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**

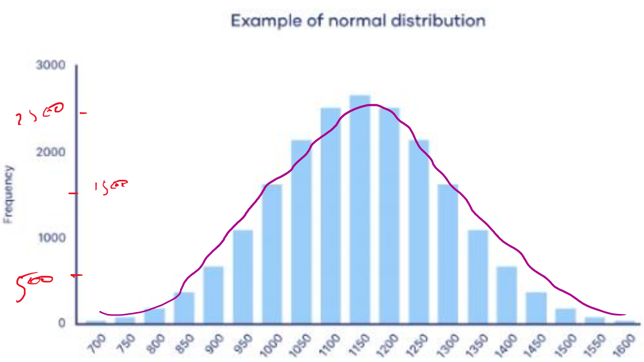
- **Normal or Gaussian Distribution**

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.

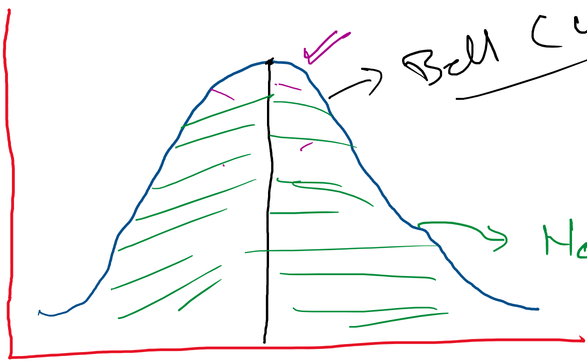
Normal distributions are also called Gaussian distributions or bell curves because of their shape.

Eg: Height
=> weight

-> Age
-> IQ



mean ≠ median ≠ mode



Bell Curve

mean = median = mode = μ

Normal / Gaussian Distribution

Notation: $N(\mu, \sigma^2)$

Parameters: $\mu \in R$ = mean

$\sigma^2 \in R > 0$ = variance

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

mean (μ) = Average Value

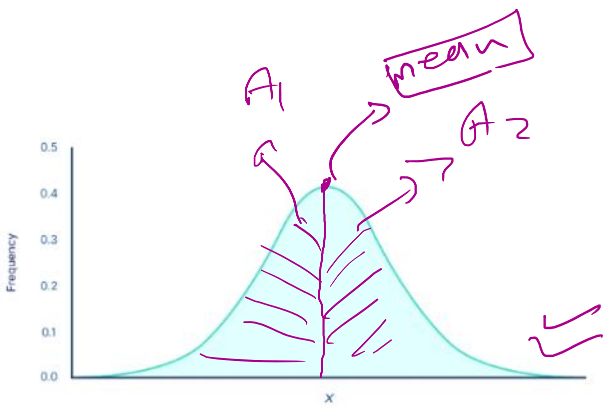
variance & std

$var = \sigma^2$ $\qquad$ $std = \sqrt{var}$

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.



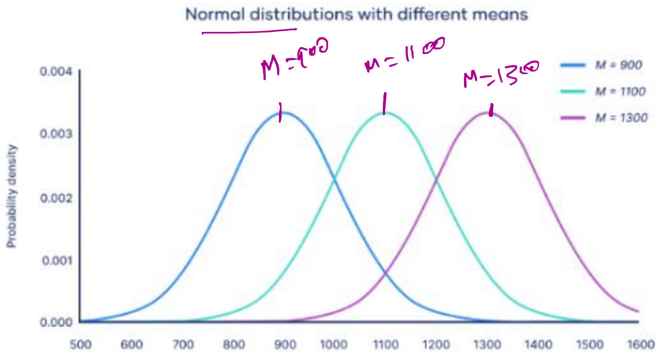
$A_1 = A_2 = $ Area

(μ, std)

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.*

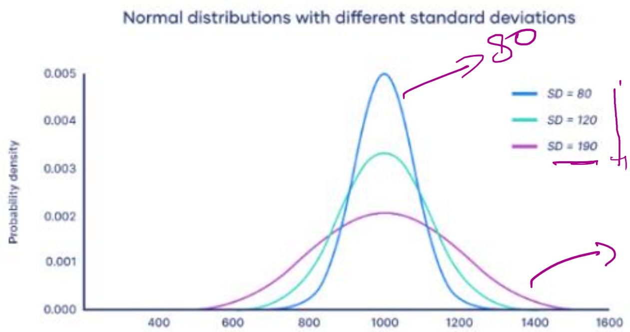
M



M↑

M↓

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.*



↑std → wide curve

↓std → narrow

Empirical rule

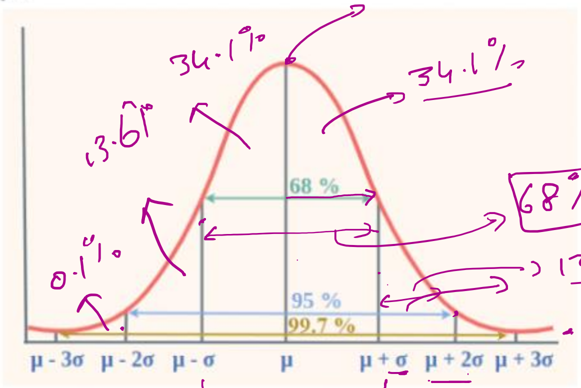
The **empirical rule**, or the 68-95-99.7 rule, tells you where most of your values lie in a normal distribution:

- 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.

Normal Distribution Graph

mean

1st → std σ



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.**

$\boxed{68\%}$

$$Pr(\mu-\sigma \le x \le \mu+\sigma) \approx 68\%$$
$$Pr(\mu-2\sigma \le x \le \mu+2\sigma) \approx 95\%$$
$$Pr(\mu-3\sigma \le x \le \mu+3\sigma) \approx 99.75\%$$

eg IRIS → (Sepal width, Sepal length) →

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.

