X = x=2) = (e^{\wedge}(-1.5) * 1.5^{\wedge}2) / 2!$$

$$\mathbf{P(X = 2) \approx 0.2510}$$

Numerical

- Solution (2)- Binomial distribution problem

  $p = 0.05$, $n = 200$, $x < 10$

  $P(X<10) = 0.98$

Continuous Distribution

- Describes the distribution of continuous random variables.
- A continuous random variable can take on any value within a range or interval of values, as opposed to a discrete random variable that can only take on distinct values.
- It is characterized by Probability Density Function (PDF).

Probability Density Function (PDF)

- Describes the probability distribution of a continuous random variable.
- Gives the relative likelihood of a random variable (X) taking on a particular value (x) within a given range of values (a, b).
- PDF=

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

Types of Continuous Distribution

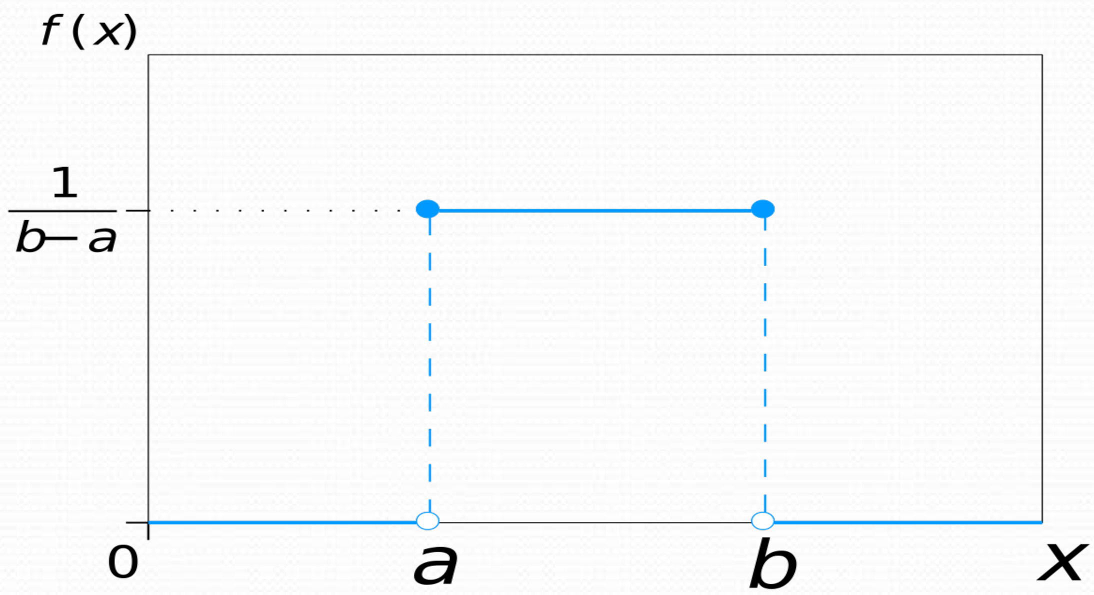
- **Uniform Distribution**
- **Normal or Gaussian Distribution**
- **Student t-Test Distribution**
- **Exponential Distribution**

Uniform Distribution

- It is a continuous or rectangular distribution.
- It describes an experiment where an outcome lies between certain boundaries.
- **Example-**
- Time to fly from Delhi to Hyderabad ranges from 120 to 150 minutes if we monitor the fly time for many commercial flights it will follow more or less the uniform distribution.

Uniform Distribution

- **PDF** $\quad f(x) = 1 / (b - a) \quad$ **for** $a \le x \le b$
- f(x) is the probability density function of X
- a and b are the lower and upper bounds of the distribution, respectively.



