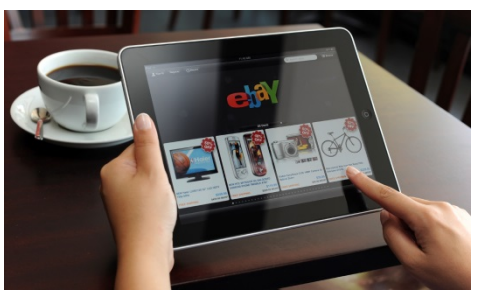
Introduction to Data Science

Dr. Parul Patel
Assistant Professor
Department of ICT, VNSGU, Surat

Introduction to Data Science

Lots of data is being collected and warehoused

- Web data, e-commerce
- Financial transactions, bank/credit transactions
- Online trading and purchasing
- Social Network

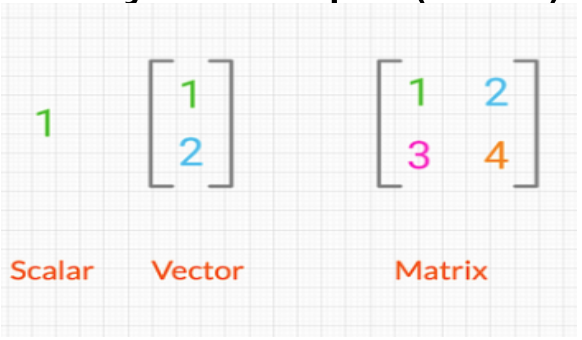


Vector & Matrix

Scalar: Any single numerical value is a scalar as shown in the image above. It is simply denoted by lowercase and italics. For example: n

Vector: An array of numbers(data) is a vector. You can assume a column in a dataset to be a feature vector.

Matrix: A matrix is a 2-D array of shape (m×n) with m rows and n columns.



Descriptive Statistics

Descriptive statistics is about describing and summarizing data. It uses two main approaches:

The quantitative approach describes and summarizes data numerically. The visual approach illustrates data with charts, plots, histograms, and other graphs.

Types of Measures

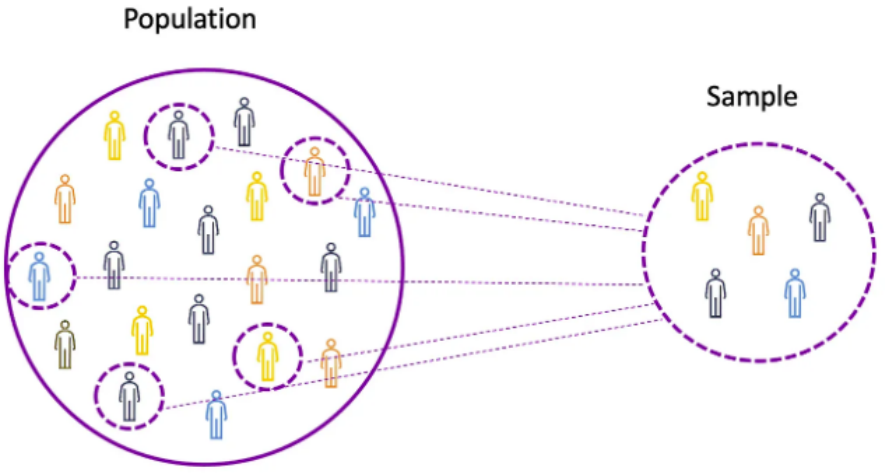
1. **Central tendency** tells you about the centers of the data. Useful measures include the mean, median, and mode.

2. **Variability** tells you about the spread of the data. Useful measures include variance and standard deviation.

3. **Correlation or joint variability** tells you about the relation between a pair of variables in a dataset. Useful measures include covariance and the correlation coefficient.

Mean, Median, Mode

The population is the set of all observations (individuals, objects, events, or procedures) and is usually very large and diverse, whereas a sample is a subset of observations from the population that ideally is a true representation of the population.

