Probability Class Notes

Department of Mathematics
Jadavpur University
Kolkata

1 Introduction

Random Experiment (r.e.): An experiment E is called a random experiment if

(i) all possible outcomes are known in advance,
(ii) it is impossible to predict which outcome will occur at a particular performance of E.
(iii) E can be repeated, at least conceptually, under identical conditions infinite number of times.

Events: Outcomes of a random experiment are called events. Events which cannot be decomposed are called simple events and events which can be decomposed into simple events are called composite events.

Mutually Exclusive Events: If the two events be such that they cannot occur simultaneously then they are said to be mutually exclusive events.

Event Space: The set of all possible outcomes of a given random experiment is called the event space of E and is denoted by S.

Class: The set of sets is called a class.

Borel Field/$\sigma-$Field: A class Δ of a given set S satisfying the following three axioms is called a Borel Field or a $\sigma-$Field or a $\sigma-$Algebra:

(i) $S \in \Delta$,

(ii) if $A \in \Delta$ then $A^c \in \Delta$,

(iii) if $A_1, A_2, ..., A_n, ... \in \Delta$ then $\sum_{n=1}^{\infty} \in \Delta$.

Any member of Δ will be called an event of the given random experiment E.

Classical Definition of Probability: Let E be a random experiment such that its event space S contains n event points which are equally likely. If an event A connected with E contains m event points then

$$P(A) = \frac{m}{n}.$$

Drawbacks: The definition cannot be applied
(i) if the simple events are not equally likely,
(ii) if the event space contains infinite number of event points.

Statistical Regularity and Frequency Definition of Probability:

Let a random experiment E be repeated N times under identical conditions, in which an event A of E occurs $N(A)$ times. Then the ratio $\frac{N(A)}{N}$ is called the frequency ratio of A and is denoted by $f(A)$. Now if the random experiment E is repeated large number of times, it is seen that the frequency ratio $f(A)$ gradually stabilizes and tends to a constant number. This tendency of stability is called statistical regularity.

On the basis of statistical regularity, we can assume that $\lim_{N\to\infty} \frac{N(A)}{N}$ exists finitely and the value of this limit is called the probability of the event A and is denoted by $P(A)$. Thus,

$$P(A) = \lim_{N\to\infty} f(A) = \lim_{N\to\infty} \frac{N(A)}{N}.$$

Limitations: From observation we obtain the frequency ratio $\frac{N(A)}{N}$, whereas $\lim_{N\to\infty} \frac{N(A)}{N}$ is a rigorous analytical concept. This combination of empirical and analytical concepts leads to mathematical deficiencies.

Axiomatic Definition of Probability: Let S be the event space connected with the random experiment E. Also let Δ be the class of subsets of S forming the class of events of E. A mapping $P : \Delta \to R$ is called a probability function defined on Δ and the unique real number $P(A)$ determined by P is called the probability of the event $A, A \in \Delta$, if the following assumptions are satisfied:

(i) $P(A) \geq 0$,

(ii) $P(S) = 1$,

(iii) if $A_1, A_2, ..., A_n, ...$ be countably infinite number of pairwise mutually exclusive events (i.e., $A_i \bigcap A_j = \emptyset, i \neq j, A_i, Aj \in \Delta$) then

$$P(A_1 + A_2 + ... + A_n + ...) = P(A_1) + P(A_2) + ... + P(A_n) + ...$$

Deductions of some results from the axioms:

(i) Probability of an impossible event is zero $P(\emptyset) = 0$.
(ii) If A^c is the complimentary event of an event A then $P(A^c) = 1 - P(A)$.
(iii) For any event A, $0 \leq P(A) \leq 1$.
(iv) If $A \subset B$ (i.e. occurrence of B implies occurrence of A) then $P(A) \subset P(B)$.
(v) For any two events A and B, $P(A + B) = P(A) + P(B) - P(A \cap B)$.

Proof: For any event A, we have $A \cap B = \emptyset$. It implies that A and ϕ are mutually exclusive. Again,

$$\begin{aligned} A \cup \emptyset &= A, \\ P(A \cup \emptyset) &= P(A), \\ P(A) + P(\emptyset) &= P(A) \\ P(\emptyset) &= 0. \end{aligned} \tag{1.1}$$

This proves the first case.

We have,

$$\begin{aligned} A \cup A^c &= S, \text{ where S is the sure event} \\ P(A \cup A^c) &= P(S) = 1, \text{ by axiom (ii)} \\ P(A) + P(A^c) &= 1, \text{ by axiom (iii)} \\ P(A^c) &= 1 - P(A). \end{aligned} \tag{1.2}$$

This proves the second case.

The events $(A \cap B)$ and $(A \cap B^c)$ are mutually exclusive. Again, we have $P(A + B) = P(A) + P(B) - P(A \cup B)$.
By axiom (i), we have

$$P(A^c) \geq 0 \Rightarrow 1 - P(A) \geq 0 \Rightarrow P(A) \leq 1.$$

Again by axiom (i), we have $P(A) \geq 0$. Thus, $0 \leq P(A) \leq 1$. This proves the third case.