Uniform Distribution

- The Expected value or Mean

$$E(X) = (a + b) / 2$$

- Variance

$$Var(X) = (b - a)^2 / 12$$

Normal Distribution

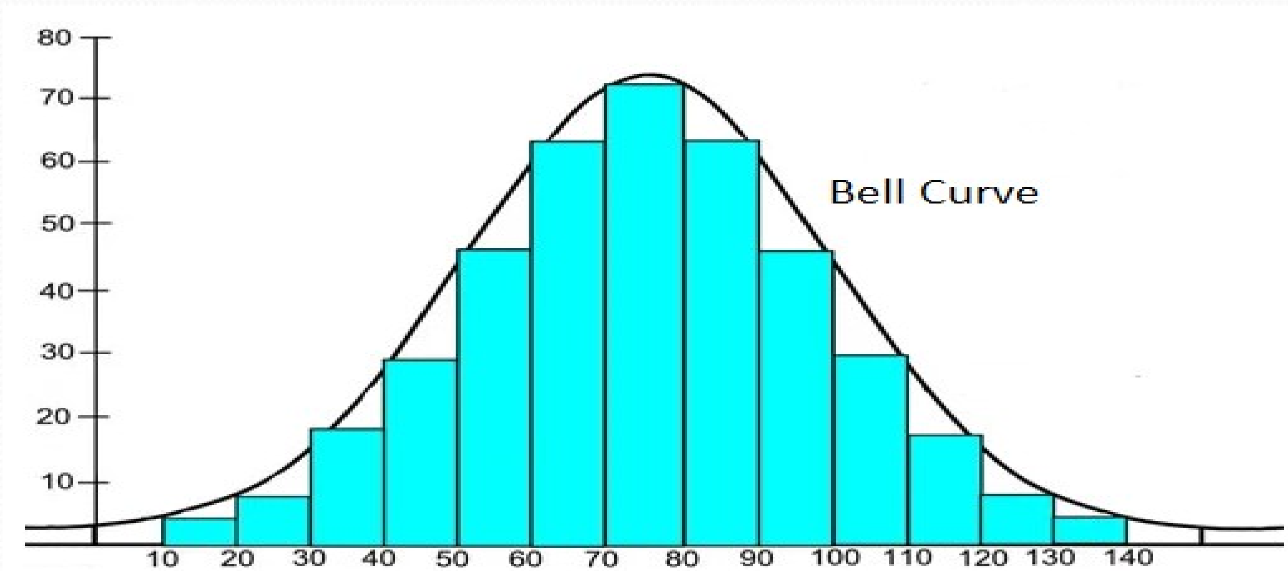
- **Symmetric Distribution of Values Around the Mean**
- Also called as Gaussian or Bell curve distribution.
- It is most commonly used in data science.
- Describes the probability of a continuous random variable that takes real values.
- When plotted, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.

Normal Distribution

- **Example-**
- Average weight of a population
- The scores of a quiz

The scores of a quiz

- Many of the students scored between 60 and 80.
- The students with scores that fall outside this range (outliers) are deviating from the center.