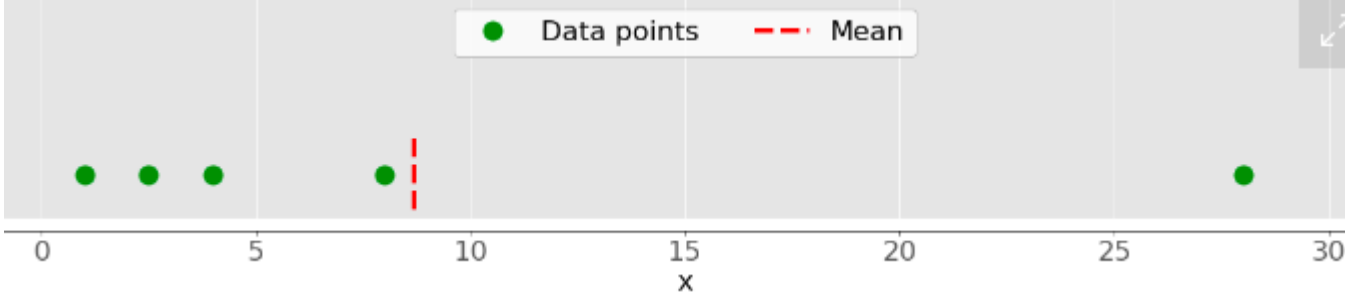


Mean

The mean, also known as the average, is a central value of a finite set of numbers.

$$x_1, x_2, x_3, \ldots, x_N$$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$



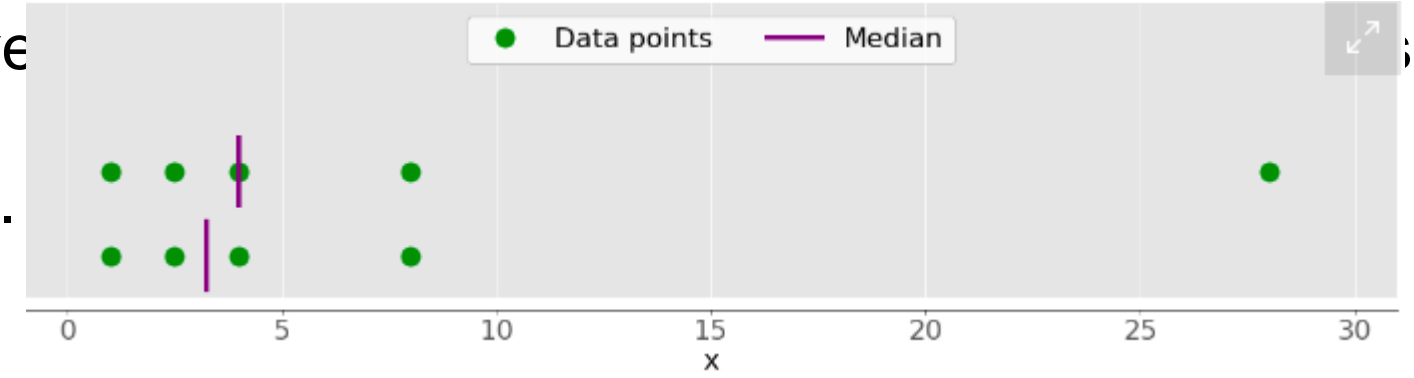
The green dots represent the data points 1, 2.5, 4, 8, and 28.

Median

- The sample median is the middle element of a sorted
- dataset. The dataset can be sorted in increasing or
- decreasing order.
- If the number of elements n of the dataset is odd, then the
- median is the value at the middle position: $0.5(n + 1)$.
- If n is even, then the median is the arithmetic mean of the \
- two values in the middle, that is, the items at the positions
- $0.5n$ and $0.5n + 1$.

Median

- For example, if you have the data points 2, 4, 1, 8, and 9, then
- the median value is 4, which is in the middle of the sorted dataset
- (1, 2, 4, 8, 9).
- If the data points are 2, 4, 1, and 8, then the median is 3, which is
- the ave sequence (2
- and 4).



Mode

- The sample mode is the value in the dataset that occurs
- most frequently.
- If there isn’t a single such value, then the set is multimodal
- since it has multiple modal values.
- For example, in the set that contains the points 2, 3, 2, 8,
- and 12, the number 2 is the mode because it occurs twice,
- unlike the other items that occur only once.

