For continuous random variable X assuming all real numbers and $f(x)$ is its probability density function (p.d.f), then expectation is defined as

$$\text{Var(X)} = \int_{-\infty}^{\infty} x^2 f(x)dx - \left(\int_{-\infty}^{\infty} xf(x)dx\right)^2, \quad \text{provided the integrals exist.}$$

Binomial distribution: The Binomial distribution of a random variable X is defined by the p.m.f $f(x)$ given by

$$f(x) = \binom{n}{x} p^x q^{n-x},\ x = 0, 1, 2, ...n \in \mathbb{N},\ 0 < p < 1,\ p + q = 1.$$

Binomial distribution is symbolically expressed as $X \sim B(n, p)$. The constants n and p are called parameters of the binomial distribution.
Here, we find the mean and variance of Binomial distribution.

Property: If $X \sim B(n, p)$, then $E(X) = np$ and $Var(X) = npq$.

We know

$$\begin{aligned}
E(X) &= \Sigma x_i f(x_i) \\
&= \Sigma_{i=0}^{n} i \binom{n}{i} p^i (1-p)^{n-i} \\
&= \Sigma_{i=1}^{n} i \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \\
&= np\ \Sigma_{i=1}^{n} \frac{(n-1)!}{(i-1)!(n-i)!} p^{i-1} (1-p)^{n-i} \\
&= np\ \Sigma_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-i-1},\ j = i-1 \\
&= np\ (p + 1 - p)^{n-1} \\
&= np.
\end{aligned}$$

Again, we have

$$\begin{aligned}
E(X(X-1)) &= \Sigma_{i=0}^{n} i(i-1) \binom{n}{i} p^i (1-p)^{n-i} \\
&= n(n-1)p^2\ \Sigma_{i=2}^{n} \binom{n-2}{i-2} p^{i-2} (1-p)^{n-i} \\
&= n(n-1)p^2\ \Sigma_{j=0}^{n-2} \binom{n-2}{j} p^j (1-p)^{n-j-2},\ j = i-2 \\
&= n(n-1)p^2\ (p + 1 - p)^{n-2} = n(n-1)p^2.
\end{aligned}$$

Now, $Var(X) = E(X(X-1)) - E(X)(E(X)-1)$
$= n(n-1)p^2 - np(np-1)$

$$= np(np - p - np + 1)$$
$$= np(1 - p) = npq.$$

Example on Binomial distribution

Example: Two fair dice are rolled 100 times. find the probability of getting at least once a double six.

Solution: Casting two dice 100 times may be considered as 100 trials, each trials resulting two outcomes, viz., getting and not getting a double six.
Hence, if X denotes the number of times a double six is obtained, then clearly $X \sim B(100, \frac{1}{36})$, since the probability of getting a double six in a single throw is $\frac{1}{36}$.

$$\begin{aligned} \therefore P(\text{at least double six}) &= P(X \geq 1) \\ &= 1 - P(X = 0) \\ &= 1 - \binom{100}{0}\left(\frac{1}{36}\right)^0\left(\frac{35}{36}\right)^{100} \\ &= 1 - \left(\frac{35}{36}\right)^{100}. \end{aligned}$$

Example: What is the probability of obtaining multiples of 3, twice in a throw of 6 dice?

Solution: Let p = probability of getting a multiple of 3= $\frac{2}{6} = \frac{1}{3}$.
$\therefore q = 1 - p = 1 - \frac{1}{3} = \frac{2}{3}$. Here $n = 6$ and $x = 2$.
$\therefore$, by Binomial law, the required probability is

$$\begin{aligned} \binom{n}{x}p^x q^{n-x} &= \binom{6}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^{6-2} \\ &= \frac{5 \times 2^4}{3^5} \end{aligned}$$

Example: 10 % of screws produced in a certain factory turn out to be defective. Find the probability that in a sample of 10 screws chosen at random, exactly two will be defective.

Solution: Here, p = probability of a defective=$\frac{10}{100} = \frac{1}{10}$ and $q = 1 - \frac{1}{10} = \frac{9}{10}$, $n = 10$, $x = 2$. Then by Binomial law, the required probability is

$$\begin{aligned} \binom{n}{x}p^x q^{n-x} &= \binom{10}{2}\left(\frac{1}{10}\right)^2\left(\frac{9}{10}\right)^{10-2} \\ &= \frac{1}{2}\left(\frac{9}{10}\right)^9. \end{aligned}$$

Example: The probability that a man aged 60 will live to be 70 is 0.65. What is the probability that out of 10 men, now 60, at least 7 will live to be 70?

Solution: The probability that a man aged 60 will live to 70=p=0.65.
$\therefore q = 1 - p = 1 - 0.65 = 0.35$ and $n = 10$.
The probability that at least 7 man will live to 70

$= \binom{10}{7}p^7q^3 + \binom{10}{8}p^78q^2 + \binom{10}{9}p^9q^1 + \binom{10}{10}p^10q^0$

$= \frac{10!}{7!3!}(0.65)^7(0.35)^3 + \frac{10!}{8!2!}(0.65)^8(0.35)^2 + \frac{10!}{9!1!}(0.65)^9(0.35)^1 + \frac{10!}{10!0!}(0.65)^10(0.35)^0$

$= 0.5137.$

Example: In a basket there are 1 red, 2 white and 3 black balls. One ball is drawn three times in succession and each time the ball is being replaced before the next draw.Find the probability that (i) All balls are white (ii) two balls are white?

Solution: From the condition of the problem it is evident that trials are independent. The probability of drawing a white ball in a trial is 2/6=1/3. Drawing of a white ball may be considered as success. Thus $p = 1/3$ and $q = 1 - p = 2/3$.

(i) Here, $p = 1/3, q = 2/3, n = 3, x = 2$.

Hence the required probability is $\binom{3}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^0 = \frac{1}{27}$.

(ii) Here $p = 1/3, q = 2/3, n = 3, x = 2$.

Hence, the required probability is $\binom{3}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^1 = \frac{2}{9}$.

Example: The probability that a fighter plane will return safely after an operation is 0.95. Find the probability that the plane fails to survive in 5 operations.

Solution: The probability that the plane will return safely in all five operations is $(0.95)^5$ (as the operations are mutually independent).
The event 'plane fails to survive' means it will be destroyed in one of the operations. Hence, the events 'plane will safely return in all five operations' and 'plane fails to survive' are complementary to each other.
$\therefore P$(plain will fail to survive) $= 1 - (0.95)^5$.

Poisson Distribution: A discrete random variable X having enumerable set $\{0, 1, 2, ...\}$ as the spectrum is said to have *Poisson distribution* with parameter $\mu(> 0)$, if the

p.m.f is given by

$$\begin{aligned} f(x) &= \frac{e^{-\mu}\mu^x}{x!}, \quad \text{for } x = 0, 1, 2, ... \\ &= 0, \text{elsewhere.} \end{aligned}$$

Now we find the mean and variance of Poisson distribution.

Property: If $X \sim P(\mu)$, then $E(X) = \mu$ and $\text{Var}(X) = \mu$.

Then

$$\begin{aligned} \text{Mean} &= E(X) \\ &= \Sigma_{i=0}^{\infty}\ ie^{-\mu}\ \frac{\mu^i}{i!} = \mu e^{-\mu}\ \Sigma_{i=1}^{\infty}\frac{\mu^{i-1}}{(i-1)!} \\ &= \mu e^{-\mu}\ \Sigma_{j=0}^{\infty}\frac{\mu^j}{j!}, j = i-1 \\ &= \mu e^{-\mu}e^{\mu} = \mu. \end{aligned}$$

Again, we have

$$\begin{aligned} E\{X(X-1)\} &= \Sigma_{i=0}^{\infty} i(i-1)e^{-\mu}\frac{\mu^i}{i!} \\ &= e^{-\mu}\mu^2\ \Sigma_{i=2}^{\infty}\frac{\mu^{i-2}}{(i-2)!} \\ &= e^{-\mu}\mu^2\ \Sigma_{j=0}^{\infty}\frac{\mu^j}{j!}, \quad j = i-2 \\ &= e^{-\mu}\mu^2 e^{\mu} = \mu^2 \end{aligned}$$

Hence, $Var(X) = E\{X(X-1)\} - E(X)(E(X)-1) = \mu^2 - \mu(\mu-1) = \mu$.

