Missing Values.

Missing values occurs in dataset when some of the informations is not stored for a variable There are 3 mechanisms

1 Missing Completely at Random, MCAR:

Missing completely at random (MCAR) is a type of missing data mechanism in which the probability of a value being missing is unrelated to both the observed data and the missing data. In other words, if the data is MCAR, the missing values are randomly distributed throughout the dataset, and there is no systematic reason for why they are missing.

For example, in a survey about the prevalence of a certain disease, the missing data might be MCAR if the survey participants with missing values for certain questions were selected randomly and their missing responses are not related to their disease status or any other variables measured in the survey.

2. Missing at Random MAR:

Missing at Random (MAR) is a type of missing data mechanism in which the probability of a value being missing depends only on the observed data, but not on the missing data itself. In other words, if the data is MAR, the missing values are systematically related to the observed data, but not to the missing data. Here are a few examples of missing at random:

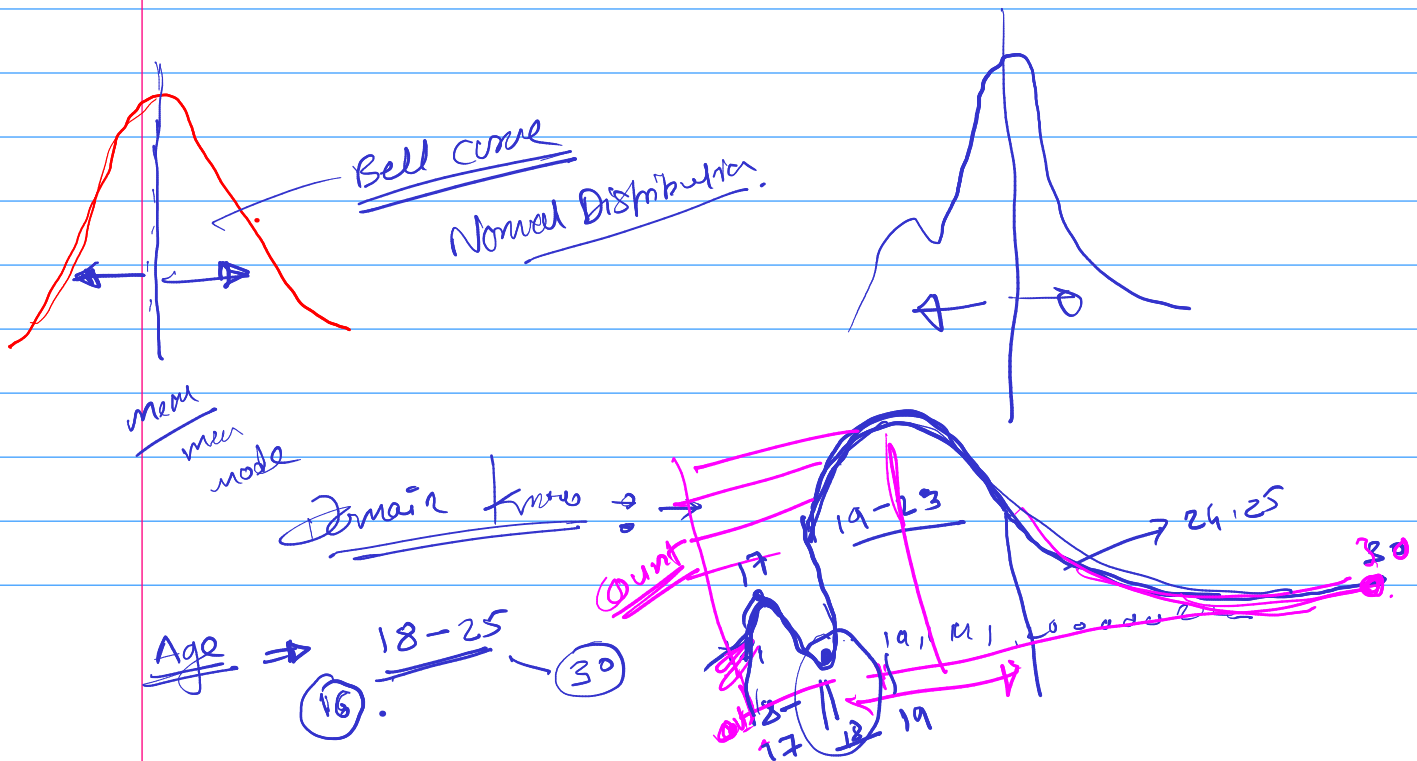
Income data: Suppose you are collecting income data from a group of people, but some participants choose not to report their income. If the decision to report or not report income is related to the participant's age or gender, but not to their income level, then the data is missing at random.