The events $(A \cap B)$ and $(A \cap B^c)$ are mutually exclusive. Again, we have

$$\begin{aligned} (A \cap B) \cup (A \cap B^c) &= A \\ P[(A \cap B) \cup (A \cap B^c)] &= P(A) \\ P(A \cap B) + P(A \cap B^c) &= P(A), \text{ by axiom (iii)} \\ P(A \cap B^c) &= P(A) - P(A \cap B). \end{aligned} \tag{1.3}$$

Similarly, the events $(A \cap B)$ and $(A^c \cap B)$ are mutually exclusive. Again, we have

$$\begin{aligned} (A \cap B) \cup (A^c \cap B) &= A \\ P[(A \cap B) \cup (A^c \cap B)] &= P(B) \\ P(A \cap B) + P(A \cap B^c) &= P(B), \text{ by axiom (iii)} \\ P(A^c \cap B) &= P(B) - P(A \cap B). \end{aligned} \tag{1.4}$$

Again, the events $(A \cap B)$, $(A \cap B^c)$ and $(A^c \cap B)$ are mutually exclusive. Thus,

$$\begin{aligned}
(A \cap B) \cup (A \cap B^c) \cup (A^c \cap B) &= A \cup B \\
P[(A \cap B) \cup (A \cap B^c) \cup (A^c \cap B)] &= P(A \cup B) \\
P(A \cap B) + P(A \cap B^c) + P(A^c \cap B) &= P(A \cup B), \text{ by axiom (iii)} \\
P(A \cap B) + P(A) - P(A \cap B) + P(B) - P(A \cap B) &= P(A \cup B).
\end{aligned}$$

Thus, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Remark: For any three events A, B, C this result becomes

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(B \cap C) - P(C \cap A) - P(A \cap B) + P(A \cap B \cap C).$$

Conditional Probability: Let S be the event space connected with the random experiment E. Let A and B be two events connected with E where $P(B) \neq 0$. The conditional probability of the event A on the hypothesis that the event B has already occurred is denoted by $P(A/B)$ and is defined by

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, P(B) \neq 0.$$

Theorem: The conditional probability satisfies all the axioms of probability.

Proof: (i) We see that for any two events A, B with $P(B) \neq 0$, $P(A/B) = \frac{P(AB)}{P(B)}$. Now by axiom (i) $P(AB) \geq 0$, $P(B) > 0$. Hence $P(A/B) \geq 0$.

(ii) Let S be the event space. If $P(B) \neq 0$, then

$$P(S/B) = \frac{P(SB)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

(iii) Let $A_1, A_2, ..., A_n, ...$ be countably infinite number of pairwise mutually exclusive points connected to the random experiment E. We have for any event B with $P(B) \neq 0$,

$$P\{(A_1 + A_2 + ... + A_n + ...)/B\} = \frac{P\{(A_1 + A_2 + ... + A_n + ...)B\}}{P(B)}$$

$$= \frac{P(A_1B + A_2B + ... + A_nB + ...)}{P(B)}$$

$$= \frac{P(A_1B) + P(A_2B) + ... + P(A_nB) + ...}{P(B)}, \quad \text{[by axioms (iii)]}$$

[Since events $A_1, A_2, ..., A_n, ...$ are mutually exclusive, $A_1B, A_2B, ..., A_nB, ...$ are also mutually exclusive events]

Therefore,

$$P\{(A_1 + A_2 + ... + A_n + ...)/B\} = \frac{P(A_1B)}{P(B)} + \frac{P(A_2B)}{P(B)} + ... + \frac{P(A_nB)}{P(B)} + ...$$

$$= P(A_1/B) + P(A_2/B) + ... + P(A_n/B) + ...$$

Hence the conditional probability satisfies all the axioms of probability.

Deduction of classical definition from axiomatic definition:

Let the event space S contain n mutually exclusive events $U_1, U_2, ..., U_n$. Then

$$\begin{aligned} S &= U_1 + U_2 + ... + U_n \\ P(S) &= P(U_1) + P(U_2) + ... + P(U_n) \\ 1 &= P(U_1) + P(U_2) + ... + P(U_n), \text{ [by axioms (ii) and (iii)].} \end{aligned} \tag{1.5}$$

If we assume that the event points are equally probable then we have

$$P(U_1) = P(U_2) = ... = P(U_n) = \frac{P(U_1) + P(U_2) + ... + P(U_n)}{n} = \frac{1}{n}.$$

If any event A contains m event points, say $U_1, U_2, ..., U_m$, then

$$\begin{aligned} A &= U_1 + U_2 + ... + U_m \\ P(A) &= P(U_1 + U_2 + ... + U_m) \\ &= P(U_1) + P(U_2) + ... + P(U_m) = \frac{m}{n}. \end{aligned}$$

Compound probability: If A and B be two events of a given random experiment then

$$P(AB) = P(A/B), \text{ P(B)} \neq 0.$$

$$P(AB) = P(B/A), \text{ P(A)} \neq 0.$$

It gives us a formula for calculating the probability of the product of two events and hence is often called the multiplication rule.