Variance

The variance measures how far the data points are spread out from the average value, and is equal to the sum of squares of differences between the data values and the average (the mean).

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

Standard Deviation

The standard deviation is simply the square root of the variance and measures the extent to which data varies from its mean.

Standard deviation is often preferred over the variance because it has the same unit as the data points, which means you can interpret it more easily.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Variance

	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$

Covariance

- The covariance is a measure of the joint variability of two
- random variables and describes the relationship between \
- these two variables.
- It is defined as the expected value of the product of the two
- random variables' deviations from their means.

$$\text{Covariance}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

xi = data value of x
yi = data value of y
$\bar{x}$ = mean of x
$\bar{y}$ = mean of y
N = number of data values.

Types of Covariance

Positive Covariance

A positive covariance between two variables indicates that they are heading in the same direction. The variables, in this case, behave similarly. That is, if the values of one variable (more or smaller) correspond to the values of another variable, they are said to be in positive covariance.

Negative Covariance

When two variables have a negative covariance, the variables shift in the opposite direction. It is the inverse of positive covariance, in which higher values of one variable correlate to lower values of another and vice versa.

Correlation

Variables within a dataset can be related for lots of reasons.
For example:

- One variable could cause or depend on the values of another variable.
- One variable could be lightly associated with another variable.
- Two variables could depend on a third unknown variable.

It can be useful in data analysis and modeling to better understand the relationships between variables. **The statistical relationship between two variables is referred to as their correlation.**

A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values decrease. Correlation can also be neutral or zero, meaning that the variables are unrelated.

Correlation

A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values decrease. Correlation can also be neutral or zero, meaning that the variables are unrelated.

Positive Correlation: both variables change in the same direction.
Neutral Correlation: No relationship in the change of the variables.
Negative Correlation: variables change in opposite directions.

Correlation

The correlation is also a measure for relationship and it measures both the strength and the direction of the linear relationship between two variables.

If a correlation is detected then it means that there is a relationship or a pattern between the values of two target variables.

$$Cor\ (X, Z) = \frac{Cov\ (X,Z)}{\sigma_x \sigma_z}$$

Correlation coefficients' values range between -1 and 1. Keep in mind that the correlation of a variable with itself is always 1, that is Cor(X, X) = 1.

Correlation

The correlation is also a measure for relationship and it measures both the strength and the direction of the linear relationship between two variables.

If a correlation is detected then it means that there is a relationship or a pattern between the values of two target variables.

$$Cor\ (X, Z) = \frac{Cov\ (X,Z)}{\sigma_x \sigma_z}$$

Correlation coefficients' values range between -1 and 1. Keep in mind that the correlation of a variable with itself is always 1, that is Cor(X, X) = 1.

Probability

- Probability means possibility .
- It is a branch of mathematics that deals with occurrence of a random event.
- The value is expressed from 0 to 1.

Binomial Distribution

The binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each with the boolean-valued outcome: success (with probability p) or failure (with probability q = 1 − p).

The binomial distribution is useful when analyzing the results of repeated independent experiments, especially if one is interested in the probability of meeting a particular threshold given a specific error rate.

Let's assume a random variable X follows a Binomial distribution, then the probability of observing k successes in n independent trials can be expressed by the following probᵃ

$$Pr\,(\mathrm{X} = \mathrm{k}) = \binom{n}{k} p^k q^{n-k}$$

Binomial Distribution

- Charactertistics
  - A Trial has only two possible outcomes – “Success” or “Failure”, “Head” or “Trail”, “Win” or “Lose” , “even” or “odd”
  - There is a fixed number n of independent trials.
  - The trials of experiment are independent from each other.
  - The probability of success, p , remains constant from trial to trial
  - If p presents the probability of success then,
    - (1-p)= q is the probability of failure.