Medical data: Suppose you are collecting medical data on patients, including their blood pressure, but some patients do not report their blood pressure. If the patients who do not report their blood pressure are more likely to be younger or have healthier lifestyles, but the missingness is not related to their actual blood pressure values, then the data is missing at random.

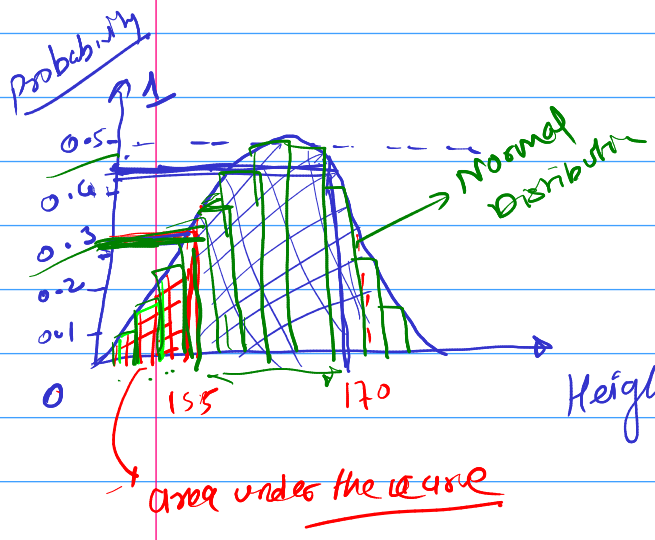
3. Missing data not at random (MNAR)

It is a type of missing data mechanism where the probability of missing values depends on the value of the missing data itself. In other words, if the data is MNAR, the missingness is not random and is dependent on unobserved or unmeasured factors that are associated with the missing values.

For example, suppose you are collecting data on the income and job satisfaction of employees in a company. If employees who are less satisfied with their jobs are more likely to refuse to report their income, then the data is not missing at random. In this case, the missingness is dependent on job satisfaction, which is not directly observed or measured.



1. **Probability Density Function (PDF):** - It is a statistical term that describes the probability distribution of a continuous random variable. The probability associate with a single value is always Zero. Below is the formula for PDF.