Independent events: An event B is said to be independent of an event A if the probability that B occurs is independent by whether A has or has not occurred. In other words, if the probability of B equals the conditional probability of B given A, i.e., $P(B/A) = P(B)$. Then from $P(AB) = P(A)P(B/A)$, we have

$$P(AB) = P(A)P(B).$$

Bayes Theorem: Let $A_1, A_2, ..., A_n$ be n pairwise mutually exclusive and exhaustive events connected with the random experiment E. Let X be an arbitrary event

connected with E, where $P(X) \neq 0$. Also, let the probabilities $P(X/A_1), P(X/A_2)$, ..., $P(X/A_n)$ be all known. Then

$$P(A_i/X) = \frac{P(A_i)P(X/A_i)}{\sum_{r=1}^{n} P(A_r)P(X/A_r)},\ i = 1, 2, 3, ...$$

Proof: Let S be the event space connected with the random experiment E. Therefore,

$S = A_1 + A_2 + ... + A_n$($\because A_1, A_2, ..., A_n$form an exhaustive set of events)

$$\begin{aligned}
\therefore\ XS &= X(A_1 + A_2 + ... + A_n) \\
X &= XA_1 + XA_2 + ... + XA_n \\
P(X) &= P(XA_1 + XA_2 + ... + XA_n)\ [\because (XA_i)(XA_j) = X(A_iA_j) = X\emptyset = \emptyset] \\
&= P(XA_1) + P(XA_2) + ... + P(XA_n) \\
&= P(A_1)P(XA_1) + P(A_2)P(XA_2) + ... + P(A_n)P(XA_n) \\
&= \sum_{r=1}^{n} P(A_r)P(X/A_r).
\end{aligned}$$

On the other hand, for any i, the conditional probability of A_i given X is defined by

$$\begin{aligned}
P(A_i/X) &= \frac{P(A_iX)}{P(X)}\ [\because P(X) \neq 0] \\
&= \frac{P(A_i)P(X/A_i)}{\sum_{r=1}^{n} P(A_r)P(X/A_r)}.
\end{aligned}$$

Properties of set:

(i) $\overline{A + B}$ = Portion shaded by horizontal line (see Fig. 1)

$\bar{A}\bar{B}$ = Common of horizontal and vertical lines (see Fig. 2) = horizontal line of Fig. 1 $= \overline{A + B}$

Thus, $\overline{A + B} = \bar{A}\bar{B}$.

Generalizing, we get

$$\overline{A_1 + A_2 + ... + A_n} = \bar{A}_1\bar{A}_2.....\bar{A}_n$$

Expending Sequence: A sequence $\{A_n\}$ is said to be expanding or monotonic non-decreasing if $A_n \subseteq A_{n+1}$ for all n and $\{A_n\}$ is said to be contracting or monotonic nonincreasing if $A_{n+1} \subseteq A_n$ for all n.

Illustrations: Let $A_1 = \{1, 2\}$, $A_2 = \{1, 2, 3\}$, $A_3 = \{1, 2, 3, 4, 5\}$.
Then $A_1 + A_2 = \{1, 2, 3\} = A_2$.
$(A_1 + A_2) + A_3 = \{1, 2, 3, 4, 5\} = A_3$.
By the method of induction, we have

$$A_1 + A_2 + ... + A_n = A_n.$$

Thus, if $\{A_n\}$ be an expanding sequence of sets then

$$\lim_{n\to\infty} A_n = \sum_{n=1}^{\infty} A_n.$$

Again consider $A_1 = \{1, 2, 3, 4, 5\}$, $A_2 = \{1, 2, 3\}$, $A_3 = \{1, 2\}$.
Then $A_1A_2 = \{1, 2, 3\} = A_2$.
$(A_1A_2)A_3 = \{1, 2\} = A_3$.
By the method of induction, we have

$$A_1A_2.....A_n = A_n.$$

Thus, if $\{A_n\}$ be an contracting sequence of sets then

$$\lim_{n\to\infty} A_n = \Pi_{n=1}^{\infty} A_n \ (i.e \cap_{n=1}^{\infty} A_n).$$

Observations:
(i) If $\{A_n\}$ is an expanding sequence then $\{\bar{A}_n\}$ is a contracting sequence.
(ii) If $\{A_n\}$ is an contracting sequence then $\{\bar{A}_n\}$ is an expanding sequence.

Theorem: Prove that If $\{A_n\}$ is either expanding or contracting sequence then

$$\overline{\lim_{n\to\infty} A_n} = \lim_{n\to\infty} \bar{A}_n.$$

Proof: Let $\{A_n\}$ be an expanding sequence of sets. Then we have

$$\lim_{n\to\infty} A_n = \sum_{n=1}^{\infty} A_n.$$

Therefore,

$$\overline{\lim_{n\to\infty} A_n} = \overline{\sum_{n=1}^{\infty} A_n} = \overline{A_1 + A_2 + ... + A_n + ...} = \bar{A}_1\bar{A}_2.....\bar{A}_n.....,$$

or

$$\overline{\lim_{n\to\infty} A_n} = \Pi_{n=1}^{\infty} \bar{A}_n = \lim_{n\to\infty} \bar{A}_n.$$

[Since $\{A_n\}$ is an expanding sequence, $\{A_n\}$ is a contracting sequence and for contracting sequence

$$\lim_{n\to\infty} A_n = \Pi_{n=1}^{\infty} A_n.]$$

Again let $\{A_n\}$ be an contracting sequence of sets. Then we have

$$\lim_{n\to\infty} A_n = \Pi_{n=1}^{\infty} A_n.$$

Therefore,

$$\overline{\lim_{n\to\infty} A_n} = \overline{\Pi_{n=1}^{\infty} A_n} = \overline{A_1 A_2 .....A_n .....} = \bar{A}_1 \bar{A}_2 .....\bar{A}_n ....., = \sum_{n=1}^{\infty} \bar{A}_n$$

[Since for a contracting sequence, $A_1 A_2 .....A_n = A_n$.]