Normal Distribution

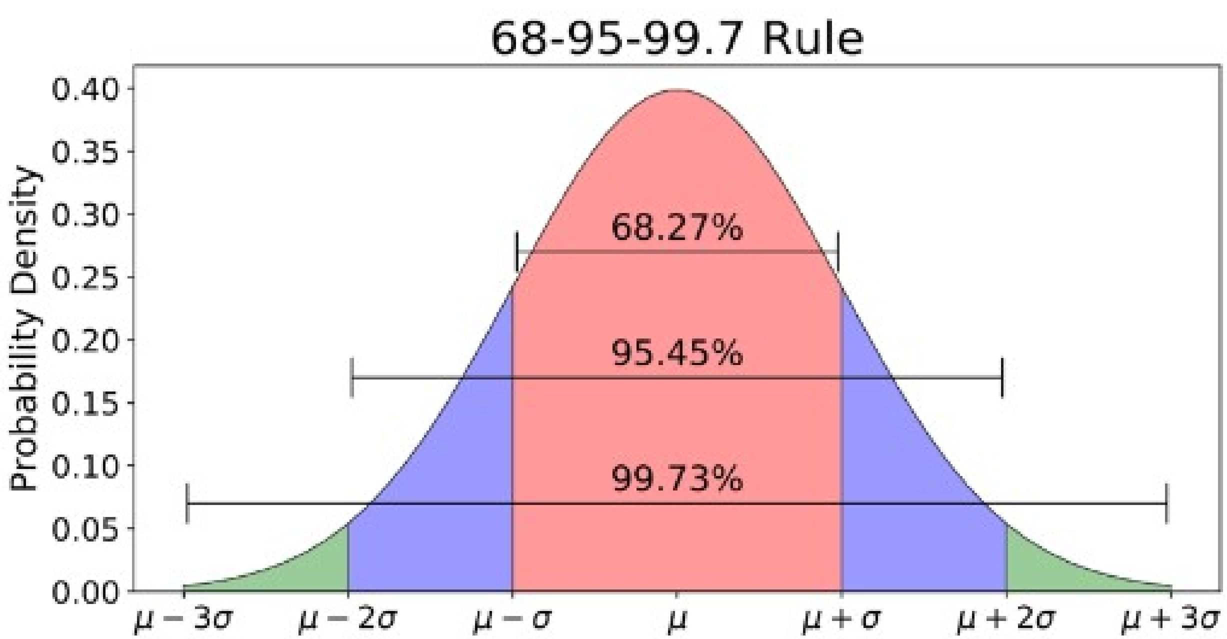
- **Characteristics-**
- The random variable takes values from $-\infty$ to $+\infty$.
- Mean, mode and median (measures of central tendency) coincide with each other.
- The distribution curve is symmetrical to the centre.
- The area under the curve is equal to 1.

Normal Distribution- 68-95-99.7 Rule

- While plotting a graph for a normal distribution, 68% of all values lie within one standard deviation from the mean.
- Similarly, 95% of the values lie within two standard deviations from the mean, and 99.7% lie within three standard deviations from the mean.
- This last interval captures almost all matters. If a data point is not included, it is most likely an outlier.

Normal Distribution- 68-95-99.7 Rule

- If the mean is 70 and the standard deviation is 10, 68% of the values will lie between 60 and 80, and so on for 95% and 99.7%.



Normal Distribution

- **PDF of normal distribution-**

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$-\infty < \mathbf{x} < +\infty, \ -\infty < \boldsymbol{\mu} < +\infty, \ \boldsymbol{\sigma} > \mathbf{0}$$

- μ is the mean (or expectation) of the distribution
- σ is the standard deviation of the distribution
- ‘x’ is the specific value of the random variable ‘X’

Normal Distribution

- The expected value or Mean of a Normal distribution:

$$E(x) = \mu$$

- Variance of a Normal distribution:

$$Var(x) = \sigma^2$$

Standard Normal Distribution

- Has a mean of zero and a standard deviation of one.
- The x values of the standard normal distribution are called z-scores.
- Z-score is used to determine the probability of a given value occurring in a normal distribution, using standard normal distribution.

Z-SCORE

- The z-score equals an X minus the population mean (μ) all divided by the standard deviation (σ).

$$Z = \frac{X - \mu}{\sigma}$$

Standard Normal Distribution

- **PDF :**

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

where $-\infty < x < +\infty$

- **Expected value or Mean:**

$E(x)=0$

- **Variance:**

$Var(X)=1$

Numerical

- The marks of students (X) *in a class of 70 students follows normal distribution with mean 50 units and variance 225 units. Find the probability that* $P(40 < X < 60)$.