Binomial Distribution

Mean $\mu = n \cdot p$

Variance $\sigma^2 = n \cdot p \cdot q$

Std. Dev. $\sigma = \sqrt{n \cdot p \cdot q}$

Where

n **= number of fixed trials**

p **= probability of success in one of the *n* trials**

q **= probability of failure in one of the *n* trials**

Binomial Distribution

Binomial Distribution Formula

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

where

n = the number of trials (or the number being sampled)

x = the number of successes desired

p = probability of getting a success in one trial

$q = 1 - p$ = the probability of getting a failure in one trial

Poisson Distribution

The Poisson distribution is the discrete probability distribution of the number of events occurring in a specified time period, given the average number of times the event occurs over that time period.

$$Pr\,(\mathrm{X} = \mathrm{k}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where e is Euler's number and λ lambda, the arrival rate parameter is the expected value of X.

Poisson Distribution

Few Cases where n is large and p (chances of success) is very small then we go for Poisson distribution, but such cases arise in case of rare events.

Requirements for the Poisson Distribution:

- The random variable x is the number of occurrences of an event over some interval.
- The occurrences must be random
- The occurrences must be independent of each other.
- The occurrences must be uniformly distributed over the interval being used.

Poisson Distribution

- An event can occur any number of times in a given time period.
- An event occur independently . In other words, if an event occurs, it
- does not affect the probability of another event occuring in the same
- time period.
- The rate of occuring is constant. e.g the rate does not change based on
- time
- The probability of an event occuring is propotional to the length of the
- time period. For example, it should be twice as likely for an event to
- occur in a 2 hour time period than it is for an event to occur in 1 hour
- period.

Poisson Distribution

- A Poisson Distribution is appropriate for modeling the number of phone calls an office would receive during the noon hour. If they know that they got average 4 calls per hour during that time period.
- Although, the average is 4 calls, they could theoretically get any number of calls during that time period.
- The events are effectively independent since there is no reason to expect a caller to affect the chances of another person calling.
- It is reasonable to assume that (for example) the probability of getting a call in the first half hour is the same as the probability of getting a call in another half hour.

Poisson Distribution (More Examples)

- For example, Poisson distribution can be used to model the number of customers arriving in the shop between 7 and 10 pm, or the number of

- patients arriving in an emergency room between 11 and 12 pm.
- The number of thefts recorded in an area on a day
- The number of customers arriving at salon in an hour.

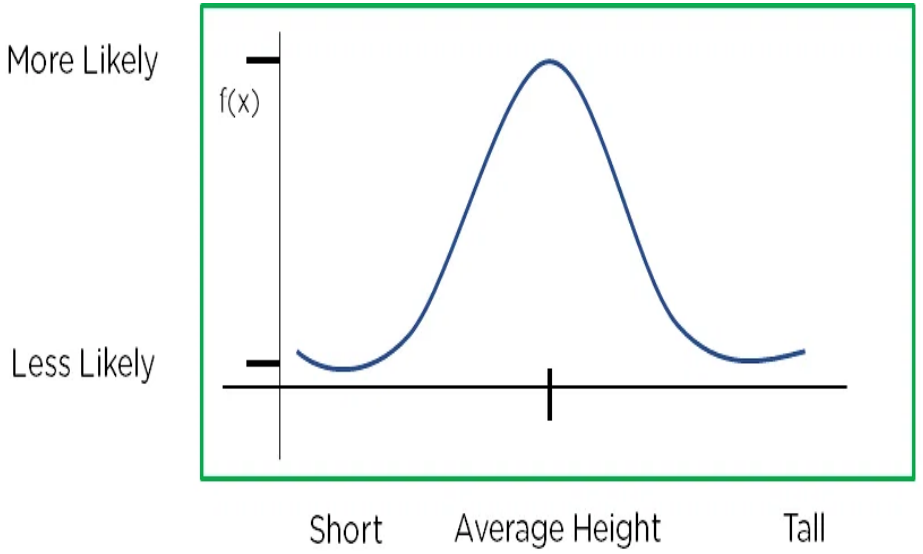
Poisson Distribution

➔

BINOMIAL DISTRIBUTION	POISSON DISTRIBUTION
Binomial distribution is one in which the probability of repeated number of trials are studied.	Poisson Distribution gives the count of independent events occur randomly with a given period of time.
The probability of repeated number of trails are studied: **Fixed**	The count of independent events occur randomly with a given period of time: **Infinite**
Only two possible outcomes, i.e. success or failure.	Unlimited number of possible outcomes.
Mean > Variance	Mean = Variance
Example: Coin tossing experiment.	Example: Printing mistakes/page of a large book.

