Session-16 Statistics

09 August 2024 01:44 PM

Agenda

- **Normal or Gaussian Distribution**
- **Properties of Normal Distribution**
- **Empirical Rule in Normal Distribution**
- **Central Limit Theorem**
- **Covariance**
- **Pearson Coefficient Correlation**

- **Normal or Gaussian Distribution**

In a normal distribution, data is symmetrically distributed with no skew. When plotted on a graph, 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 distributions are also called Gaussian distributions or bell curves because of their shape.

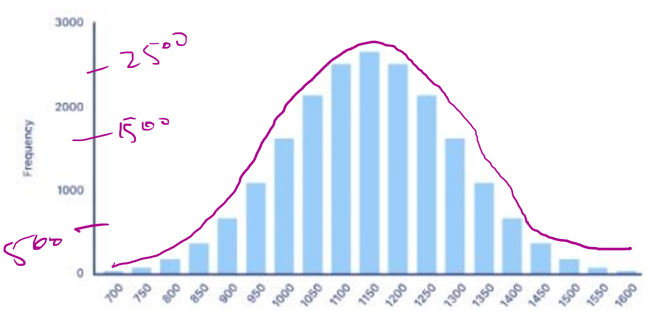
$\boxed{\text{mean = median = mode}}$

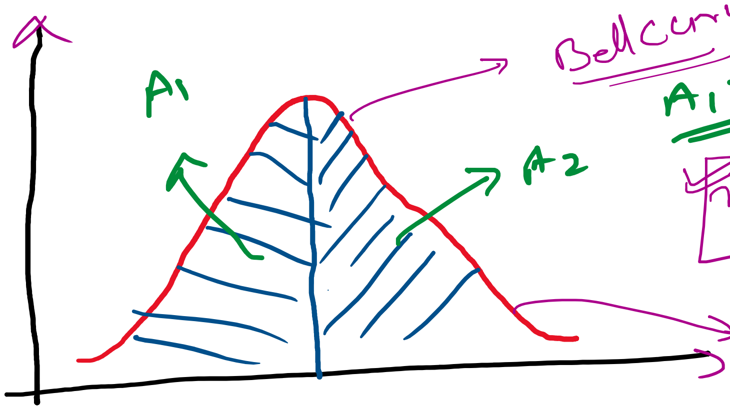
eg = Height

170 -
180 -
165
163
173
⋮
180

100

Example of normal distribution





Bell curve

$A_1 = A_2$

$\boxed{\text{mean = median = mode}}$

Normal/Gaussian Distribution

* It is a type of continous Probability distribution for a real-valued random variable.

Notation $N(\mu, \sigma^2)$ → mean, → std

Parameter → $\mu \in R$ = mean

$\sigma^2 \in R^+$ = variance.

$x \in R$

$$\boxed{PDF = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-1/2\left(\frac{x_i-\mu}{\sigma}\right)^2}}$$

Probability density function

Mean of Normal Distribution

mean (μ) => Average value

variance & std, var = σ^2

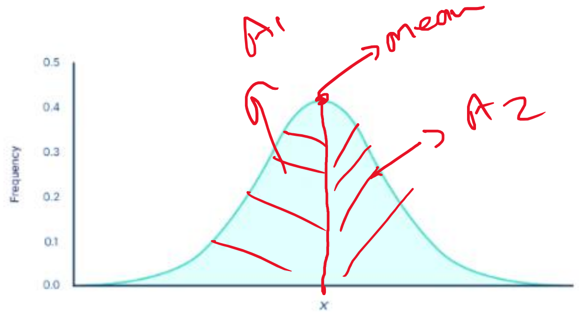
std = $\sqrt{var}$

What are the properties of normal distributions?

Normal distributions have key characteristics that are easy to spot in graphs:

- The mean, median and mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.