Therefore,

$$\overline{\lim_{n\to\infty} A_n} = \lim_{n\to\infty} \bar{A}_n.$$

[Since $\{A_n\}$ is a contracting sequence, $\{A_n\}$ is an expanding sequence and for expanding sequence

$$\lim_{n\to\infty} A_n = \sum A_n.]$$

Theorem: If $\{A_n\}$ be a monotonic sequence of events then

$$P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n).$$

Proof: First suppose that $\{A_n\}$ is monotonic increasing, so we have

$$\lim_{n\to\infty} A_n = \sum_{n=1}^{\infty} A_n .....(1)$$

Now set $B_1 = A_1, B_n = A_n - An - 1, n \geq 2$.

Therefore, $\{A_n\}$ is a sequence of pairwise mutually exclusive events such that

$$\sum_{n=1}^{\infty} A_n = \sum_{n=1}^{\infty} B_n .....(2)$$

Also, for every $n \in N$(set of natural numbers)

$$A_n = \sum_{i=1}^{n} B_i .....(3)$$

Therefore,

$$P(\lim_{n\to\infty} A_n) = P(\sum_{n=1}^{\infty} A_n) \ \ [by(1)] = P(\sum_{n=1}^{\infty} B_n) \ \ [by(2)] = \sum_{n=1}^{\infty} P(B_n) \ \ [by\ axiom\ (iii)]$$

That is

$$P(\lim_{n\to\infty} A_n) = \lim_{n\to n} \sum_{i=1}^{\infty} P(B_i) = \lim_{n\to n} P(\sum_{i=1}^{\infty} B_i) \ [by\ axiom\ (iii)] \ = \lim_{n\to\infty} P(A_n) \ [by\ (iii)].$$

Next we suppose that $\{A_n\}$ is nonincreasing sequence of events. Therefore, $\{\bar{A}_n\}$ is expanding sequence. Thus, from previous results, we have

$$P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(\bar{A}_n).....(4)$$

We also know that for any monotonic sequence $\{A_n\}$,

$$\overline{\lim_{n\to\infty} A_n} = \lim_{n\to\infty} \bar{A}_n.$$

That is left hand side of (4) becomes

$$P(\lim_{n\to\infty} \bar{A}_n) = P(\overline{\lim_{n\to\infty} A_n}) = 1 - P(\lim_{n\to\infty} \bar{A}_n) \;\; [since P(\bar{A}) = 1 - P(A)].$$

Right hand side of (4) can be written as

$$\lim_{n\to\infty} P(\bar{A}_n) = \lim_{n\to\infty} (1 - P(A_n)) = 1 - \lim_{n\to\infty} P(A_n).$$

Thus, using (4), we have

$$1 - P(\lim_{n\to\infty} A_n) = 1 - \lim_{n\to\infty} P(A_n).$$

Therefore,

$$P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n).$$

Random Variable: In many experiments, we are interested not in knowing which of the outcomes has occurred, but in the numbers associated with them. For example, when n coins are tested, we want to know the number of heads obtained. As another example, when two dice are thrown, we seek information about the sum of points and are not really concerned with the separate number on each face. Thus, we associate a real number with each outcome of an experiment. Such a real-valued function defined on a sample space S is called a random variable.

Definition Let E be a random experiment and S be the event space connected with E. Let Δ be the $\sigma-$field of subsets of S. A mapping $X : S \to R$ is called a random variable if the inverse image under X of all semi-closed intervals of the form $(-\infty, x], x \in R$ are events in Δ, i.e.,

$$X^{-1}(-\infty, x] = \{\omega \in S : \; -\infty < X(\omega) \leq x\} \in \Delta.$$

The range of the mapping $X : S \to R$ is called the spectrum of the random variable X. The spectrum may be discrete and continuous and accordingly the r.v. is said to be discrete or continuous.

Notes: (i) Note that the definition uses only the notions of sample space S and the $\sigma-$field Δ associated with S. It has not used the notion of probability.
(ii) Note that X is not a variable and it is not random either. Observe that X assigns values to the outcomes ω of the sample space S, which are random.

(iii) If S is a discrete sample space having finite or countable number of sample points, then Δ can be taken as the set of all subsets of S. If S is a continuous sample space like $I = [a, b]$ or the whole real line F then the associated $\sigma-$field Δ is governed by the collection of all semi-closed intervals of the forms $(a, b]$ contained in it.