Numerical

- **Solution- Normal Distribution**

Mean (μ) of 50 units

Variance (σ^2) of 225 units,

Standardize the distribution using the Z-score

So, to find the probability $P(40 < X < 60)$, first find the Z-score for $X = 40$ and $X = 60$:

$Z1 = (40 - 50) / 15 = -0.67$

$Z2 = (60 - 50) / 15 = 0.67$

Numerical

- **Solution-**

Using a calculator, the probability of Z being between -0.67 and 0.67.

$$P(-0.67 < Z < 0.67) = 0.7486 - 0.2514$$

$$= 0.4970$$

Student t-Test Distribution

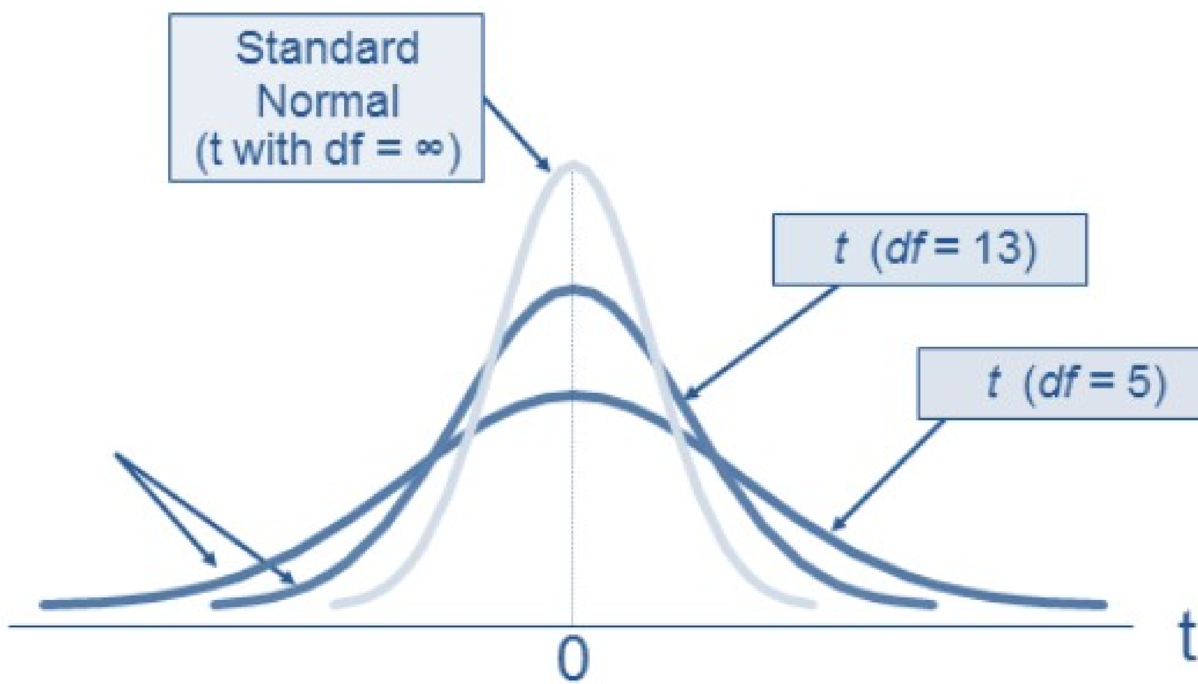
- **Small sample size approximation of a normal distribution.**
- It is also known as the 't' distribution.
- Similar to the standard normal distribution with its bell shape but has heavier tails.
- The shape of the t-distribution depends on the degrees of freedom 'n', which is equal to the sample size 'k' minus one.
- Degree of freedom 'n' = k-1

Student t-Test Distribution

- **Example-**
- Suppose we deal with the total apples sold by a shopkeeper in a month.
- In that case, we will use the normal distribution.
- Whereas, if we are dealing with the total amount of apples sold in a day, i.e., a smaller sample, we can use the ‘t’ distribution.

Student t-Test Distribution

- As the sample size increases, the t-distribution approaches the normal distribution, and, the t-distribution can be used for larger sample sizes as well.



Student t-Test Distribution

- PDF =

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi},\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

- n= degree of freedom
- Γ is the gamma function, which is a generalization of the factorial function to complex numbers

Student t-Test Distribution

- **Expected value or Mean**

$$E(x) = 0$$

- **Variance**

$$Var(x) = n/(n-2)$$

n= degree of freedom

Exponential Distribution

- It models elapsed time between two events.
- It is concerned with the amount of time until some specific event occurs.