$\boxed{\mu \text{ \& std}}$



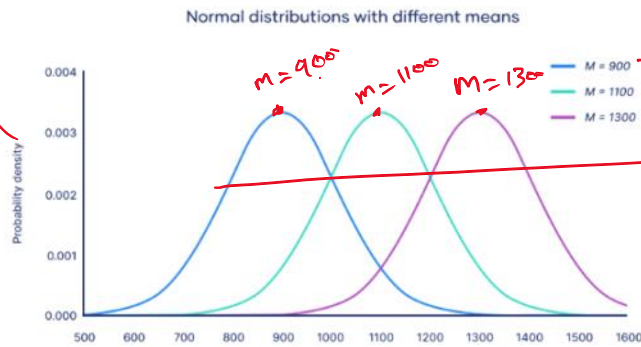
$A_1 = A_2 =$ Area

$\boxed{\mu, \text{ \& std}}$

Note:

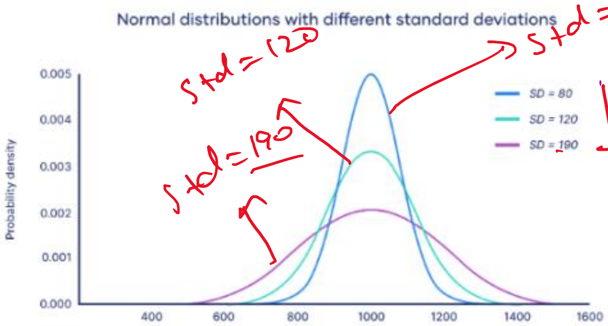
The mean is the location parameter while the standard deviation is the scale parameter.

1. *The mean determines where the peak of the curve is centered. Increasing the mean moves the curve right, while decreasing it moves the curve left.*



$\mu \uparrow \propto$ location

2. *The standard deviation stretches or squeezes the curve. A small standard deviation results in a narrow curve, while a large standard deviation leads to a wide curve.*



Empirical rule

68%, 95%, & 99.7%

The **empirical rule**, or the 68-95-99.7 rule, tells you where most of your values lie in a normal distribution:

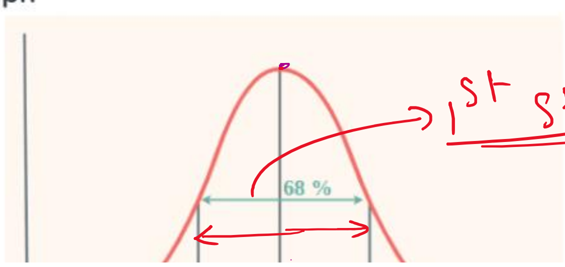
- Around 68% of values are within 1 standard deviation from the mean.
- Around 95% of values are within 2 standard deviations from the mean.
- Around 99.7% of values are within 3 standard deviations from the mean.

1st std

2nd std

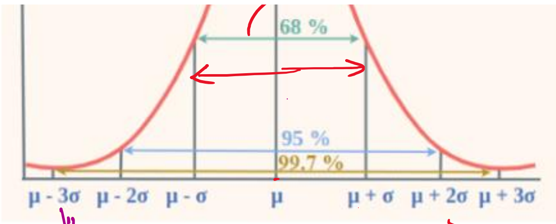
3rd std

Normal Distribution Graph



1st std = 68%

$\sigma^2 =$



$\sigma^2 = \sigma$ std

Studying the graph it is clear that using Empirical Rule we distribute data broadly in three parts. And thus, empirical rule is also called "68 – 95 – 99.7" rule.

68%

95%

99.7

Probability

① $Pr(\mu - \sigma \le x \le \mu + \sigma) \approx 68\%$

② $Pr(\mu - 2\sigma \le x \le \mu + 2\sigma) \approx 95\%$

③ $Pr(\mu - 3\sigma \le x \le \mu + 3\sigma) \approx 99.7\%$

Central Limit Theorem

The **central limit theorem** states that if you take sufficiently large samples from a population, the samples' means will be normally distributed, even if the population isn't normally distributed.

The central limit theorem relies on the concept of a **sampling distribution**, which is the probability distribution of a **statistic** for a large number of samples taken from a population.

Imagining an experiment may help you to understand sampling distributions:

- Suppose that you draw a random sample from a population and calculate a statistic for the sample, such as the mean.
- Now you draw another random sample of the same size, and again calculate the mean.
- You repeat this process many times, and end up with a large number of means, one for each sample.