Example: A hospital switchboard receives on average 4 emergency calls in a five-minute interval. What is the probability that here are (i) at most two emergency calls in a five-minute interval, (ii) exactly 3 emergency calls in a five minute interval?

Solution: Let X denote the number of calls received in a five minute interval. We know that $X \sim P(\mu)$, where μ is the average number of calls in a five-minute interval. Here $\mu = 4$.
Now

$$\begin{aligned} P(\text{at most 2 emergency calls}) &= P(X \leq 2) \\ &= P(X=0) + P(X=1) + P(X=2) \\ &= \frac{e^{-4}4^0}{0!} + \frac{e^{-4}4^1}{1!} + \frac{e^{-4}4^2}{2!} = 13e^{-4} \end{aligned}$$

Again, P(exactly 3 emergency calls)=$P(X=3)$

$$= \frac{e^{-4}4^3}{3!} = \frac{32}{3}e^{-4}.$$

Example: The number of emergency admission each day to a hospital is found to have a Poisson distribution with parameter 2. (i) Evaluate the probability that on a particular day there will be no emergency admission. (ii) At the beginning of one day the hospital has five beds available for emergency. Calculate the probability that this will be an insufficient number for the day.

Solution: The probability for i admissions on any day= $P(X = i) = \frac{e^{-2}2^i}{i!}$, i is a non-negative integer.
(a) Required probability $=P(X = 0) = e^{-2}$.
(b) Required probability $= P(X > 5) = 1 - P(X \leq 5)$
$= 1 - e^{-2}\left(1 + 2 + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!} + \frac{2^5}{5!}\right) \approx 0.0166.$

Example: The probability of a product produced by a machine to be defective is 0.01. If 30 products are taken at random, find the probability that exactly 2 will be defective. Approximate by Poisson distribution and evaluate the error in the approximation.

Solution: Since the probability of success is small, we approximate by Poisson distribution, the parameter of the distribution being $\mu = np = 30 \times 0.01 = 0.3$.
Hence the probability of getting exactly 2 defective

$$= \frac{\mu^2}{2!}\ e^{-\mu} = \frac{(0.3)^2}{2!}\ e^{-0.3} = 0.03337.$$

Example: If there is a war every 15 years on the average, then find the probability that there will be no war in 25 years.

Solution: λ= number of changes per unit of time on the average=$\frac{1}{15}$. Let X be the random variable denoting the number of wars in the interval $(0, 25)$, when the unit of time is one year, then X is Poisson distributed with parameter $\mu = \lambda t = \frac{1}{15} \times 25 = \frac{5}{3}$.
$\therefore$ probability of no war in the given interval of time

$$= P(X = 0) = \frac{e^{-\mu}\mu^0}{0!} = e^{-\frac{5}{3}}.$$

Example: A car-hire firm has two cars, which it hires out by the day. The number of demands for a car on each day is Poisson distributed with parameter 1.5. Calculate th proportion of days on which neither of the cars is used, and the proportion of days on which some demand cannot be met for lack of cars.

Solution: Let X be the random variable denoting the number of demands for a car on nay day. Then X is Poisson distributed with parameter 1.5.

(a) Proportion of days on which neither car is used

$= P(X = 0) = e^{-1.5} = 0.223.$

(b) Proportion of days on which some demands is refused

$= P(X > 2) = 1 - P(X \leq 2)$
$= 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\}$
$= 1 - e^{-1.5}\left\{1 + 1.5 + \frac{(1.5)^2}{2}\right\}$
$= 0.1916.$

Normal Distribution The most commonly encountered physical phenomena provide plenty of normal distributions. Carl Fredrich Gauss (1777-1855) while studying the nature of errors made in any scientific measurement came out with a peculiar distribution, presently, known as *Gaussian distribution or normal distribution.* This distribution plays a vital role in statistical decision theory and estimation.
The normal distribution of a random variable X with parameter μ or m and $\sigma(> 0)$ is defined by the probability function $f(x)$ given by

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

A random variable with the above probability function is called normal variate. The fact that X follows normal distribution with parameter μ and σ is expressed symbolically by $X \sim N(\mu, \sigma^2)$.
In particular, the random variable Z with $\mu = 0$ and $\sigma = 1$ is called the *standard normal variate.* Thus, if Z is the standard normal variate then $Z \sim N(0, 1)$.

Property If $X \sim N(\mu, \sigma^2)$, then $E(X) = \mu$ and $Var(X) = \sigma^2$.
To see the above, we note that

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{1.6}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{1.7}$$

Putting $\frac{x-\mu}{\sqrt{2}\sigma} = t$, we see $dx = \sqrt{2}\ \sigma dt$ and $t \to \infty$ as $x \to \infty$, also $t \to -\infty$ as $x \to -\infty$.

$$\therefore E(X) = \frac{1}{\sqrt{2\pi}\sigma} \sqrt{2}\sigma \int_{-\infty}^{\infty} (\mu + \sqrt{2}\sigma t) e^{-t^2} dt$$

$$= \frac{\mu}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt + \frac{\sqrt{2}\sigma}{\sqrt{\pi}} \int_{-\infty}^{\infty} t e^{-t^2} dt$$

$$= \frac{\mu}{\sqrt{\pi}} \times \sqrt{\pi} + \frac{\sqrt{2}\sigma}{\sqrt{\pi}} \times 0 = \mu$$

since te^{-t^2} is and odd function and $\int_{-\infty}^{\infty} e^{-t^2} dt = \sqrt{\pi}$.

Now

$$E(X^2) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} x^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{1.8}$$

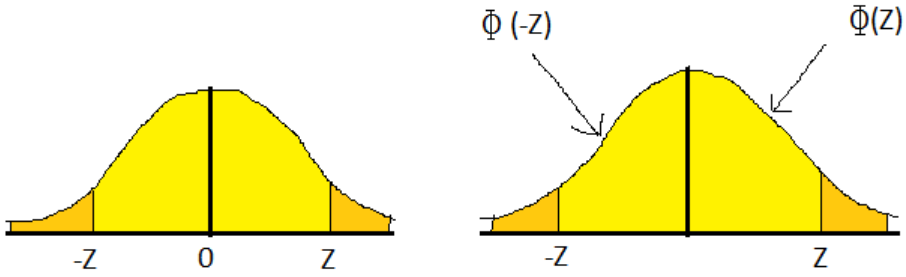
$$= \frac{1}{\sqrt{2\pi}\sigma}\sqrt{2}\sigma\int_{-\infty}^{\infty}(\mu+\sqrt{2}\sigma t)^2 e^{-t^2} dt$$

$$= \frac{\mu^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-t^2} dt + \frac{2\sqrt{2}\mu\sigma}{\sqrt{\pi}}\int_{-\infty}^{\infty} te^{-t^2} dt + \frac{2\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} t^2 e^{-t^2} dt$$

$$= \mu^2 + 0 + \sigma^2, \text{ since } \int_{-\infty}^{\infty} t^2 e^{-t^2} dt = \frac{1}{2}\sqrt{\pi}. \tag{1.9}$$

Hence,
$$\begin{aligned} Var(X) &= E(X^2) - \{E(X)\}^2 \\ &= \mu^2 + \sigma^2 - \mu^2 \\ &= \sigma^2. \end{aligned}$$

The Normal Probability Table A table showing the probability that Z is less than or equal to a particular value of z is known as the probability table. The value is usually denoted by $\Phi(z)$. Thus,



$$\phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{Z} e^{-\frac{t^2}{2}}\, dt$$

Clearly,

$$\Phi(0) = 0.5,\ \Phi(-z) = 1 - \Phi(z) \text{ by symmetry.}$$

Also

$$P(a \le x \le b) = \Phi(b) - \Phi(a) \text{ since } P(X = a) = 0.$$

Example. If $X \sim N(30, 25)$, then find

(i)$P(X \ge 45)$, (ii) $P(26 \le X \le 40)$, (iii) $P(|X - 30| > 5)$.

Solution: Since $X \sim N(30, 25)$, $Z = \frac{X-30}{5} \sim N(0, 1)$.