Formula

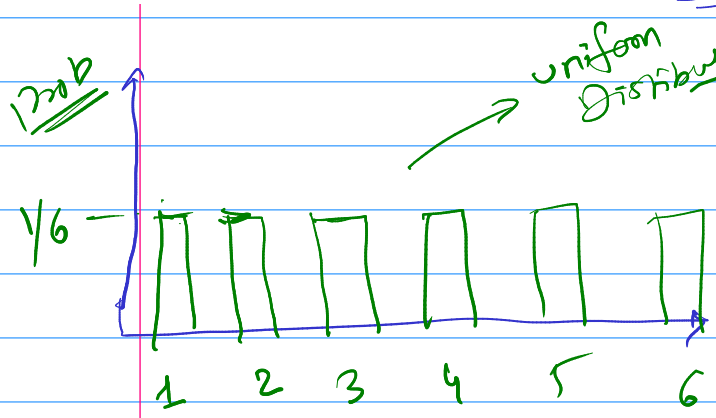
$Pr(H \leq 155) = 0.3$

$= P(H \geq 155 \text{ and } H \leq 170)$

$= 0.3 + 0.5 = 0.8$

5'11

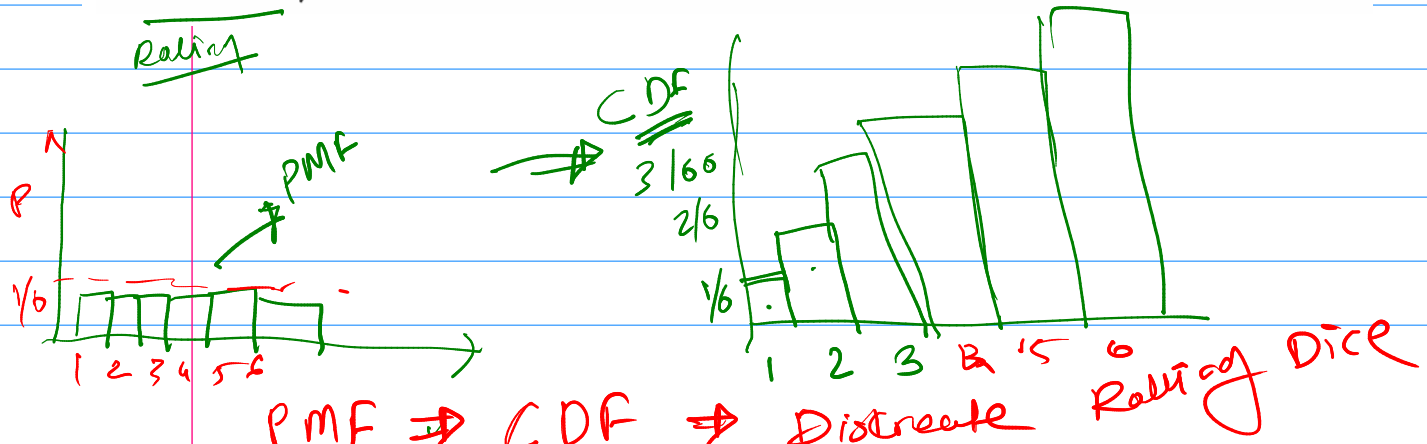
B. **Probability Mass Function (PMF):**- It is a statistical term that describes the probability distribution of a discrete random variable.



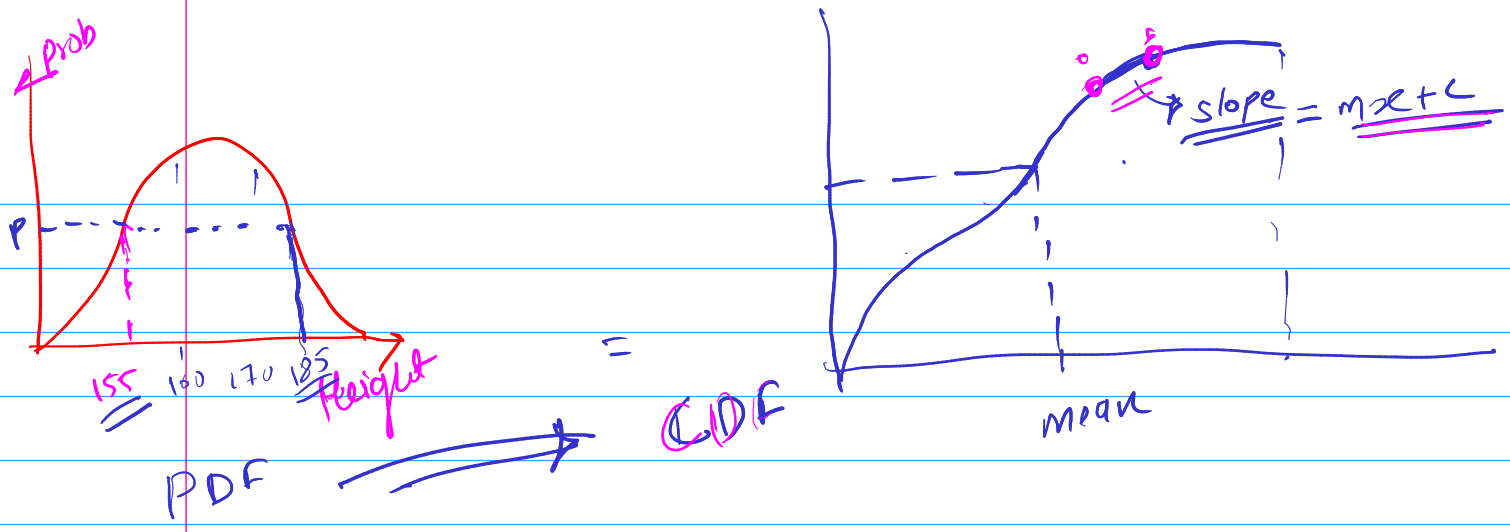
$P(1) = 1/6$
$P(6) = 1/6$
} example Dice