Example: Consider the experiment of tossing a coin. Then $S = \{H, T\}$. Let X be the function that defines the "number of heads" in the outcome ω. Then $X\{H\} = 1, \quad X\{T\} = 0$. Then we have

$$X^{-1}(-\infty, x] = \begin{cases} \phi, \ if \ x < 0 \\ \{T\}, \ if \ 0 \leq x < 1, \\ \{H, T\}, \ if \ x \geq 1. \end{cases}$$

Example: A coin is tossed twice. Here

$$S = \{(HH) = \omega_1, (HT) = \omega_2, (TH) = \omega_3, (TT) = \omega_4\}.$$

A mapping $X : S \to R$ is defined as follows:

$$X(\omega_i) = k, \text{ where k is the number of heads, } i = 1, 2, 3, 4$$

Thus, $X(\omega_1) = 2, X(\omega_2) = X(\omega_3) = 1, X(\omega_4) = 0$.
Then we have

$$X^{-1}(-\infty, x] = \begin{cases} \phi, \ if \ x < 0 \\ \{TT\}, \ if \ 0 \leq x < 1, \\ \{TT, HT, TH\}, \ if \ 1 \leq x < 2, \\ \{TT, HT, TH, HH\} = S, \ if \ x \geq 2. \end{cases}$$

Here X is a r.v. defined in the domain S and the spectrum (range) of X is $\{0, 1, 2\}$. According to our notation, $X = 0$ represents the event $\{(TT)\}$; $0 \leq X \leq 2$ is the certain event and $1 < X < 2$ represents the impossible event $\emptyset$.

Remark: The above random variable $X : S \to R$ is also described in the following manner: The r.v. X defined on S denotes the total number of heads in two tosses of the coin.

Example: Consider the experiment of throwing a fair die. Then $S = \{1, 2, 3, 4, 5, 6\}$. We can define a function X on S as

$$X(\omega) = \begin{cases} 1, \ if \ \omega \ is \ even \\ -1, \ if \ \omega \ is \ odd. \end{cases}$$

be the function that defines the "number of heads" in the outcome ω. Then $X\{H\} = 1, \quad X\{T\} = 0$. Then we have

$$X^{-1}(-\infty, x] = \begin{cases} \phi, \ if \ x < -1 \\ \{1, 3, 5\}, \ if \ -1 \leq x < 1, \\ S, \ if \ x \geq 1. \end{cases}$$

Thus, it follows that X is a random variable.

Theorem: Let S be a sample space and Δ be the σ-field of subsets of S. Then the following properties hold:
(i) X is a random variable on S if and only if for each real x,

$$X^{-1}(-\infty, x] = \{\omega \in S : \ -\infty < X(\omega) \leq x\} \in \Delta.$$

(ii) If X_1 and X_2 are random variables and α is a real constant then $X_1+X_2, \alpha X_1, X_1X_2$ are also random variables.
(iii) If X is a random variables then X^2 and $|X|$ also are random variables.
(iv) If X is a random variable taking nonnegative values then $\sqrt{X}$ also is a random variable.
(v) If X is a random variable then $\frac{1}{X}$ also is a random variable provided $X^{-1}\{0\} \neq \phi$, i.e., there exists no $\omega \in S$ for which $X(\omega) = 0$.
(vi) If X_1 and X_2 are random variables then $\max(X_1, X_2)$ and $\min(X_1, X_2)$ are also random variables.
(vii) If X is a random variable and $g : R \to R$ is a continuous function then $g(X)$ also is a random variable.

Distribution function of a random variable: Till now we have learned how to associate events with the random variable. However, random variable will be of real interest only when they are defined on probability space (S, Δ, P).

Definition: Let (S, Δ, P) be a probability space and X is a random variable defined on S. The function $F : R \to R$ defined by

$$F(X) = P(-\infty, x] = P\{\omega \in S : \ -\infty < X(\omega) \leq x\}, \forall x \in R.$$

is called the distribution function or cumulative distribution function (CDF) of the random variable X.

Notes: (i) In practice, $F(x)$ is abbreviated as $F(x) = P\{X \leq x\}$. Sometimes we use the notation $F_X(x)$ to emphasize that the distribution function is associated with the random variable X.
(ii) Clearly, the domain of the distribution function is $(-\infty, \infty)$ and its range is $[0, 1]$.

Example: Consider the experiment of tossing three coins. Let X be a random variable that defines the number of heads in the outcome. Then

$$S = \{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT\}.$$

Now

$$P\{X = 0\} = P\{TTT\} = \frac{1}{8}$$

$$P\{X = 1\} = P\{HTT, THT, TTH\} = \frac{3}{8}$$

$$P\{X = 2\} = P\{HHt, HTH, THH\} = \frac{3}{8}$$

$$P\{X = 3\} = P\{HHH\} = \frac{1}{8}.$$

Hence the distribution function of X is given by

$$F(x) = \begin{cases} 0, \ if \ x < 0 \\ \frac{1}{8}, \ if \ 0 \leq x < 1, \\ \frac{4}{8}, \ if \ 1 \leq x < 2, \\ \frac{7}{8}, \ if \ 2 \leq x < 3, \\ \frac{8}{8} = 1, \ if \ x \geq 3. \end{cases}$$

If we plot the distribution function, we get the following graph:

Properties of distribution function:
(a) $0 \leq F(x) \leq 1$,
(b) $P(a < X < b) = F(b) - F(a)$,
(c) F is monotonically increasing.
(d) $F(\infty) = 1$ (e) $F(-\infty) = 0$.
(f) F is right continuous.