The distribution of the sample means is an example of a **sampling distribution.**

The central limit theorem says that the sampling distribution of the mean will always be **normally distributed**, as long as the sample size is large enough. Regardless of whether the population has a normal, Poisson, binomial, or any other distribution, the sampling distribution of the mean will be normal.

A normal distribution is a symmetrical, bell-shaped distribution, with increasingly fewer observations the further from the center of the distribution.

What is Z-Score?

(μ & std)

ML

Normalization

100
200
300

-3 to $+3$

Z-score, also known as the **standard score**, **tells us the deviation of a data point from the mean** by expressing it in terms of standard deviations above or below the mean. It gives us an idea of how far a data point is from the mean. Hence, the Z-Score is measured in terms of standard deviation from the mean. For example, a **Z-score of 2 indicates the value is 2** standard deviations **away from the mean**. To use a z-score, we need to know the population mean (μ) and also the population standard deviation (σ).

Z-score is a statistical measure that describes a value's position relative to the mean of a group of values. It is population standard deviation (σ).

-3 to +3

Z-score is a statistical measure that describes a value's position relative to the mean of a group of values. It is expressed in terms of standard deviations from the mean. The Z-score indicates how many standard deviations an element is from the mean.

Z-Score Formula

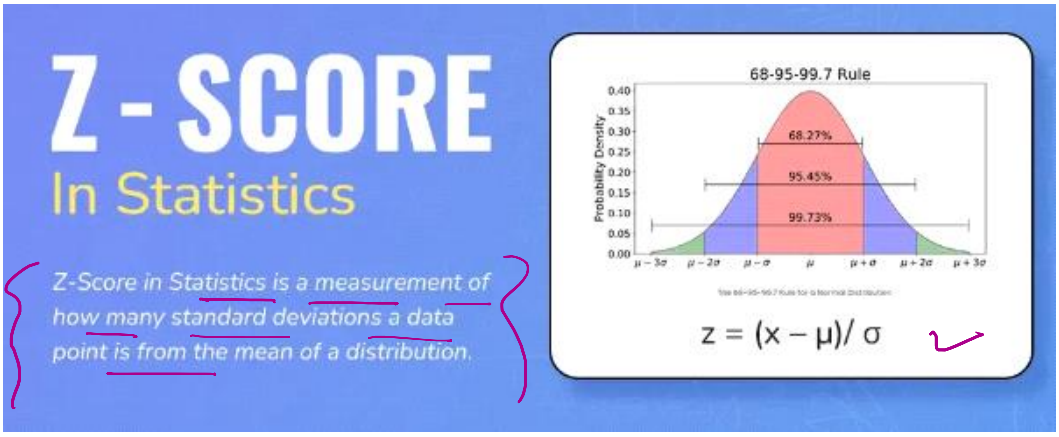
To calculate the z- score for any given data we need the value of the element along with the mean and standard deviation. A z-score can be calculated using the following Z- score formula.

$$z = (X - \mu) / \sigma$$

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

where,

- z = Z-Score
- X = Value of Element
- μ = Population Mean
- σ = Population Standard Deviation



Ex

	A	B	C
1	Factor (x)	Mean (μ)	St. Dev. (σ)
2	3	12.17	6.4
3	13	12.17	6.4
4	8	12.17	6.4
5	21	12.17	6.4
6	17	12.17	6.4
7	11	12.17	6.4

Z score

=> same distribution

-3 to +3

	A	B	C	D
1	Factor (x)	Mean (μ)	St. Dev. (σ)	Z-Score
2	3	12.17	6.4	-1.43

1	Factor (x)	Mean (μ)	St. Dev. (σ)	Z-Score
2	3	12.17	6.4	-1.43
3	13	12.17	6.4	0.13
4	8	12.17	6.4	-0.65
5	21	12.17	6.4	1.38
6	17	12.17	6.4	0.75
7	11	12.17	6.4	-0.18