Now,

$$\begin{aligned} P(X \geq 45) &= P\left(\frac{X-30}{5} \geq \frac{45-30}{5}\right) \\ &= P(Z \geq 3) = 1 - P(Z < 3) = 1 - P(Z \leq 3), \text{ since } P(Z = 3) = 0 \\ &= 1 - \Phi(3) = 1 - 0.99865 \text{ from normal probability table} \\ &= 0.00135. \end{aligned}$$

Again,

$$\begin{aligned} P(26 \leq X \leq 40) &= P\left(\frac{26-30}{5} \leq \frac{X-30}{5} \leq \frac{40-30}{5}\right) \\ &= P(-0.8 \leq Z \leq 2) \\ &= \Phi(2) - \Phi(-0.8) \\ &= 0.7653 \end{aligned}$$

Finally,

$$\begin{aligned} P(|X-30| > 5) &= P\left(\left|\frac{X-30}{5}\right| > 1\right) \\ &= P(Z > 1) \\ &= 1 - P(|Z| \leq 1) = 1 - P(-1 \leq Z \leq 1) \\ &= 1 - \{\Phi(1) - \Phi(-1)\} \\ &= 0.3174. \end{aligned}$$

Example. Assuming that the lifespan of a type of transistor is normal, find the mean and standard deviation if 84 % of the transistors have lifespan less than 65.2 months and 68 % have lifespan lying between 65.2 and 62.8 months.

Solution: Let X denote lifespan of the said type of transistor and let its mean be μ and variance be σ^2. By the assumption $X \sim N(\mu, \sigma^2)$.

$$P(X \leq 65.2) = 0.84 \text{ and } P(62.8 \leq X \leq 65.2) = 0.68.$$

Then,

$$p\left(\frac{X-\mu}{\sigma} \leq \frac{65.2-\mu}{\sigma}\right) = 0.84 \text{ and } p\left(\frac{62.8-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{65.2-\mu}{\sigma}\right) = 0.68$$

or,

$$P\left(Z \leq \frac{65.2-\mu}{\sigma}\right) = 0.84 \text{ and } p\left(Z \leq \frac{62.8-\mu}{\sigma}\right) = 0.16$$

But from the normal probability table,

$$P(Z \leq 0.9) = 0.84 \text{ and } P(Z \leq -0.9) = 0.16.$$

Hence,

$$\frac{65.2-\mu}{\sigma} = 0.9 \text{ and } \frac{62.8-\mu}{\sigma} = -0.9$$

or,

$$\mu + 0.9\sigma = 65.2 \text{ and } \mu - 0.9\sigma = 62.8.$$

$$\therefore \mu = 64 \text{ months}, \ \sigma = 1.33 \text{ months}.$$

Example. The marks obtained by 1000 students in a final examination are found to be approximately normally distributed with mean 70 and standard deviation 5. Estimate the number of students whose marks will be between 60 and 75, both inclusive, given that the area under the normal curve $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ between$z = 0$ and $z = 2$ is 0.4772 and between $z = 0$ and $z = 1$ is 0.3413.

Solution: Let X denote the marks of a student in the examination. Then X is normal variate with mean 70 and standard deviation 5.

$$\therefore Z = \frac{X-70}{5} \text{ is the standard normal variate.}$$

$$\begin{aligned}
\therefore P(60 \le X \le 75) &= P\left(\frac{60-70}{5} \le \frac{X-70}{5} \le \frac{75-70}{5}\right) \\
&= P(-2 \le Z \le 1) \\
&= \Phi(1) - \Phi(-2) \\
&= 0.8413 - 0.0228 = 0.8185, \text{ since the normal curve is symmetrical.}
\end{aligned}$$

Exercise

1. It is known that one in every 10 villagers of a certain village contract leprosy. If 7 people are selected at random from the village, find the probability that 3 of them will have leprosy in future. $Ans : \mathbf{\frac{9^4.35}{10^7}}$.

2. Two fair dice are rolled 100 times. Find the probability of getting at least once a double six. $Ans : \mathbf{1 - \left(\frac{35}{36}\right)^{100}}$.

3. Suppose the probability of a new born baby being a boy is 0.51. In a family of 8 children, calculate the probability that there are 4 or 5 boys. $Ans : \mathbf{0.5003}$.

4. A random variable X follows binomial distribution with mean $\frac{5}{3}$ and $P(X = 2) = P(X = 1)$. Find variance, $P(X \ge 1)$ and $P(X \le 1)$. $Ans : \mathbf{\frac{10}{9}, \frac{211}{243}, \frac{112}{243}}$.

5. A discrete random variable X has the mean 6 and variance 2. Assuming the distribution to be binomial, find the probability that $5 \le X \le 7$. $Ans : \mathbf{\frac{2^6 \times 73}{3^8}}$.

6. If X is Poisson variate such that $P(X = 1) = 0.2$ and $P(X = 2) = 0.2$, find $P(X = 0)$. $Ans : \mathbf{0.1}$.

7. In a certain factory of turning razor blades, there is a small chance of 1 in 500 blades to be defective. The blades are in packet of 10. Use Poisson distribution to calculate the approximate number of packets containing (i) no defective, (ii) one defective, (ii) two defective blades respectively in one consignment of 10,000 packets. Ans : (**i**)**9802**, (**ii**)**196**, (**iii**)**19604**.

Measure of Central Tendency

It is generally seen as that in a distribution, the values of the variable tend to cluster around a central value of the distribution. This tendency of the distribution is called central tendency and the measures devised to consider the tendency are called the measures of central tendency. The average is a single value in the distribution and serves as a representative of the distribution.

Kinds of Central Tendency

There are three measures of central tendency:

1. Mean: (i) Arithmetic mean (ii) Geometric mean (iii) Harmonic mean.
2. Median.
3. Mode.

When we simply say 'mean' or 'average', we generally refer to Arithmetic mean.

Arithmetic mean (A.M.)

1. The Arithmetic mean of a variable is derived by dividing the sum of its values by the number of values. If x denotes the variable under consideration and its values are $x_1, x_2, ..., x_n$, the arithmetic mean of x is denoted by $\bar{x}$ and given by $\bar{x} = \frac{1}{n}\Sigma_{i=1}^{n} x_i$.
2. Weighted A.M. for ungrouped frequency distribution is given by

$$\bar{x} = \frac{1}{N}\Sigma_{i=1}^{n} f_i x_i \text{ where } N = \Sigma_{i=1}^{n} f_i$$

   where f_i is the frequency of the value of x_i $(i = 1, 2, ..., n)$.
3. Weighted A.M. for grouped frequency distribution

$$\begin{pmatrix} x_1 - x_2 & x_2 - x_3 & ..... & x_n - x_{n-1} \\ f_1 & f_2 & ..... & f_n \end{pmatrix}$$

   is given by $\bar{x} = \frac{1}{N}\Sigma_{i=1}^{n} f_i m_i$ where $N = \Sigma_{i=1}^{n} f_i$ and m_i is the midpoint of the class-interval $(x_i - x_{i+1})$.

Some properties of A.M.

1. $\Sigma x_i = n\bar{x}$ and $f_i x_i = N\bar{x}$.
2. $\Sigma(x_i - \bar{x}) = 0$ and $\Sigma f_i(x_i - \bar{x}) = 0$.
3. If $z = ax + b$, where x and y are two variables, then $\bar{z} = a\bar{x} + b$ (a and b are constants).
4. If mean of a series $(x_{11}, x_{12}, x_{13}, ..., x_{1n_1})$ is $\bar{x_1}$ and the mean of a series $(x_{21}, x_{22}, x_{23}, ..., x_{2n_2})$ is $\bar{x_2}$, then the mean of the combined series is given by

$$\bar{a} = \frac{n_1\bar{x_1} + n_2\bar{x_2}}{n_1 + n_2}$$

Calculation of A.M. by Step Deviation Method

Example. Calculate the A.M. from the following data:

Marks:	$0-10$	$10-20$	$20-30$	$30-40$	$40-50$	$50-60$	$60-70$
Students:	5	8	3	10	8	14	2

Solution:

Marks (class) $(x_i - x_{i+1})$	Students f_i	Mid-points (m_i of classes)	$f_i m_i$
0-10	5	5	25
10-20	8	15	120
20-30	3	25	75
30-40	10	35	350
40-50	8	45	360
50-60	14	55	770
60-70	2	65	130
Total	N=50		$\Sigma f_i m_i$=1830

$$\therefore \bar{x} = \frac{1}{N}\Sigma f_i m_i = \frac{1830}{30} = 36.6.$$

Alternative

Marks (class) $(x_i - x_{i+1})$	Students f_i	Mid-points (m_i of classes)	$y_i = \frac{m_i - 35}{10}$	$f_i y_i$
0-10	5	5	-3	25
10-20	8	15	-2	120
20-30	3	25	-1	75
30-40	10	35	0	350
40-50	8	45	1	360
50-60	14	55	2	770
60-70	2	65	3	130
Total	N=50			$\Sigma f_i y_i$=8

$$\therefore \bar{y} = \frac{8}{50} = 0.16.$$

Hence, $\bar{x} = 35 + 10 \times 0.16 = 36.6$.