Proof:
(a) By axiom (i), we have

$$0 < P(-\infty < X \leq x) \leq 1.$$

So $0 \leq F(x) \leq 1, \forall x \in (-\infty, \infty)$.

(b) The events $(-\infty < X \leq a)$ and $(a < X \leq b)$ are mutually exclusive and

$$(-\infty < X \leq a) + (a < X \leq b) = (-\infty < X \leq b).$$

Therefore, by axiom (iii), we have

$$P(-\infty < X \leq a) + P(a < X \leq b) = P(-\infty < X \leq b).$$

i.e.,

$$F(a) + P(a < X \leq b) = F(b)$$
$$\Rightarrow F(b) - F(a) = P(a < X \leq b).$$

(c) Let $x_1 > x_2, x_1, x_2 \in R$. Then we have following property (b)

$$P(a < X \leq b) = F(x_2) - F(x_1).$$

Since $P(a < X \leq b) \geq 0$, so we have $F(x_2) \geq F(x_1)$. This shows that F is monotonically increasing.

(d) Let A_n denote the event $-\infty < X \leq n,\ n = 1, 2, 3.....$
Then $\{A_n\}$ is an increasing events such that

$$\lim_{n\to\infty} A_n = -\infty < X < \infty = S.$$

Therefore,

$$P(\lim_{n\to\infty} A_n) = P(S) = 1.....(1)$$

From definition, we have

$$P(A_n) = P(-\infty < X \leq n) = F(n).$$

Thus,

$$\lim_{n\to\infty} P(A_n) = F(\infty).....(2)$$

From (1) and (2), we have

$$F(\infty) = \lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n) = 1.$$

(e) Let A_n denote the event $-\infty < X \leq -n,\ n = 1, 2, 3.....$
Then $\{A_n\}$ is a contracting events such that

$$\lim_{n\to\infty} A_n = -\infty < X < -\infty = \phi.$$

Therefore,

$$P(\lim_{n\to\infty} A_n) = P(\phi) = 0.....(1)$$

From definition, we have

$$P(A_n) = P(-\infty < X \leq -n) = F(-n).$$

Thus,

$$\lim_{n\to\infty} P(A_n) = F(-\infty).....(2)$$

From (1) and (2), we have

$$F(-\infty) = \lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n) = 0.$$

(f) In this case, we have to show that

$$F(x+0) = F(x), \forall x \in R,$$

where

$$F(x+0) = \lim_{h\to 0} F(x+h).$$

The events $\{-\infty < X \leq x\}$ and $\{x < X \leq x+h\}, h > 0$ are mutually exclusive. Also

$$\{-\infty < X \leq x\} \cup \{x < X \leq x+h\} = \{-\infty < X \leq x+h\}$$

$$\therefore \quad P(-\infty < X \leq x) + P(x < X \leq x+h) = P(-\infty < X \leq x+h)$$

$$\Rightarrow \quad P(x < X \leq x+h) = F(x+h) - F(x)$$

$$\therefore \quad \lim_{h\to 0}[F(x+h) - F(x)] = \lim_{h\to 0} P(x < X \leq x+h) = P(\lim_{h\to 0}\{x < X \leq x+h\})$$

[Since the events $(x < X \leq x + h_n)$ forms a decreasing sequence of events as $h_n \to 0$, so $\lim_{n\to\infty} A_n = \cap_{n=1}^{\infty} A_n$.]

$$Or, \quad F(x+0) - F(x) = P(\phi) = 0$$

$$\therefore \quad F(x+0) = F(x).$$

Hence F is right continuous.

Theorem: A function $F : R \to R$ is a distribution function for some probability space (S, Δ, P) if and only if F satisfies the following properties:

(i) F is nondecreasing,
(ii) F is right continuous,
(iii) $F(-\infty) = 0$ and $F(+\infty) = 1$.

Theorem: Let F be a distribution function of a random variable X defined on a probability space (S, Δ, P). Then

$$F(x) - F(x-0) = P(X = x),$$

where $F(x-0) = \lim_{h\to 0} F(x-h)$.

Proof: The events $\{-\infty < X \leq x-h\}$ and $\{x-h < X \leq x\}, h > 0$ are mutually exclusive. Also

$$\{-\infty < X \leq x-h\} \cup \{x-h < X \leq x\} = \{-\infty < X \leq x\}$$

$$\therefore \quad P(-\infty < X \leq x-h) + P(x-h < X \leq x) = P(-\infty < X \leq x)$$

$$\Rightarrow \quad P(x-h < X \leq x) = F(x) - F(x-h)$$

$$\therefore \quad \lim_{h\to 0}[F(x) - F(x-h)] = \lim_{h\to 0} P(x-h < X \leq x) = P(\lim_{h\to 0}\{x-h < X \leq x\})$$

[Since the events $(x-h_n < X \leq x)$ forms an increasing sequence of events as $h_n \to 0$, so $\lim_{n\to\infty} A_n = \cup_{n=1}^{\infty} A_n$.]

$$Or, \quad F(x) - F(x-0) = P(X = x)$$

Hence the proof.

Theorem: A necessary and sufficient condition for a distribution function F of a random variable X to be continuous is that $P(X = x) = 0, \ \forall x \in R$.