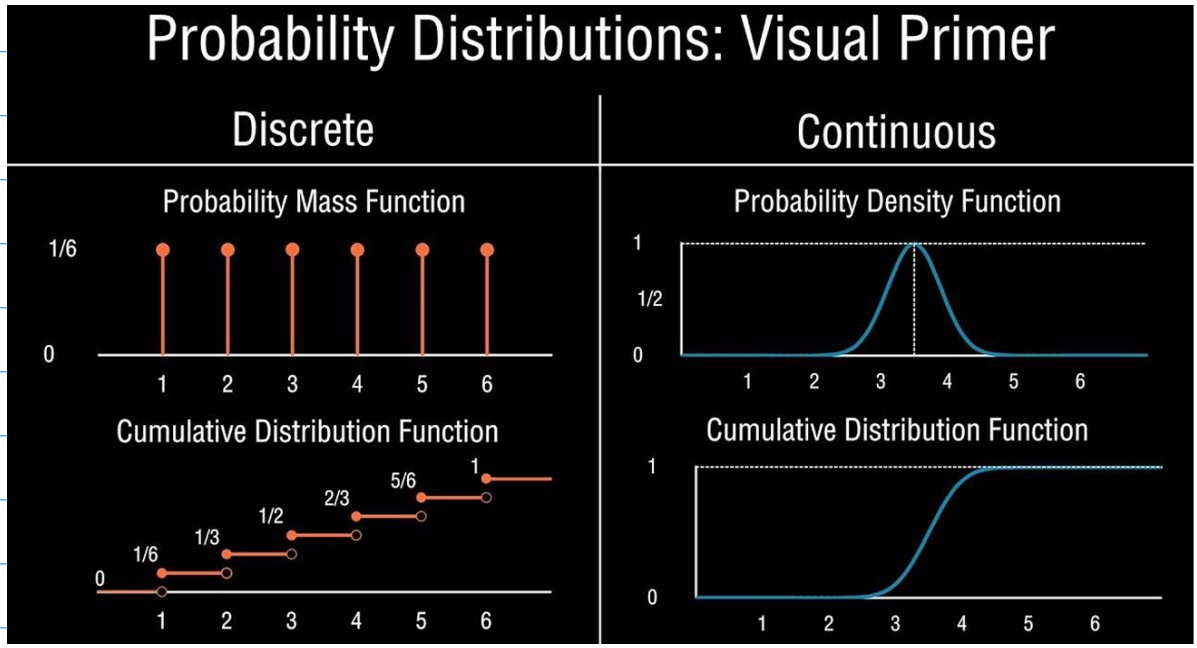
$P(x \leq 4) = P(1) + P(2) + P(3) + P(4)$
$= 1/6 + 1/6 + 1/6 + 1/6$
$= 2/3$

C. **Cumulative Distribution Function (CDF):**- It is another method to describe the distribution of a random variable (either continuous or discrete).



PMF ⇒ CDF ⇒ Discrete





Types of Probability Distribution: -

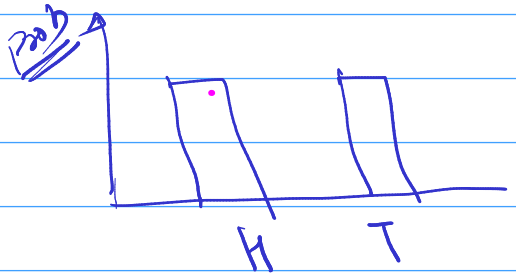
1. Normal or Gaussian Distribution
2. Bernoulli Distribution
3. Uniform Distribution
4. Poisson Distribution
5. Binomial Distribution
6. Log-Normal Distribution

1. Bernoulli Distribution: -

- Bernoulli distribution is a discrete probability distribution
- it's concerned with discrete random variables {PMF}
- Bernoulli distribution applies to events that have **one trial** and **two possible outcomes**. These are known as Bernoulli trials.