Example. Find the value of p if the mean of the distribution is 20.

x :	15	17	19	$20+p$	23
t :	2	3	4	$5p$	6

Solution:

x	f	fx
15	2	30
17	3	51
19	4	76
$20+p$	$5p$	$100p + 5p^2$
23	6	138
Total	$N = 5p + 15$	$\Sigma fx = 295 + 100p + 5p^2$

Now, given that mean= 20
$\implies \frac{\Sigma fx}{N} = 20$
$\implies \frac{295+100p+5p^2}{5p+15} = 20$
$\implies 5(p^2 - 1) = 0$. ‘ $\therefore p = \pm 1$.

Example. The following table shows the marks scored by 140 students in an examination of a certain paper:

Marks:	$0-10$	$10-20$	$20-30$	$30-40$	$40-50$
Number of students:	20	24	40	36	20

Calculate the average marks by assumed mean method.

Solution: Let the assumed mean be = 25.

class interval	Mid-value x_i	$d_i = x_i - A$ $= x_i - 25$	$u_i = \frac{x_i - 25}{10}$	Frequency f_i	$f_i u_i$
0-10	5	-20	-2	20	-40
10-20	15	-10	-1	24	-24
20-30	25	0	0	40	0
30-40	35	10	1	36	36
40-50	45	20	2	20	40
Total				$N = 140$	$\Sigma f_i u_i$=12

$$\text{Mean} = A + \frac{\Sigma f_i u_i}{N} \times h$$

$$= 25 + \frac{120}{140} \times 10 = 25 + 0.857$$

$$= 25.857.$$

Example. Find the missing frequencies in the following frequency distribution, when it is known that A.M.=36.6 and it is also known that $\Sigma f_i = 50$:

Marks:	$0-10$	$10-20$	$20-30$	$30-40$	$40-50$	$50-60$	$60-70$
Students:	5	f_2	f_3	10	8	14	2

Solution: Here c is so chosen that either y_2 or y_3 becomes 0.
We choose c so as to make $y_3 = 0$, evidently, $c = 25$. We take $d = 5$.

Marks (class) $(x_i - x_{i+1})$	Students f_i	Mid-points (m_i of classes)	$y_i = \frac{m_i - 35}{10}$	$f_i y_i$
0-10	5	5	-4	-20
10-20	f_2	15	-2	$-2f_2$
20-30	$3f_3$	25	0	0
30-40	10	35	2	20
40-50	8	45	4	32
50-60	14	55	6	84
60-70	2	65	8	16
Total	N=50			$\Sigma f_i y_i = 132 - 2f_2$

We have the formula, $\bar{x} = c + d\bar{y}$ i.e, $\bar{x} = c + d\frac{\Sigma f_i y_i}{N}$

$$\begin{aligned} \text{or,} \quad 36.6 &= 25 + \left(5 \times \frac{132 - 2f_2}{50}\right) \\ \text{or,} \quad \frac{132 - 2f_2}{50} &= 36.6 - 25 = 11.6 \\ \text{or,} \quad 132 - 2f_2 &= 116 \\ \text{or,} \quad f_2 &= 8. \end{aligned}$$

$\therefore \quad f_3 = 50 - (5 + 8 + 10 + 8 + 14 + 2) = 50 - 47 = 3.$

Median

Median is the middle-most value of the variate in a distribution.

1. For discrete variables without repetition

Arrange the data either in ascending or in descending order in their values:

1. If the number of values is odd, $2n + 1$, (say), then, the $(n + 1)$th value is the median. For example te median of the observations 3, 6, 8, 12, 13, 17, 25 is 12.
2. . If the number of values is even, $2n$, say, then A.M. of the nth and $(n + 1)$th values is the median. for example, the median of the observations 30, 25, 24, 20, 17, 14, 9, 7 is $\frac{20+17}{2} = 18.5$.

2. For ungrouped frequency distribution

Take the following steps:

1. Arrange the distribution in either ascending or in descending order.
2. Find $N/2$, where $N = \Sigma f_i$.
3. Find the cumulative frequencies.
4. Find the cumulative frequency just greater that $N/2$.
5. The corresponding value of the variable is the median.

Example. Find the median from the following distribution:

Income (Rs) :	100	150	80	200	250	180
No. of persons :	16	24	26	20	6	30

Solution: Arranging in ascending order and calculating cumulative frequencies:

Income (Rs)	No. of persons	Cumulative frequency (C.F)
80	26	26
100	16	42
150	24	66
180	30	96
200	20	116
250	6	122

Here $N = 122, \quad \therefore N/2 = 61.$
C.F. just greater than 61 is 66. The corresponding income is Rs. 150. Hence the median= Rs. 150.

2. For grouped frequency distribution

For grouped data median formula is:

$$\text{Median } = \ l + \frac{\frac{n}{2} - c}{f} \times h$$

where l= lower limit of median class, n= number of observation, f= frequency of median class, c= Cumulative frequency of preceding class, h= class width.

Example. Calculate the median of the distribution:

Wages (Rs) :	$30-40$	$40-50$	$50-60$	$60-70$	$70-80$
No. of workers :	5	3	20	10	7

Solution:

Wages (Rs)	No. of workers	Cumulative frequency (C.F)
30-40	5	5
40-50	3	8
50-60	20	28
60-70	10	38
70-80	7	45

Here $N = 45$, $N/2$=22.5. Obviously, the class corresponds to C.F. 28 is the median class, i.e. 50-60.
$\therefore l = 50, h = 10, f = 20, c = 8.$
Hence the median $= l + \frac{\frac{N}{2} - c}{f} \times h$

$$= 50 + \tfrac{10}{20}(22.5 - 8)$$

$$= 50 + 7.25 = 57.25.$$

Example. An incomplete distribution is given as follows:

Marks:	$0-10$	$10-20$	$20-30$	$30-40$	$40-50$	$50-60$	$60-70$
Students:	10	20	?	40	?	25	15

You are given that the median value is 35 and the sum of all frequencies is 170. Using median formula, fill up the missing frequencies.

Solution:

Wages (Rs)	No. of workers	Cumulative frequency (C.F)
0-10	10	10
10-20	20	30
20-30	f_1	$30+f_1$
30-40	40	$70+f_1$
40-50	f_2	$70+f_1+f_2$
50-60	25	$95+f_1+f_2$
60-70	15	$110+f_1+f_2$

Given Median $= 35$, then median class $= 30-40$
$\therefore l = 30,\ h = 10,\ f = 40,\ F = 30+f_1.$

$\therefore$ Median $= l + \frac{\frac{N}{2}-c}{f} \times h$
$\implies 35 = 30 + \frac{85-(30+f_1)}{40} \times 10$
$\implies f_1 = 35.$

Again, it is given sum of the frequencies $= 170$
$\therefore f_2 = 170-10-20-35-40-25-15 = 25.$

Mode

Mode of a distribution is the value of the variable having maximum frequency.

Mode in an ungrouped frequency distribution

Example. Calculate the mode in the following distribution:

x :	1	2	3	4	5	6
f :	9	13	28	21	8	3

Solution: Since 28 is the maximum frequency, therefore Mode=3.

Grouped frequency distribution

For a grouped frequency distribution, if f_0, f_{-1} and f_1 represent the frequencies of modal class, the class just preceding and the class just following it, then

$$\text{Mode} \;=\; l_1 + \frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1} \times c$$

where l_1= lower boundary of the modal class, c= common width of the classes.