Normal Distribution

A normal distribution has a probability distribution that is centered around the mean. This means that the distribution has more data around the mean. The data distribution decreases as you move away from the center. The resulting curve is symmetrical about the mean and forms a bell-shaped distribution.



Normal Distribution

1. It is symmetric
   - The shape of the normal distribution is perfectly symmetrical.
   - This means that the curve of the normal distribution can be divided from the middle and we can produce two equal halves.
   - Moreover, the symmetric shape exists when an equal number of observations lie on each side of the curve.
2. The mean, median and mode are equal
   - The midpoint of normal distribution refers to the point with maximum frequency i.e., it consists of most observations of the variable.
   - The midpoint is also the point where all three measures of central tendency fall. These measures are usually equal in a perfectly shaped normal distribution.
3. Empirical rule
   - In normally distributed data, there is a constant proportion of data points lying under the curve between the mean and a specific number of standard deviations from the mean.
   - Thus, for a normal distribution, almost all values lie within 3 standard deviations of the mean.

Applications Normal Distribution

Problem Statement:

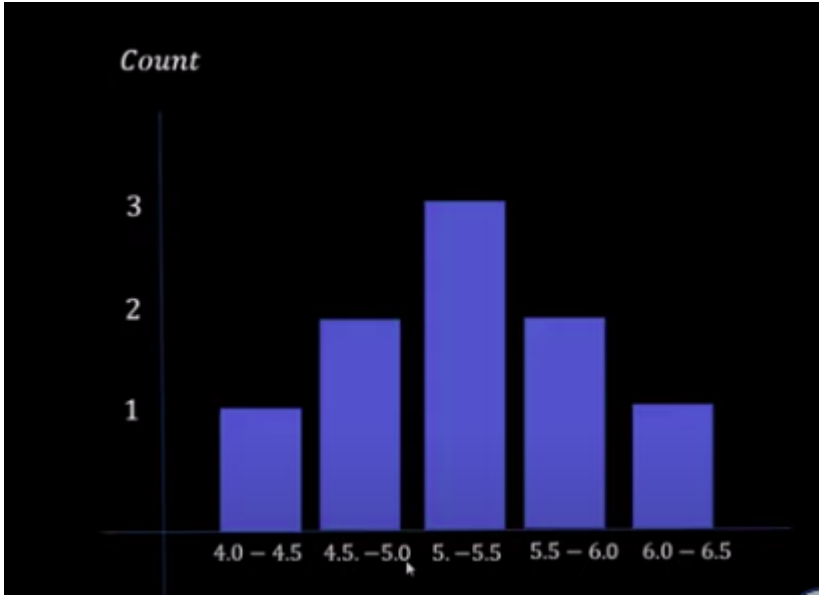
Let's have data of heights of Indian women aged 18 to 24, which is approximately normally distributed with a mean of 65.5 inches and a standard deviation of 2.5 inches.

From the empirical rule, it follows that:

– 68% of these Indian women have heights between 65.5 – 2.5 and 65.5 + 2.5 inches or between 63 and 68 inches,
– 95% of these Indian women have heights between 65.5 – 2(2.5) and 65.5 + 2(2.5) inches, or between 60.5 and 70.5 inches.
– Therefore, the tallest 2.5% of these women are taller than 70.5 inches. (The extreme 5% fall more than two standard de