$P(H) = 0.5 = p$

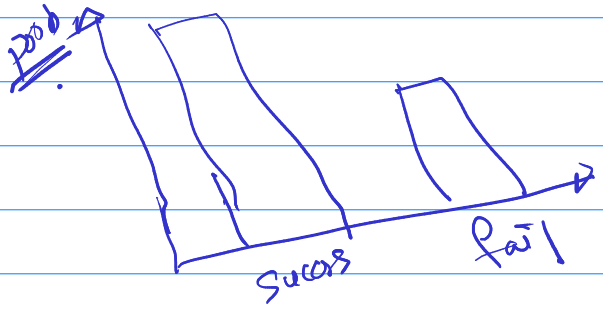
$P(T) = 0.5 = 1 - p = q$



Fail/Success

$P(Pass) = 0.85 = p$

$P(fail) = 1 - p = 1.85 = q$



2. Binomial Distribution: -

- it's concerned with discrete random variables {PMF}
- There are two possible outcomes: true or false, success or failure, yes or no.
- These Experiments is Performs for n trials
- Every trial is an independent trial, which means the outcome of one trial does not affect the outcome of another trial.

eg - Tossing a Coin 100 times.

3. Poisson Distribution: -

- it's concerned with discrete random variables {PMF}
- Describe the number of events occurring in a fixed time interval

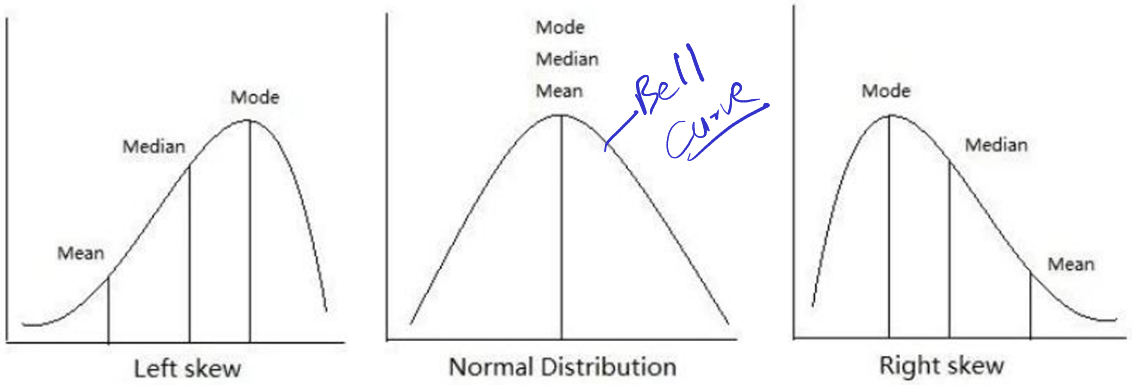
E.g.: - No. of people visiting hospital every hour
No. of people visiting bank at 11am

4. **Normal or Gaussian Distribution: -**

- it's concerned with Continuous random variables {PDF}
- Normal distributions are symmetrical, but not all symmetrical distributions are normal

Characteristics of Normal Distribution

- mean = median = mode
- Symmetrical about the center
- Unimodal
- 50% of values less than the mean and 50% greater than the mean

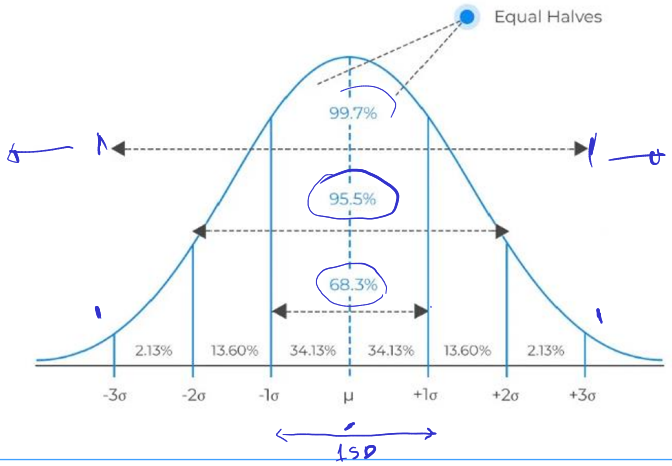


- BP
- Height
- Weight
- Errors
- Marks in test.

mean
median
mode

Skewness — refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data.

Empirical Rule of Normal Distribution: - The empirical rule in statistics, also known as the 68 95 99 rule, states that for normal distributions, 68% of observed data points will lie inside one standard deviation of the mean, 95% will fall within two standard deviations, and 99.7% will occur within three standard deviations.