Covariance & correlation

Q: what is relationship b/w x & y

x	y
2	3
4	5
6	7
8	9

(a) $x\uparrow$ $y\uparrow$

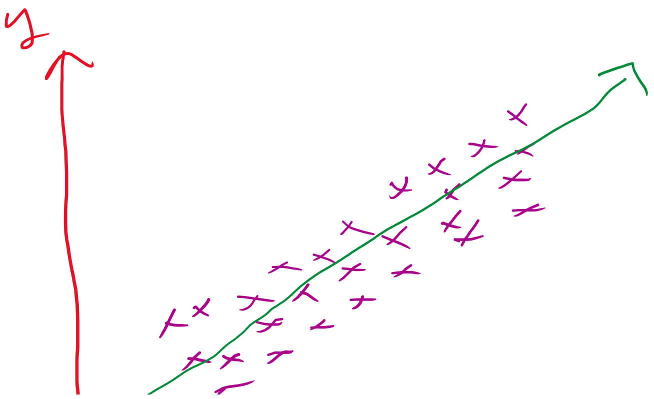
(b) $x\downarrow$ $y\downarrow$

(c) $x\uparrow$ $y\downarrow$

(d) $x\downarrow$ $y\uparrow$

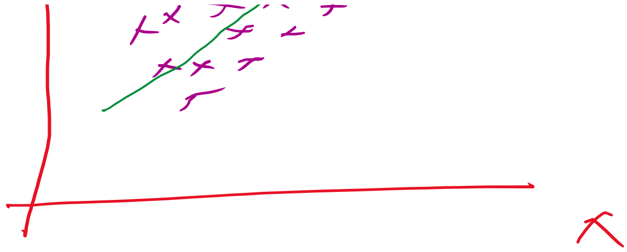
① $\downarrow\uparrow x \propto y \uparrow\downarrow$

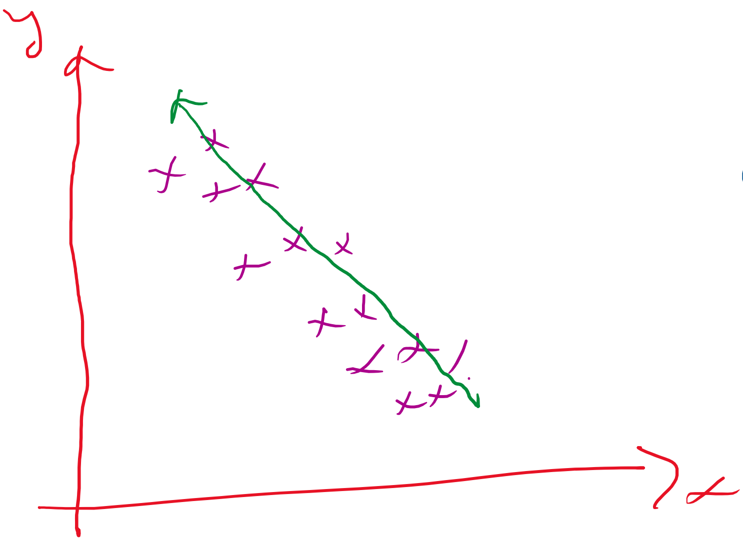
② $\downarrow\uparrow x \propto \frac{1}{y\downarrow\uparrow}$



① $x\uparrow \propto y\uparrow$

$x\downarrow \propto y\downarrow$





2

$\uparrow x \quad y \downarrow$

$\downarrow x \quad y \uparrow$

$x \propto \frac{1}{y}$

Covariance

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

$x_i \rightarrow$ Data point of x

$\bar{x} \rightarrow$ Sample mean

$y_i \rightarrow$ Data points of y

$\bar{y} \rightarrow$ Sample mean of y

1 +ve Covariance

$cov(x, y) \rightarrow \underline{+ve}$

$$\boxed{\begin{array}{cc} x\uparrow & y\uparrow \\ x\downarrow & y\downarrow \end{array}}$$

$x \propto y$

(2) $\underline{-ve \text{ Covariance}}$

$$\boxed{\begin{array}{cc} x\uparrow & y\downarrow \\ x\downarrow & y\uparrow \end{array}}$$