Example. The monthly profits in rupees of 100 shops are distributed as follows:

Profits per shop :	$0-100$	$100-200$	$200-300$	$300-400$	$400-500$	$500-600$
No. of shops :	12	18	27	20	17	6

Solution: We see that the largest class frequency is 27, lies in the class 200-300 and hence, this is the modal class. Therefore,

$$l_1 = 200,\ f_0 = 27,\ f_{-1} = 18,\ f_1 = 20,\ c = 100.$$

$$\therefore \quad \text{Mode} = l_1 + \tfrac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1} \times c$$

$$= 200 + \tfrac{9}{9+7} \times 100 = 200 + \tfrac{9}{16} \times 100 = 256.25.$$

Empirical relation between mean, median and mode

For unimodal moderately skewed distribution, the following approximate relation has been found to hold:

$$\text{Mean- Mode} \;=\; 3\ (\text{Mean-Median})$$

Sometimes this relation may be used for the calculation of mode. When distribution is symmetrical, mean, median and mode coincide. For the normal distribution mean and median are equal.

In most frequency distribution, it has been observed that the three measures of central tendency, viz., mean, median and mode, obey the approximate relation provided the distribution is not very skew.

Therefore, this relation is applied to estimate one of them when the values of the other two are known.

Example. Find the mode of the following distribution:

Marks obtained :	$0-10$	$10-20$	$20-30$	$30-40$	$40-50$
No. of students :	2	4	9	7	3

Solution: For calculation of the mode, we construct the following table:

Marks (class) $(x_i - x_{i+1})$	Students f_i	Mid-points (m_i of classes)	$f_i m_i$
0-10	2	5	10
10-20	4	15	60
20-30	9	25	225
30-40	7	35	245
40-50	3	45	135
Total	N=25		$\Sigma f_i m_i$=675

We see that the largest class frequency is 9, lies in the class 20-30 and hence, this is the modal class. Therefore,

$$l_1 = 20,\ f_0 = 9,\ f_{-1} = 4,\ f_1 = 7,\ c = 10.$$

$$\therefore \quad \text{Mode} = l_1 + \frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1} \times c$$

$$= 20 + \frac{5}{5+2} \times 10 = 20 + \frac{50}{7} = 27.14.$$

Example: Find the mean, median and mode of the following data:

Marks obtained :	$0-20$	$20-40$	$40-60$	$60-80$	$80-100$	$100-120$	$120-140$
No. of students :	6	8	10	12	6	5	3

Solution:

Class interval $(x_i - x_{i+1})$	Mid value x	Frequency f	fx	Cumulative frequency
0-20	10	6	60	6
20-40	30	8	240	14
40-60	50	10	500	24
60-80	70	12	840	36
80-100	90	6	540	42
100-120	110	5	550	47
120-140	130	3	390	50
Total		N=50		$\Sigma fx = 3120$

$\therefore$ Mean= $\frac{\Sigma fx}{N} = \frac{3120}{50} = 62.4$.

We have, $N = 50$.
Then, $\frac{N}{2} = 25$.

The cumulative frequency is just greater than $\frac{N}{2}$ is 36, then the median class is 60-80 such that

$$l = 60,\ h = 20,\ f = 12,\ c = 24.$$

$$\therefore \quad \text{Median} = l + \frac{\frac{N}{2} - c}{f} \times h$$
$$= 60 + \frac{25 - 24}{12} \times 20$$
$$= 60 + \frac{20}{12}$$
$$= 61.67.$$

And, $\text{Mode} = l_1 + \frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1} \times c$

$= 60 + \frac{12-10}{2\times12-10-6} \times 20 = 60 + \frac{2\times20}{8} = 65.$

Measures of dispersion

The dispersion is the measure of variation in the values of the variable. It measures the degree of scatteredness of the observations in a distribution around the central value.

Following are commonly used for measures of dispersion:
(i) Range, (ii) Quartile deviation, (iii) Mean deviation, (iv) Standard deviation.

Range : The range is the difference between two extreme observations of the distribution. If A and B are the greatest and smallest values respectively of the observations in a distribution, then its range is $A - B$.
Thus,

Range of a distribution = Maximum value − Minimum value.

For example, consider

Match :	1	2	3	4	5	6	7	8	9
Batsman 1 :	30	91	0	64	42	80	30	5	117
Batsman 2 :	53	46	48	50	53	53	58	60	57

Range of scores of batsman 1= 117-0=117 and Range of scores of batsman 2=60-46=14.

Range is the simplest but crude measurement of dispersion. As it is based on two extreme observations so it does not measures the dispersion of a data from its central value.

Quartile deviation : A median divides a given dataset (which is already sorted) into two equal halves similarly, the quartiles are used to divide a given dataset into four equal halves. Therefore, logically there should be three quartiles for a given distribution, but if one observes it, the second quartile is equal to the median itself. The **first quartile** or the **lower quartile** or the 25th percentile, also denoted by Q_1, corresponds to the value that lies halfway between the median and the lowest value in the distribution (when it is already sorted in the ascending order). Hence, it marks the region which encloses 25 % of the initial data. Similarly, the **third quartile** or

the **upper quartile** or 75th percentile, also denoted by Q3, corresponds to the value that lies halfway between the median and the highest value in the distribution (when it is already sorted in the ascending order). It, therefore, marks the region which encloses the 75 % of the initial data or 25 % of the end data.

The difference $Q_3 - Q_1$ is called the inter quartile range and the quartile deviation (Q.D) is defined by

$$\text{Q.D} \;=\; \frac{Q_3 - Q_1}{2}.$$

A relative measure of dispersion based on the quartile deviation is known as the coefficient of quartile deviation. It is characterized as

$$\textit{Coefficient of Quartile Deviation} \;=\; \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100.$$

For grouped data:

For the case of a grouped-data distribution, we can find the quartiles through the following steps-

1. Construct a cumulative frequency table for the given data alongside the given distribution.

2. From the total number of data values, estimate the groups/classes of the Lower and Upper Quartiles

Use the following formulae to then calculate the quartiles:

$$\text{The lower quartile } Q_1 \;=\; LB + \frac{\frac{n}{4} - f_c}{f} \times h$$

$$\text{The upper quartile } Q_3 \;=\; LB + \frac{\frac{3n}{4} - f_c}{f} \times h$$

where, LB – the lower bound of the class in which the respective quartile lies.
h – the class width.
f_c – the cumulative frequency up to that class.
f – the frequency corresponding to that particular class.

Example : The number of vehicles sold by a major Toyota Showroom in a day was recorded for 10 working days. The data is given as –

Day	Frequency
1	20
2	15
3	18
4	5
5	10
6	17
7	21
8	19
9	25
10	28

Find the Quartile Deviation and its coefficient for the given discrete distribution case.

Solution : We first need to sort the frequency data given to us before proceeding with the quartiles calculation –

Sorted Data – 5, 10, 15, 17, 18, 19, 20, 21, 25, 28 and $n = 10$.

Now, to find the quartiles, we use the logic that the first quartile lies halfway between the lowest value and the median; and the third quartile lies halfway between the median and the largest value.

$$\begin{aligned}\text{First Quartile } Q_1 &= \tfrac{n+1}{4} \text{ th term} \\ &= \tfrac{10+1}{4} \text{ th term } = 2.75 \text{ th term} \\ &= 2 \text{ nd term} + 0.75 \times (3\text{rd term} - 2\text{nd term}) \\ &= 10 + 0.75 \times (15 - 10) = 13.75.\end{aligned}$$

$$\begin{aligned}\text{Third Quartile } Q_3 &= \tfrac{3(n+1)}{4} \text{ th term} \\ &= \tfrac{3(10+1)}{4} \text{ th term } = 8.25 \text{ th term} \\ &= 8 \text{ th term} + .25 \times (9\text{th term} - 8\text{th term}) \\ &= 21 + 0.25 \times (25 - 21) = 22.\end{aligned}$$

Using the values for Q_1 and Q_3, now we can calculate the Quartile Deviation and its coefficient as follows –

$$\begin{aligned}\text{Quartile deviation} &= \text{Semi-Inter quartile range} \\ &= \tfrac{Q_3 - Q_1}{2} \\ &= \tfrac{22 - 13.75}{2} \\ &= 4.125.\end{aligned}$$

$$\begin{aligned}\text{Coefficient of Quartile deviation} &= \tfrac{Q_3 - Q_1}{Q_3 + Q_1} \times 100 \\ &= \tfrac{22 - 13.75}{22 + 13.75} \times 100 \\ &= \tfrac{8.25}{35.75} \times 100 \\ &= 23.08.\end{aligned}$$

Example: For the following data, calculate the Quartile Deviation and its coefficient.