Exponential Distribution

- Example-
- How long do we need to wait before a customer enters a shop?
- How long will it take before a call center receives the next phone call?
- How long will a piece of machinery work without breaking down?

Exponential Distribution

- All these questions concern the time we need to wait before a given event occurs.
- If the waiting time is unknown, it is often appropriate to think of it as a random variable having an exponential distribution.

Exponential Distribution

- PDF of Exponential Distribution:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

'x' is the time between events

$\lambda > 0$, is the rate parameter and it is inversely proportional to expected duration (μ)

Exponential Distribution

- The CDF of the exponential distribution gives the probability that the time between events is less than or equal to a specific value x.
- CDF of Exponential Distribution:

$$\mathbf{F(x; \lambda) = P(X \leq x) = 1 - e^{\wedge}(-\lambda x) \text{ for } x \geq 0}$$

'x' is the time between events

$\lambda > 0$ is the rate parameter and it is inversely proportional to expected duration (μ)

Exponential Distribution

- **Expected value or Mean of Exponential Distribution-**

$$E(X) = 1/\lambda$$

- **Variance-**

$$Var(X) = 1/\lambda^2$$

Exponential Distribution

- Let's say find the probability that the time between events is less than or equal to ‘1’ minute if $\lambda = 10$ events per hour.

Exponential Distribution

- Solution- $P(X \leq 1)$, $\lambda = 10$

Convert 1 minute into hour

Using the CDF of the exponential distribution:

$$= 1 - e^{(-10*0.0167)}$$

$$= 1 - e^{(-0.167)}$$

$$= 0.15$$

Exponential Distribution

- The time (in hours) required to repair a machine is an exponentially distributed random variable with parameter $\lambda = 1/2$. What is the probability that a repair time exceeds 2 hours?

Exponential Distribution

- Solution- $\lambda = ½$, $P(X \geq 2)$

  Complement rule

  $P(X \geq x) = 1 - P(X \leq x)$

  $= 1 - [1 - e^{(-\lambda x)}]$

  $= e^{(-\lambda x)}$

  $P(X \geq 2) = 1 - P(X \leq 2)$

  $= e^{(-\lambda x)}$

  $= e^{(-1/2 * 2)}$

  $= 0.367$

Numerical

- The length of life of an instrument produced by a machine has a normal distribution with a mean of 12 months and standard deviation of 2 months. Find the probability that an instrument produced by this machine will last less than 7 months.

Numerical

- Solution-
- X is the value to standardize, X = 7 months
- μ is the mean, μ = 12 months
- σ is the standard deviation, σ = 2 months
- Substituting the given values, in the Z-score formula
- z = (7 - 12) / 2 = -2.5
- =0.0062 or 0.62%

Exponential Distribution

- Suppose that the time between machine breakdowns at a factory follows an exponential distribution with a mean of 10 hours. Calculate the probability that the time between breakdowns is between 5 and 10 hours.

Exponential Distribution

Find $P(5 \le X \le 10)$

- **Use Interval rule-**
- **Probability of being inside the interval is complement of being outside the interval.**
- **The probability of being outside the interval is the composite event of being too low $P(X \le 5)$ for the interval and being too high $P(X \ge 10)$ for the interval.**
- **$P(5 \le X \le 10) = 1 - [P(X \le 5) + P(X \ge 10)]$**

Exponential Distribution

Compute $P(5 \le X \le 10)$ with $\lambda = 1/10$

$P(5 \le X \le 10) = 1 - [P(X \le 5) + P(X \ge 10)]$

Too low $P(X \le 5) = 0.3934$

Too high $P(X \ge 10) = 0.3678$

Outside $= [P(X \le 5) + P(X \ge 10)] = 0.7612$

Inside $= P(5 \le X \le 10) = 1 - 0.7612 = 0.2388$