Proof: We know that the distribution function is right continuous, i.e.,

$$F(x+o) = F(x), \ \forall x \in R.$$

Hence F is continuous if and only if F is also left continuous, i.e.,

$$F(x-0) = F(x), \ \forall x \in R.$$

$$i.e. \ if \ F(x) - F(x-0) = 0$$

$$or, \ P(X = x) = 0.$$

Hence the proof.

Example: A random variable X assumes the values -1,0,1 with probabilities $\frac{1}{3}, \frac{1}{2}, \frac{1}{6}$ respectively. Determine the distribution.

Solution: Let $F(X)$ be the distribution function of the random variable X.
In $-\infty < x < -1$, $F(X) = P(-\infty < X \leq x) = 0$
In $-1 \leq x < 0$, $F(X) = P(X = -1) = \frac{1}{3}$
In $0 \leq x < 1$, $F(X) = P(X = -1) + P(X = 0) = \frac{1}{3} + \frac{1}{2} = \frac{5}{6}$
and in $1 \leq x < \infty$, $F(X) = P(X = -1) + P(X = 0) + P(X = 1) = \frac{1}{3} + \frac{1}{2} + \frac{1}{6} = 1$.

Example: Let $F(X) = 0, \quad -\infty < x < 0$
$= \frac{1}{5}, \quad 0 \leq x < 1.$
$= \frac{3}{5}, \quad 1 \leq x < 3.$
$= 1, \quad 3 \leq x < \infty.$
Show that $F(X)$ is a possible distribution function and determine spectrum and probability mass function of the distribution.

Solution: Since $F(X)$ is non-decreasing function everywhere, continuous on right at every point and $F(-\infty) = 0$, $F(\infty) = 1$. Hence it is a possible distribution function. The spectrum of $F(X)$ is 0,1,3 and the probability mass function is given by

$$f_0 = P(X = 0) = F(0) - F(0 - 0) = \frac{1}{5} - 0 = \frac{1}{5}$$

$$f_1 = P(X = 1) = F(1) - F(1 - 0) = \frac{3}{5} - \frac{1}{5} = \frac{2}{5}$$

$$f_2 = P(X = 3) = F(3) - F(3 - 0) = 1 - \frac{3}{5} = \frac{2}{5}.$$

Expectation: If X is a discrete random variable assuming values $x_1, x_2, ..., x_n$ with probabilities $p_1, p_2, ..., p_n$, then its expectation $E(X)$ is defined as the sum $\Sigma x_i p_i$.
For discrete random variable, if $f(x)$ is the probability mass function (p.m.f) of X, then $E(x) = \Sigma x_i f(x_i)$.
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

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx, \quad \text{provided the integral exists.}$$

Properties of expectation:
(i) $E(X \pm Y) = E(X) \pm E(Y)$
(ii) $E(XY) = E(X)E(Y)$, if X, Y are independent variable
(iii)$E(C) = C$, where C is constant
(iv)$E(aX) = aE(X)$
(v)$[E(X)]^2 \leq E(X^2)]$ (Schwarz's inequality)
(vi) $\{E(XY)^2\} \leq E(X^2)E(Y^2)$ (Cauchy-Schwarz inequality).

Variance: For a random variable X, variance is defined by $Var(X) = E(X^2) - \{E(X)\}^2$.
If X is a discrete random variable assuming values $x_1, x_2, ..., x_n$ with probabilities $p_1, p_2, ..., p_n$, then its variance Var(X) is defined as the $\Sigma x_i^2 p_i - (\Sigma x_i p_i)^2$.

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

$$
\begin{aligned}
&= P(X > 2) = 1 - P(X \leq 2) \\
&= 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\} \\
&= 1 - e^{-1.5}\left\{1 + 1.5 + \tfrac{(1.5)^2}{2}\right\} \\
&= 0.1916.
\end{aligned}
$$

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

$$
\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \qquad (1.6) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \qquad (1.7)
\end{aligned}
$$

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 $x \to -\infty$.

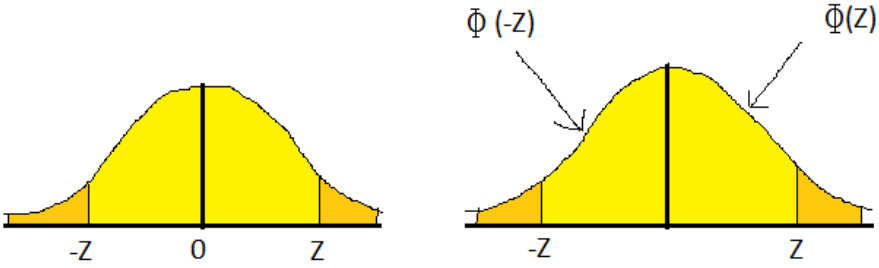
$$
\begin{aligned}
\therefore E(X) &= \frac{1}{\sqrt{2\pi}\sigma} \sqrt{2}\sigma \int_{-\infty}^{\infty} (\mu + \sqrt{2}\sigma t) e^{-t^2} dt \\
&= \frac{\mu}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt + \frac{\sqrt{2}\sigma}{\sqrt{\pi}} \int_{-\infty}^{\infty} t e^{-t^2} dt \\
&= \frac{\mu}{\sqrt{\pi}} \times \sqrt{\pi} + \frac{\sqrt{2}\sigma}{\sqrt{\pi}} \times 0 = \mu \\
&\quad \text{since } te^{-t^2} \text{ is and odd function and } \int_{-\infty}^{\infty} e^{-t^2} dt = \sqrt{\pi}.
\end{aligned}
$$