Marks	Number of students
$0-10$	10
$10-20$	20
$20-30$	30
$30-40$	50
$40-50$	40
$50-60$	30

Solution: For the given data, we can form the required table with the cumulative frequency as –

Solution:

Marks	Frequency	Cumulative Frequency
0-10	10	10
10-20	20	30
20-30	30	60
30-40	50	110
40-50	40	150
50-60	30	180

Since the total number of students is 180, the first quartile must lie at the position of $180/4 = 45$th student. Similarly, the third quartile must lie at the position of $180 \times 3/4 = 135$th student. By the distribution of our data into groups, we can note that the first quartile will lie in the $20-30$ marks range.

Here, $LB = 20; h = 10 f_c = 30; f = 30; n = 180.$

$$\text{The lower quartile } Q_1 \;=\; 20 + \frac{\frac{180}{4} - 30}{30} \times 10 \;=\; 25.$$

Similarly, the third quartile will lie in the 40-50 marks range. Hence,
Here, $LB = 40; h = 10 f_c = 110; f = 40; n = 180.$

$$\text{The upper quartile } Q_3 \;=\; 40 + \frac{\frac{3}{4} \times 180 - 110}{40} \times 10 \;=\; 46.25.$$

Now, using the values for Q1 and Q3, now we can calculate the Quartile Deviation and its coefficient as follows –

$$\begin{aligned} \text{Quartile deviation} &= \text{Semi-Inter quartile range} \\ &= \frac{Q_3 - Q_1}{2} \\ &= \frac{46.25 - 25}{2} \end{aligned}$$

$$= 10.625.$$

$$\begin{aligned}\text{Coefficient of Quartile deviation} &= \frac{Q_3-Q_1}{Q_3+Q_1} \times 100 \\ &= \frac{46.25-25}{46.25+25} \times 100 \\ &= \frac{21.25}{71.25} \times 100 \\ &= 29.82.\end{aligned}$$

Mean Deviation

In this section, we will learn how to calculate mean deviation about mean and median for various types of data.

Mean deviation for ungrouped data

$x_1, x_2, ..., x_n$ be n values of a variable X , then the mean deviation from an average A (median or mean) is given by

$$\text{Mean deviation} = \frac{1}{n} \Sigma_{i=1}^{n} |x_i - A|.$$

Example. Calculate the mean deviation about median from the following data : 340, 150, 210, 240, 300, 310, 320.

Solution: Arranging the observations in ascending order of magnitude, we have 150, 210, 240, 300, 310, 320, 340. Clearly, the middle observation is 300. So, median is 300.

x_i	$d_i = \lvert x_i - A \rvert$
340	40
150	150
210	90
240	60
300	0
310	10
320	20
Total	$\Sigma d_i = \Sigma \lvert x_i - 300 \rvert = 370$

$$\therefore \text{Mean deviation} = \frac{1}{n} \Sigma_{i=1}^{n} |x_i - A| = \frac{370}{7} = 52.8.$$

Example. Calculate the mean deviation about the mean of the set of first n natural numbers when n is odd natural number.

Solution: Since n is odd natural number, we consider $n = 2m + 1$, where m is some natural number. Let $\bar{X}$ be the mean of first n natural numbers. Then

$$\bar{X} = \frac{1 + 2 + ... + (n-1) + n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}.$$

$\therefore \bar{X} = \frac{2m+1+1}{2} = m+1.$

The mean deviation (M.D.) about the mean is given by

$$\begin{aligned}
\text{M.D.} &= \frac{1}{n}\Sigma_{r=1}^{n}|r-\bar{X}| \\
&= \frac{1}{2m+1}\Sigma_{r=1}^{2m+1}|r-(m+1)| \\
&= \frac{1}{2m+1}\left\{\Sigma_{r=1}^{m}|r-(m+1)| + \Sigma_{r=m+1}^{2m+1}|r-(m+1)|\right\} \\
&= \frac{1}{2m+1}\left\{\Sigma_{r=1}^{m}(m+1-r) + \Sigma_{r=m+1}^{2m+1}(r-(m+1))\right\} \\
&= \frac{1}{2m+1}\left\{-\frac{m(m+1)}{2} + m(m+1) + \frac{1}{2}(m+1)(3m+2) - (m+1)^2\right\} \\
&= \frac{m(m+1)}{2m+1} = \frac{\left(\frac{n-1}{2}\right)\left(\frac{n-1}{2}+1\right)}{n} = \frac{n^2-1}{4n}.
\end{aligned}$$

Example. Calculate the mean deviation about mean of the following data :

x_i :	3	9	17	23	27
f_i :	8	10	12	9	5

Solution: We first calculate the mean deviation about mean:

x_i	f_i	f_ix_i	$\lvert x_i-15\rvert$	$f_i\lvert x_i-15\rvert$
3	8	24	12	96
9	10	90	6	60
17	12	204	2	24
23	9	207	8	72
27	5	135	12	60
Total	$N=\Sigma f_i = 44$	$\Sigma f_ix_i = 660$		$\Sigma f_i\lvert x_i-15\rvert = 312$

$\therefore$, Mean $= \bar{X} = \frac{1}{N}\Sigma f_ix_i = \frac{660}{44} = 15$

$\therefore$ Mean deviation = M.D.$= \frac{1}{N}\Sigma f_i|x_i-15| = \frac{312}{44} = 7.09.$

Example. Calculate the mean deviation from the median of the following data:

Wages per week :	$10-20$	$20-30$	$30-40$	$40-50$	$50-60$	$60-70$	$70-80$
No. of workers :	4	6	10	20	10	6	4

Solution:

Wages per week	Mid-value x_i	Frequency f_i	Cumulative Frequency	$\|d_i\| = \|x_i - 45\|$	$f_i\|d_i\|$
10-20	15	4	4	30	120
20-30	25	6	10	20	120
30-40	35	10	20	10	100
40-50	45	20	40	0	0
50-60	55	10	50	10	100
60-70	65	6	56	20	120
70-80	75	4	60	30	120
Total		$N = \Sigma f_i = 60$			$\Sigma f_i\|d_i\|$=680

Here $N = 60$, so $\frac{N}{2} = 30$. The cumulative frequency just greater than $\frac{N}{2} = 30$ is 40 and the corresponding class is 40-50. So 40-50 is the median class.

$$\therefore l = 40, f = 20, h = 10, c = 20.$$

So, Median $= l + \frac{\frac{N}{2} - c}{f} \times h = 40 + \frac{30-20}{20} \times 10 = 45.$

Thus, we have

$$\Sigma f_i|x_i - 45| = \Sigma f_i|d_i| = 680 \text{ and } N = 60.$$

$$\therefore \text{Mean deviation from median } = \frac{\Sigma f_i|d_i|}{N} = \frac{680}{60} = 11.33.$$

Example. Find the mean deviation about the mean for the following data

Marks obtained :	$10-20$	$20-30$	$30-40$	$40-50$	$50-60$	$60-70$	$70-80$
No. of students :	2	3	8	14	8	3	2

Solution: In order to avoid the tedious calculation of computing mean $(\bar{X})$, we compute $\bar{X}$ by step-deviation method. The formula for step-deviation method is given by

$$\bar{X} = a + h\left(\frac{1}{N}\Sigma_{i=1}^{n} f_i d_i\right)$$

where $d_i = \frac{x_i - a}{h}$, a = assumed mean and h = common factor.

We consider the assumed mean $a = 45$ and $h = 10$ for the following table.