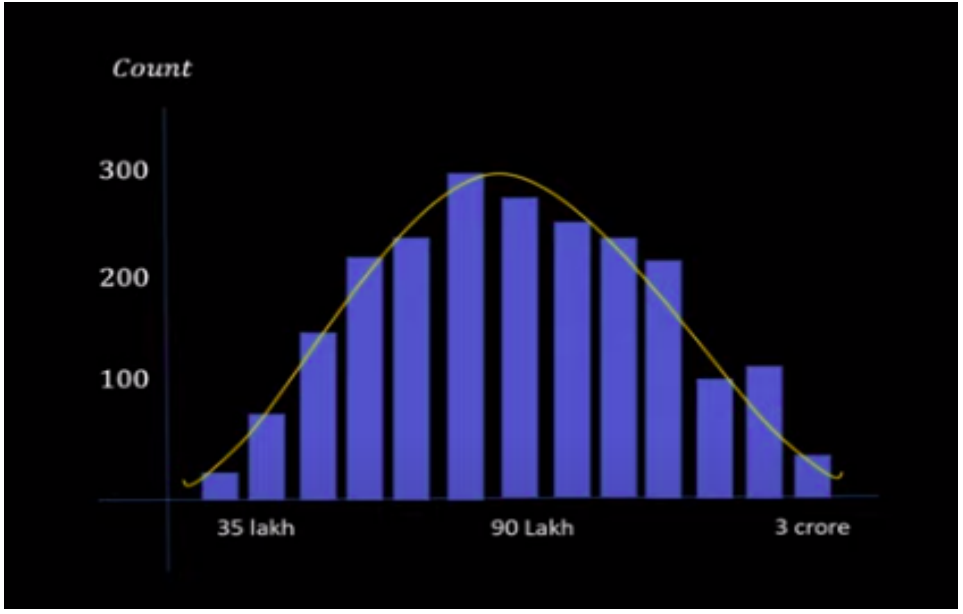
Normal Distribution

Name	Height
Om	6.2
Dhyan	5.7
Shlok	4.6
Meera	5.4
Thomas	5.9
Rina	4.3
Mohan	5.1
Rob	5.2
Meena	4.9