$A_1 \neq A_2$

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.

Population

$A_1 \neq A_2$ pdf + Area under curve

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.

Pdf → Probability distribution curve

Covariance & Correlation

Q what is relationship, x & y

x	y
2	3
4	5
6	7
8	9

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

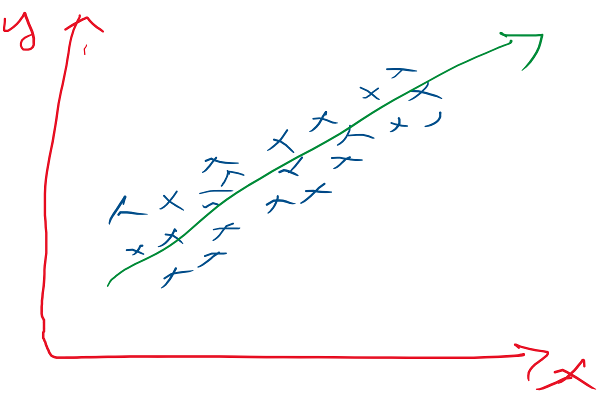
(b) $x\downarrow$ $y\uparrow$

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

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

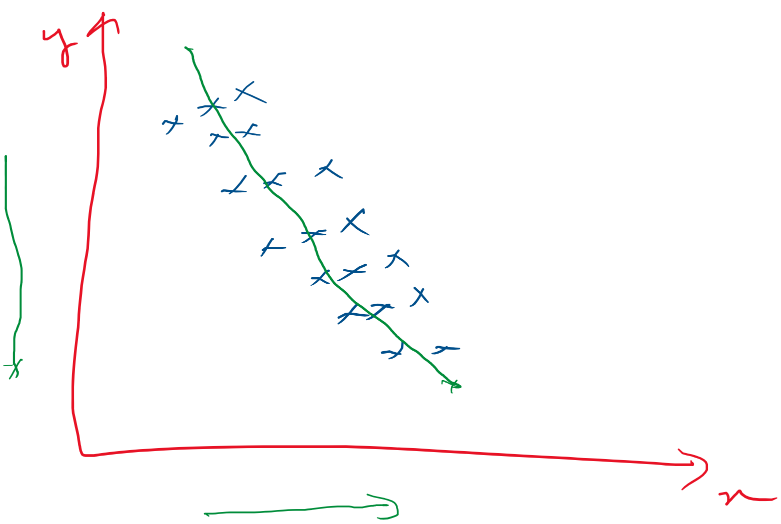
$x \propto y$

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



a) $x\uparrow$ $y\uparrow$
$x\downarrow$ $y\downarrow$

$x \propto y$



b) $x\uparrow$ $y\downarrow$
$x\downarrow$ $y\uparrow$

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

Covariance

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

x_i = Datapoints of x
$\bar{x}$ → Sample mean
y_i → Datapoint of y
$\bar{y}$ → Sample mean of y

cov(x,y)

$x\uparrow$	$y\uparrow$
$x\downarrow$	$y\downarrow$

→ +ve covariance

$x\uparrow$	$y\downarrow$
$x\downarrow$	$y\uparrow$

→ −ve covariance

eg

x	y
2	3
4	5
6	7
$\bar{x}=4$	$\bar{y}=5$

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

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

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

+ve covariance

{ x & y having +ve covariance }

+ve covariance

+ve covariance

Note (i) Relationship x & y
(+ve or -ve value)

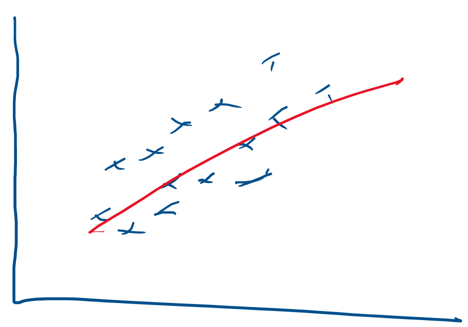
(ii) covariance does not have a specific limit value.

Pearson Correlation Coefficient (r)

[-1 to 1]

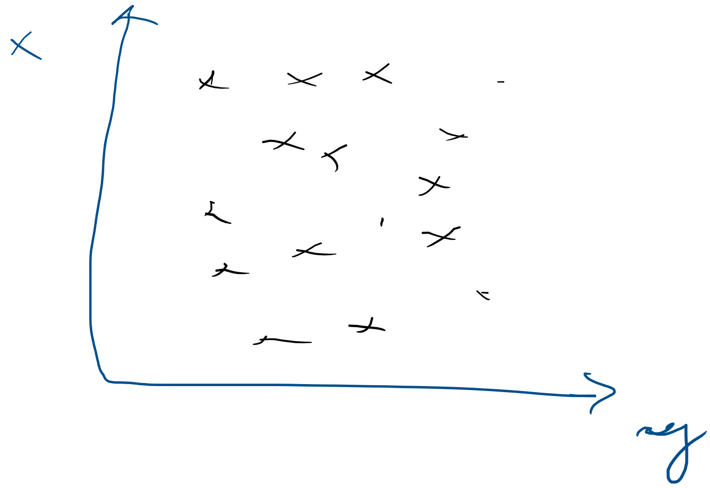
$$\boxed{\rho_{(x,y)} = \frac{cov(x,y)}{\sigma_x \sigma_y}}$$

① Between 0 & 1 -> Positive correlation



② 0 No correlation

No relationship



(3) Between 0 & -1 Negative correlation