Marks obtained	Number of students f_i	Mid-points x_i	$d_i = \frac{x_i - 45}{10}$	$f_i d_i$	$\lvert x_i - 45 \rvert$	$f_i \lvert x_i - 45 \rvert$
10-20	2	15	-3	-6	30	60
20-30	3	25	-2	-6	20	60
30-40	8	35	-1	-8	10	80
40-50	14	45	0	0	0	0
50-60	8	55	1	8	10	80
60-70	3	65	2	6	20	60
70-80	2	75	3	6	30	60
	$N = 40$			$\Sigma f_i d_i = 0$		$\Sigma f_i \lvert x_i - 45 \rvert = 400$

Clearly, $N = 40$, $\Sigma f_i d_i = 0$.

$\therefore \bar{X} = a + h\left(\frac{1}{N}\Sigma_{i=1}^{n} f_i d_i\right) = 45 + 10 \times \frac{0}{40} = 45.$

$\therefore$ M.D. $= \frac{1}{N}\Sigma f_i |x_i - 45| = \frac{400}{40} = 10.$

Limitations of mean deviation

Following are the limitation of the mean deviation.

1. In frequency distribution, the sum of the absolute values of the deviations from the mean is always more than the sum of the deviations from median. Therefore, the mean deviation about mean is not very scientific. Thus, in many cases, mean deviation may give unsatisfactory results.

2. In a distribution, where the degree of variability is high, the median is not a representative central value. Thus, the mean deviation about the median calculated for such series cannot be fully relied.

3. In the computation of mean deviation we use absolute values of deviations. Therefore, it cannot be subjected to further algebraic treatment.

Variance and Standard Deviation

Variance The variance of a variate X is the arithmetic mean of the squares of all deviations of X from the arithmetic mean of the observations and is denoted by $\mathrm{Var}(X)$ or σ^2.
The positive square root of the variance of a variate X is known as its standard deviation and is denoted by σ. Thus, Standard deviation$= +\sqrt{\mathrm{Var}(X)}$.

Variance of ungrouped observation

If $x_1, x_2, ..., x_n$ are n values of a variable X , then

$$Var(X) = \frac{1}{n}\left\{\Sigma_{i=1}^{n}(x_i - \bar{X})^2\right\}.$$

After some simplifications of this summation formula, one can get

$$\text{Var } (X) \;=\; \frac{1}{n}\Sigma_{i=1}^{n}x_i^2 - \left\{\frac{1}{n}\Sigma_{i=1}^{n}x_i\right\}^2.$$

Example. Find the variance and standard deviation for the following data:

$$65, 68, 58, 44, 48, 45, 60, 62, 60, 50.$$

Solution: Let $\bar{X}$ be the mean of the given set of observations. Then

$$\bar{X} = \frac{65 + 68 + 58 + 44 + 48 + 45 + 60 + 62 + 60 + 50}{10} = \frac{560}{10} = 56.$$

x_i	$x_i - \bar{X} = x_i - 56$	$(x_i - \bar{X})^2$
65	9	81
68	2	4
58	12	144
44	-12	144
48	-8	64
45	-11	121
60	4	16
62	6	36
60	4	16
50	-6	36
		$\Sigma(x_i - \bar{X})^2 = 662$

Here $n = 10$, and $\Sigma(x_i - \bar{X})^2 = 662$.

Variance $= \frac{1}{n}\Sigma(x_i - \bar{X})^2 = \frac{662}{10} = 66.2$.

Hence, Standard deviation $(\sigma) = \sqrt{\text{Variance}} = \sqrt{66.2} = 8.13$.

Properties of variance

1. Let $x_1, x_2, x_3, ..., x_n$ be n values of the variable X. If these values are changed to $x_1 + a, x_2 + a, ...x_n + a$, where $a \in R$, then the variance remains unchanged.

2. Let $x_1, x_2, x_3, ..., x_n$ be n values of the variable X and a be a non-zero real number. Then the variance of the observations $ax_1, ax_2, ax_3, ..., ax_n$ is a^2 Var (X).

Example. If for a distribution of 18 observations, $\Sigma(x_i - 5) = 3$ and $\Sigma(x_i - 5)^2 = 43$, find the mean and standard deviation.

Solution: We have

$$\Sigma_{i=1}^{18}(x_i - 5) = 3 \text{ and } \Sigma_{i=1}^{18}(x_i - 5)^2 = 43.$$

$$\implies \Sigma_{i=1}^{18}x_i - \Sigma_{i=1}^{18}5 = 3 \text{ and } \Sigma_{i=1}^{18}x_i^2 - 10\Sigma_{i=1}^{18}x_i + \Sigma_{i=1}^{18}25 = 43$$

$$\implies \Sigma_{i=1}^{18}x_i = 93 \text{ and } \Sigma_{i=1}^{18}x_i^2 = 523.$$

$$\therefore \text{Mean } = \tfrac{1}{18}\Sigma_{i=1}^{18}x_i = \tfrac{93}{18} = 5.17.$$

$$\therefore \text{S.D.} = \sqrt{\tfrac{1}{18}\Sigma_{i=1}^{18}x_i^2 - \left(\tfrac{1}{18}\Sigma_{i=1}^{18}x_i\right)^2} = \tfrac{\sqrt{765}}{18} = 1.536.$$

Example. For a group of 200 candidates the mean and S.D were found to be 40 and 15 respectively. Later it was found that the score 43 was misread as 34. Find the correct mean and correct Standard deviation (S.D.).

Solution: We have $n = 200$, $\bar{X} = 40$ and $\sigma = 15$.

$$\therefore \bar{X} = \frac{1}{n}\Sigma x_i \implies \Sigma x_i = n\bar{X} = 8000.$$

Now, Corrected Σx_i = Incorrect $\Sigma x_i -$ (Sum of incorrect values)+(Sum of correct values) $= 8000 - 34 + 43 = 8009$.

$$\therefore \text{Corrected mean } = \frac{\text{Corrected } \Sigma x_i}{n} = \frac{8009}{200} = 40.045.$$

and $\sigma = 15$
$\implies$ Variance $= 15^2$
$\implies 15^2 = \frac{1}{200}(\Sigma x_i^2) - \left(\frac{1}{200}\Sigma x_i\right)^2$
$\implies 225 = \frac{1}{200}(\Sigma x_i^2) - 1600$
$\implies$ incorrect $\Sigma x_i^2 = 365000$.
$\therefore$ Corrected Σx_i^2 = Incorrect $\Sigma x_i^2 -$ (Sum of squares of incorrect values)
+(Sum of squares of correct values)
$\implies$ corrected $\Sigma x_i^2 = 365693$.
$\therefore$ corrected $\sigma = \sqrt{\frac{1}{n} \text{ corrected } \Sigma x_i^2 - \left(\frac{1}{n} \text{ corrected } \Sigma x_i\right)^2}$
$= \sqrt{\frac{365693}{200} - \left(\frac{8009}{200}\right)^2} = 14.995.$

Example. Find the mean and standard deviation of first n terms of an arithmetic progression (A.P.) whose first term is a and common difference is d.

Solution: The terms of the A.P. are: $a, a+d, a+2d, a+3d, ..., a+(r-1)d, ..., a+(n-1)d$. Let $\bar{X}$ be the mean of the terms. Then

$$\bar{X} = \frac{1}{n}\{a + (a+d) + (a+2d) + ... + (a+(n-1)d)\} = \frac{1}{n}\left[\frac{n}{2}\{2a + (n-1)d\}\right]$$

$$= a + (n-1)\tfrac{d}{2}.$$

Let σ be the standard deviation of n terms of A.P. then,

$$\begin{aligned}
\sigma^2 &= \frac{1}{n}\Sigma_{r=1}^{n}\left[\{a+(r-1)d\} - \bar{X}\right]^2 \\
&= \frac{1}{n}\Sigma_{r=1}^{n}\left[\{a+(r-1)d\} - \{a+(n-1)d\}\right]^2 \\
&= \frac{d^2}{4n}\Sigma_{r=1}^{n}\left[2r-(n+1)\right]^2 \\
&= \frac{d^2}{4n}\Sigma_{r=1}^{n}\left[4r^2 - 4(n+1)r + (n+1)^2\right] \\
&= \frac{d^2}{4n}\left[4\Sigma_{r=1}^{n}r^2 - 4(n+1)\Sigma_{r=1}^{n}r + \Sigma_{r=1}^{n}(n+1)^2\right] \\
&= \frac{d^2}{4n}\left\{\frac{4n(n+1)(2n+1)}{6} - \frac{4(n+1)n(n+1)}{2} + n(n+1)^2\right\} \\
&= \frac{d^2}{4n}n(n+1)\{2(2n+1) - 3(n+1)\} = \frac{(n^2-1)d^2}{12} \\
&= \sigma = d\sqrt{\frac{n^2-1}{12}}.
\end{aligned}$$

Example. Find the variance and standard deviation of the following frequency distribution.