Standard Normal Distribution Z-Score: - The standard normal distribution is a specific type of normal distribution where the mean is equal to 0 and the standard deviation is equal to 1.

The normal distribution is the most commonly used probability distribution in statistics.

It has the following properties:

- Symmetrical
- Bell-shaped
- Mean and median are equal; both located at the center of the distribution

The mean of the normal distribution determines its location and the standard deviation determines its spread.

What is a "Z-score"?

The number of **standard deviations from the mean** is also called the "Standard Score", "sigma" or "Z-score". Simply, a Z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units.

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

- **Z** is the "z-score" (Standard Score)
- **x** is the value to be standardized
- **μ** (mu) is the mean
- **σ** (sigma) is the standard deviation

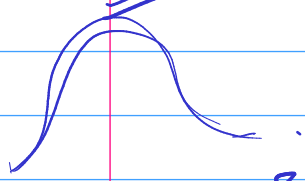
$x = \{1, 2, 3, 4, 5\}$

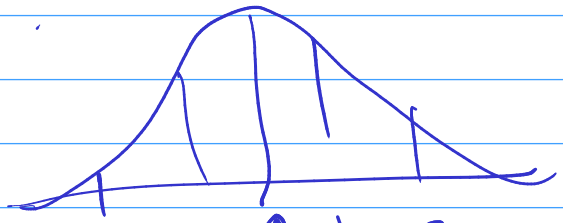
$\mu = \frac{1+2+3+4}{4} = \frac{10}{4} = 2.5 \approx 3$

$\sigma = 1.414 \approx 1$

$z\text{-score} = \frac{x_i - \mu}{\sigma}$

$= \{-2, -1, 0, 1, 2\}$





$z\text{-score} = \frac{1-3}{1} = -2$

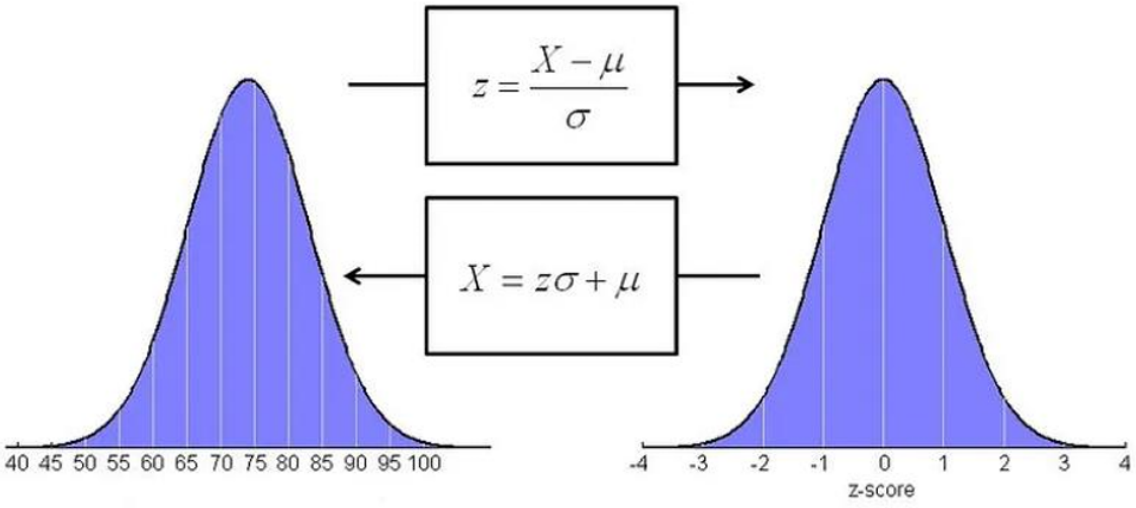
$= \frac{2-3}{1} = -1$

$= \frac{3-3}{1} = 0$

$\mu = 0$

$\frac{4-3}{1} = 1$

$\frac{5-3}{1} = 2$



6. **Log-Normal Distribution:** - A log-normal distribution is a continuous distribution of random variable y whose natural logarithm is normally distributed. For example, if random variable $y = \exp\{y\}$ has log-normal distribution then $x = \log(y)$ has normal distribution.