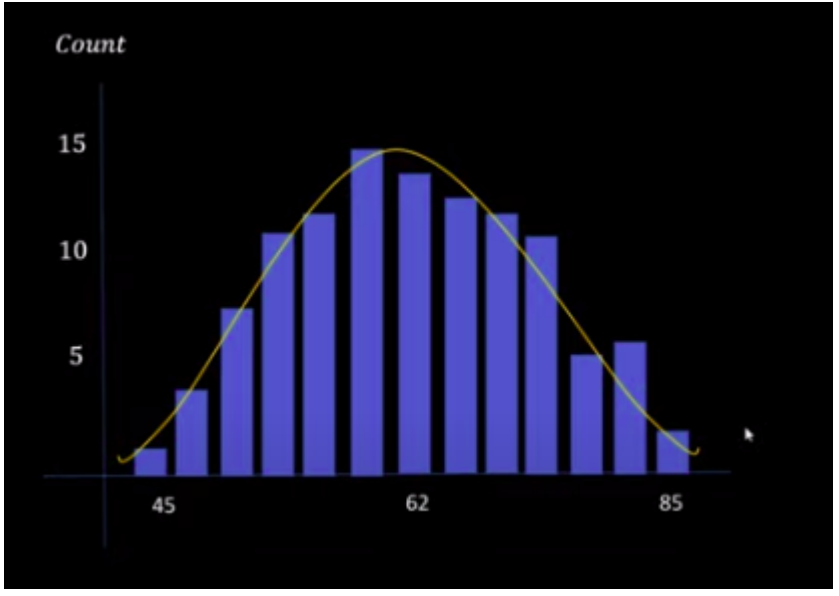
Normal Distribution

2 BHK Flat price in Pune



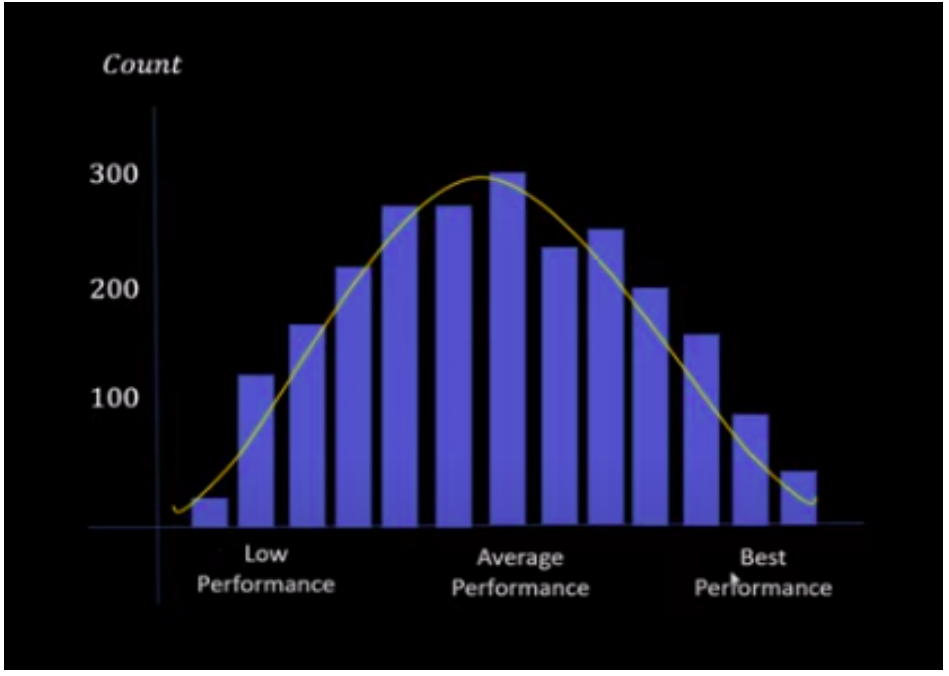
Normal Distribution

Test Score out of 100



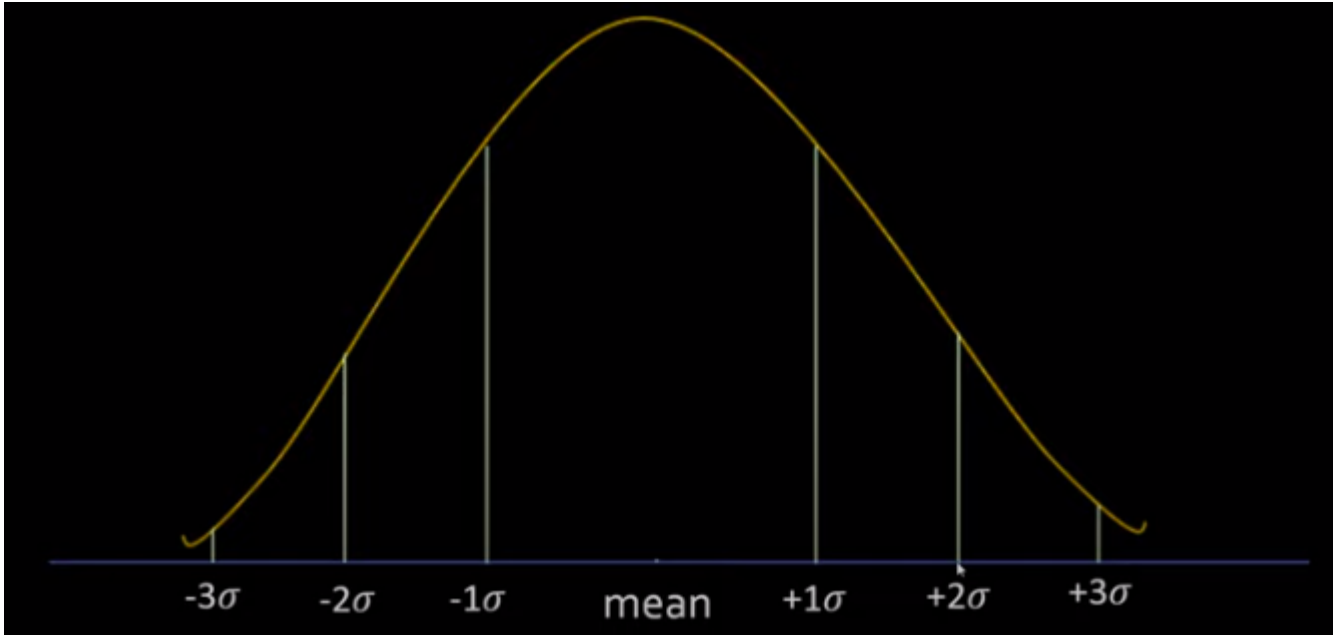
Normal Distribution

Employee Performance



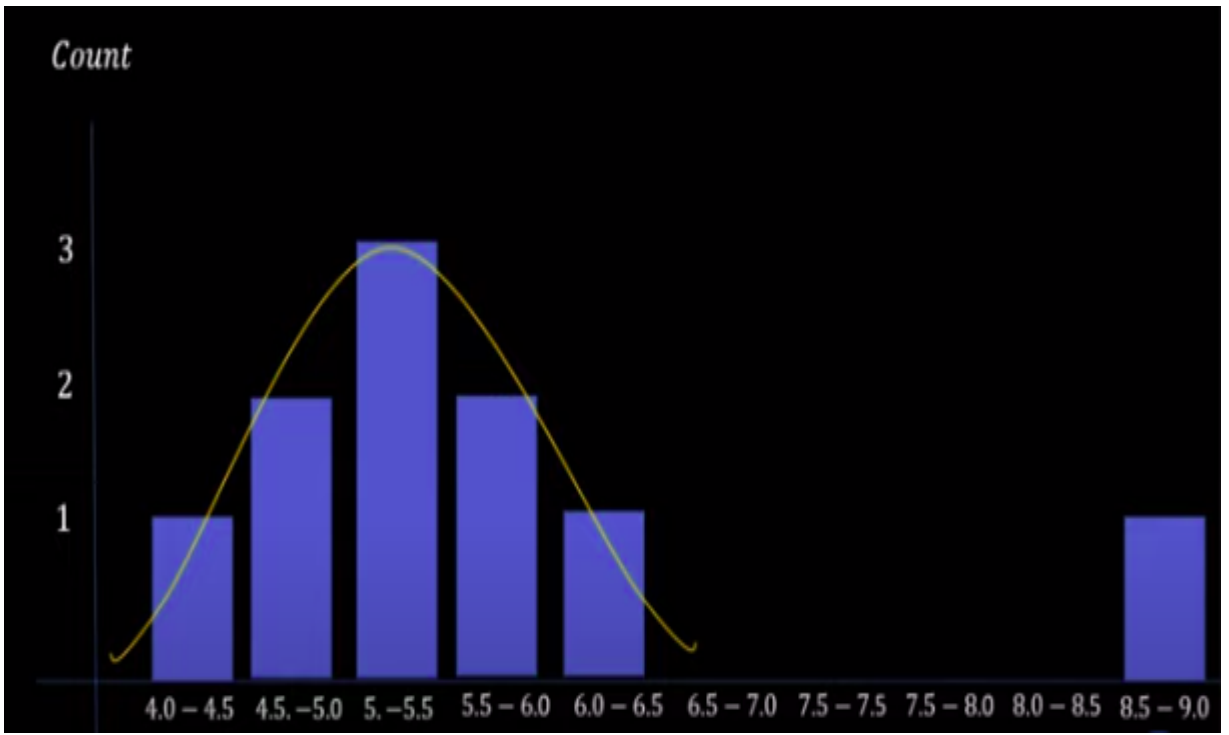
Normal Distribution

Employee Performance



Normal Distribution – Outlier Removal

Name	Height
Om	6.2
Dhyan	5.7
Shlok	4.6
Meera	5.4
Thomas	5.9
Rina	4.3
Mohan	5.1
Rob	5.2
Meena	4.9
Sudeha	9.0



Normal Distribution

The Normal probability distribution is the continuous probability distribution for a real-valued random variable. Normal distribution, also called Gaussian distribution is arguably one of the most popular distribution functions that are commonly used in social and natural sciences for modeling purposes, for example, it is used to model people's height or test scores.

$$Pr\,(\mathrm{X} = \mathrm{k}) = \frac{1}{\sigma\sqrt{2\pi}}\; e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

where the parameter μ (mu) is the mean of the distribution also referred to as the location parameter,
parameter σ (sigma) is the standard deviation of the distribution also referred to as the scale parameter.
The number π (pi) is a mathematical constant approximately equal to 3.14.