$x \propto \frac{1}{y}$

Ex

x	y
2	3
4	5
6	7

$\bar{x} = 4$ $\bar{y} = 5$

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

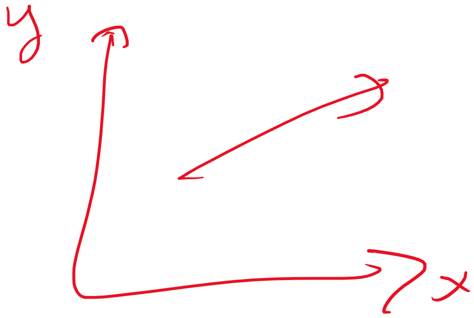
$$\Rightarrow \frac{[(2-4)(3-4) + (4-4)(5-5) + (6-4)(7-5)]}{n-1}$$

$$= \frac{4+0+4}{2} = \frac{8}{2} = 4$$

4

+ve

covariance



(x & y having +ve covariance)

Note

(i) Relationship x & y (+ve or −ve)

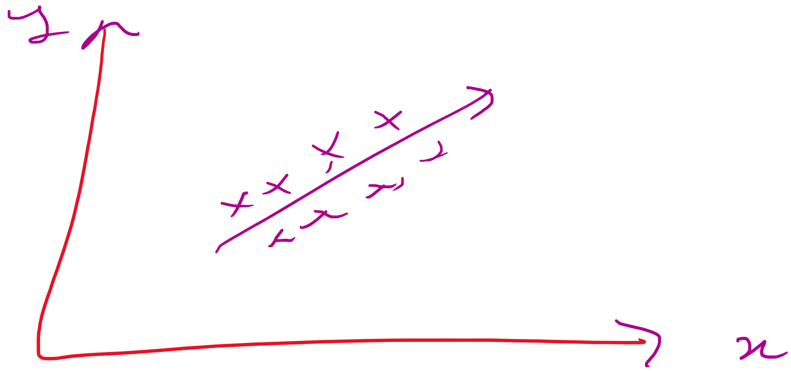
(ii) covariance does not have a specific limit value

Pearson Correlation coefficient

[−1 to 1)

$$\boxed{\rho_{(x,y)} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}}$$

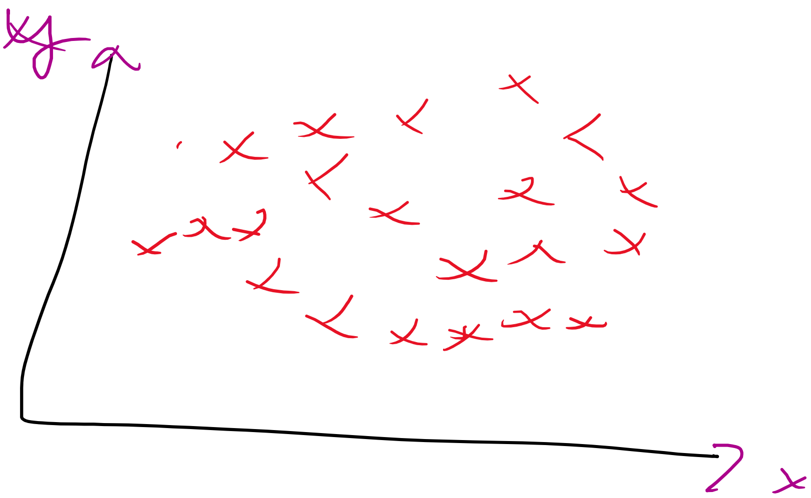
① Between 0 & 1 → +ve correlation



② 0 ⇒ No correlation

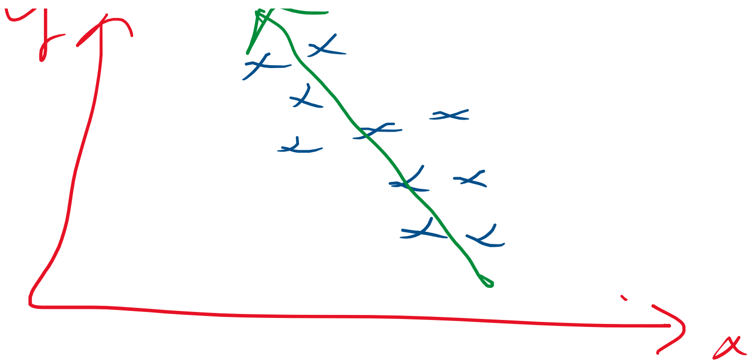
⇓

No relationship



③ Between 0 & −1

Negative correlation


y
x