Variable x_i :	2	4	6	8	10	12	14	16
Frequency f_i :	4	4	5	15	8	5	4	5

Solution:

Variable x_i	Frequency f_i	f_ix_i	$x_i - \bar{X} = x_i - 9$	$(x_i - \bar{X})^2$	$f_i(x_i - \bar{X})^2$
2	4	8	-7	49	196
4	4	16	-5	25	100
6	5	30	-3	9	45
8	15	120	-1	1	15
10	8	80	1	1	8
12	5	60	3	9	45
14	4	56	5	25	100
16	5	80	7	49	245
Total	$N = \Sigma f_i = 50$	$\Sigma f_ix_i = 450$			$\Sigma f_i(x_i - \bar{X})^2$=754

Here, $N = 50, \Sigma f_ix_i = 450$ and $\Sigma f_i(x_i - \bar{X})^2 = 754$.

$\therefore \bar{X} = \frac{1}{N}\Sigma f_ix_i = \frac{450}{50} = 9$ and Variance $(X) = \frac{1}{N}\Sigma f_i(x_i - \bar{X})^2 = \frac{754}{50} = 15.08$.

Hence S.D.= $\sqrt{\text{Var}(X)} = \sqrt{15.08} = 3.88$.

Example. Calculate the variance and standard deviation of the following frequency distribution from the data below using assumed mean.

Size of item x_i :	3.5	4.5	5.5	6.5	7.5	8.5	9.5
Frequency f_i :	3	7	22	60	85	32	8

Solution: Let the assumed mean be $A = 6.5$.

Size of item x_i	Frequency f_i	$d_i = x_i - 9$	d_i^2	f_id_i	$f_id_i^2$
3.5	3	-3	9	-9	27
4.5	7	-2	4	-14	28
5.5	22	-1	1	-22	22
6.5	60	0	0	0	0
7.5	85	1	1	85	85
8.5	32	2	4	64	128
9.5	8	3	9	24	72‘
Total	$N = \Sigma f_i = 217$			$\Sigma f_id_i = 128$	$\Sigma f_id_i^2$=362

Here $N = 217$, $\Sigma f_i d_i = 128$, and $\Sigma f_i d_i^2 = 362$.

$$\therefore \text{ Var } (X) = \frac{1}{N}\Sigma f_i d_i^2 - \left(\frac{1}{N}\Sigma f_i d_i\right)^2 = \frac{362}{217} - \left(\frac{128}{217}\right)^2 = 1.321.$$

Hence S.D. $= \sqrt{\text{Var}(X)} = \sqrt{1.321} = 1.149$.

Example. Calculate the variance and standard deviation for the following distribution.

Marks :	$20-30$	$30-40$	$40-50$	$50-60$	$60-70$	$70-80$	$80-90$
No. of students :	3	6	13	15	14	5	4

Marks obtained	Number of students f_i	Mid-points x_i	$u_i = \frac{x_i - 45}{10}$	$f_i u_i$	u_i^2	$f_i u_i^2$
20-30	3	25	-3	-9	9	27
30-40	6	35	-2	-12	4	24
40-50	13	45	-1	-13	1	13
50-60	15	55	1	0	0	0
60-70	14	65	1	14	1	14
70-80	5	75	2	10	9	20
80-90	4	85	3	12	4	36
	$N = \Sigma f_i = 60$			$\Sigma f_i u_i = 2$		$\Sigma f_i u_i^2 = 134$

Here $N = 60$, $\Sigma f_i u_i = 2$, $\Sigma f_i u_i^2 = 134$, and $h = 10$.

$$\therefore \text{ Mean } = \bar{X} = A + h\left(\frac{1}{N}\Sigma f_i u_i\right) = 55 + 10\left(\frac{2}{60}\right) = 55.333$$

and

$$\text{Var } (X) = h^2\left\{\frac{1}{N}\Sigma f_i u_i^2 - \left(\frac{1}{N}\Sigma f_i u_i\right)^2\right\} = 100\left[\frac{134}{60} - \left(\frac{2}{60}\right)^2\right] = 222.9.$$

Hence S.D. $= \sqrt{\text{Var}(X)} = \sqrt{222.9} = 14.94$.

Exercise

1. The A.M of the following distribution is 67.45. Find the missing frequency.

Height :	$60-62$	$63-65$	$66-68$	$69-71$	$72-74$
No. of students :	15	54	126	$--$	24

Ans : **81**.

2. Find the median of the following distribution:

Class interval	Frequency
130-134	5
135-139	15
140-144	28
145-149	24
150-154	17
155-159	10
160-164	1

Ans : **144.92**.

3. Find the mode of the following distribution:

Marks :	$50-59$	$60-69$	$70-79$	$80-89$	$90-99$
No. of students :	6	14	16	13	3

Ans : **73.50**.

4. Find the missing frequencies of the following distribution:

x :	$10-20$	$20-30$	$30-40$	$40-50$	$50-60$	$60-70$	$70-80$
Frequency :	12	30	f_3	65	f_5	25	18

Ans : $\mathbf{f_3 = 33.5, f_5 = 45}$.

5. Calculate the mean deviation about the mean of the set of first n natural numbers when n is even natural number. Ans : $\mathbf{M.D.} = \frac{\mathbf{n}}{\mathbf{4}}$.

6. Find the mean deviation from mean for the following data:

Classes :	$95-105$	$105-115$	$115-125$	$125-135$	$135-145$	$145-155$
Frequencies :	9	13	16	26	30	12

Ans : **12.005**.

7. Calculate the mean deviation about the median age for the age distribution of 100 persons given below:

Classes :	$16-20$	$21-25$	$26-30$	$31-35$	$36-40$	$41-45$	$46-50$	$51-55$
Frequencies :	5	6	12	14	26	12	16	9

Ans : $\mathbf{9.44, 9.56}$.

8. The mean and standard deviation of 20 observations are found to be 10 and 2, respectively. On checking it was found that an observation 8 was incorrect. Calculate the correct mean and standard deviation in each of the following cases:
   (i) If wrong item is omitted, (ii) if it is replaced by 12. $Ans : \mathbf{1.997}, \mathbf{1.98}$.

9. Calculate the mean and standard deviation for the following table of the age distribution of a group of people:

Age :	$20-30$	$30-40$	$40-50$	$50-60$	$60-70$	$70-80$	$80-90$
No. of persons :	3	51	122	141	130	51	2

$Ans : \mathbf{55.1}, \mathbf{11.8739}$.

Bivariate Analysis

When two quantities are related to each other in any manner, their relationship can often be expressed mathematically in the form of a function though not always. If the relation between them is expressible in the form $y = a + bx$ or $x = a + by$, such a relation is called linear. Similarly, if the relation between them is expressible as $y = a + bx + cx^2$ or $x = a + by + cy^2$, then the relation is called quadratic relation. The relation may even be something else, like an intricate algebraic relation. The relation may be strong or weak depending on what kind of change in one variable induces what kind of change in other. The relation may be positive or negative or even zero. When an increase in the values of one variable induces an increase in the other variable, the relation is said to be positive but if an increase in one induces a decrease in the other, the relation is called negative. If changes in one variable does not have any impact on the other variable, the relation is referred to as a zero relation.

Our discussion here is confined to only linear relation between two variables. Linear relations may be strong or weak, positive or negative. If the values of two variables are known, the set of such values is referred as *bivariate data* and their corresponding graphical representation as a set of points is called the *scatter diagram* or *dot diagram*.

From any bivariate data we can find:

1. whether or not there is any linear relationship between variables, and if there is any, whether the relation is strong or weak, and positive or negative.

2. an approximate mathematical relation (linear) between them so that it is possible to estimate one variable for a given value of the other variable.

We now derive a statistical measure, called *coefficient of correlation*, by which we can decide whether there is a (linear) relation between the variables and also whether the relation is strong or weak, and positive or negative. Thereafter, we derive the best