Now

$$\begin{aligned} E(X^2) &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} x^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx \qquad (1.8)\\ &= \frac{1}{\sqrt{2\pi}\sigma}\sqrt{2}\sigma\int_{-\infty}^{\infty}(\mu+\sqrt{2}\sigma t)^2 e^{-t^2}dt \\ &= \frac{\mu^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-t^2}dt + \frac{2\sqrt{2}\mu\sigma}{\sqrt{\pi}}\int_{-\infty}^{\infty} te^{-t^2}dt + \frac{2\sigma^2}{\sqrt{\pi}}\int_{-\infty}^{\infty} t^2 e^{-t^2}dt \\ &= \mu^2 + 0 + \sigma^2, \text{ since } \int_{-\infty}^{\infty} t^2 e^{-t^2}dt = \frac{1}{2}\sqrt{\pi}. \qquad (1.9) \end{aligned}$$

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	$\Sigma d_i = \Sigma \lvert x_i - 300 \rvert = 370$

$\therefore$ 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. $= \frac{1}{N}\Sigma f_i \lvert x_i - 45 \rvert = \frac{400}{40} = 10.$

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

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

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

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

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

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

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	$f_i x_i$	$x_i - \bar{X} = x_i - 9$	$(x_i - \bar{X})^2$	$f_i(x_i - \bar{X})^2$
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	$N = \Sigma f_i = 50$	$\Sigma f_i x_i = 450$			$\Sigma f_i(x_i - \bar{X})^2$=754

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

$\therefore \bar{X} = \frac{1}{N}\Sigma f_i x_i = \frac{450}{50} = 9$ 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	$f_i d_i$	$f_i d_i^2$
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	$N = \Sigma f_i = 217$			$\Sigma f_i d_i = 128$	$\Sigma f_i d_i^2$=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 : **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* : **1.997, 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 : **55.1, 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

possible mathematical relation between the two. Very naturally, this relation depends on the assumption of independence and dependence of variables. That is, if x is taken as independent variable, y as dependent variable, then we get one relation, called the *regression equation of y on x*, and similarly if y is taken as independent variable and x as an dependent variable, then we get another relation, called the *regression equation x on y*. It is to be noted that under some stringent conditions the two relations may be identical.

Correlation analysis

Consider the following bivariate data:

$$\begin{array}{lccccc} x: & x_1 & x_2 & x_3 & \ldots & x_n \\ y: & y_1 & y_2 & y_3 & \ldots & y_n \end{array}$$

Then the covariance of the two variables x and y is denoted by $\mathrm{Cov}(x, y)$ and defined by

$$\mathrm{Cov}(x.y) = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

where,

$$\bar{x} = \frac{1}{n}\Sigma_{i=1}^{n}x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\Sigma_{i=1}^{n}y_i.$$

Another form of the covariance formula is

$$\mathrm{Cov}(x, y) = \frac{1}{n}\Sigma_{i=1}^{n}x_i y_i - \bar{x}\bar{y}$$

which one can deduce from the above covariance formula.

In fact

$$\begin{aligned} \mathrm{Cov}(x.y) &= \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n}\Sigma_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) \\ &= \frac{1}{n}\Sigma_{i=1}^{n}x_i y_i - \bar{y}\frac{1}{n}\Sigma_{i=1}^{n}x_i - \bar{x}\frac{1}{n}\Sigma_{i=1}^{n}y_i + \frac{1}{n}\Sigma_{i=1}^{n}\bar{x}\bar{y} \\ &= \frac{1}{n}\Sigma_{i=1}^{n}x_i y_i - \bar{y}\bar{x} - \bar{x}\bar{y} + \frac{1}{n}n\bar{x}\bar{y} \\ &= \frac{1}{n}\Sigma_{i=1}^{n}x_i y_i - \bar{x}\bar{y} \end{aligned}$$

Karl Pearson correlation coefficient

The correlation coefficient of the two variables x and y is denoted by r and is defined by

$$r_{xy} = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

where $\sigma_x = \frac{1}{n}\Sigma_{i=1}^{n} x_i^2 - (\bar{x})^2$ and $\sigma_y = \frac{1}{n}\Sigma_{i=1}^{n} y_i^2 - (\bar{y})^2$.

Hence,

$$r_{xy} = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\Sigma_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\Sigma_{i=1}^{n} x_i^2 - (\bar{x})^2\right)\left(\frac{1}{n}\Sigma_{i=1}^{n} y_i^2 - (\bar{y})^2\right)}}$$

Property 1. The correlation coefficient r is a pure number and is independent of units of measurement, i.e., it has no unit.

Property 2. The correlation of coefficient r is independent of the choice of origin.

Proof: Let (x, y) and (u, v) be the two sets of bivariate data such that $u = x - a$ and $v = y - b$ where a and b are constants.

$$\therefore \bar{u} = \bar{x} - a \quad \text{and} \quad \bar{v} = \bar{y} - b.$$

$$\therefore \bar{u} = \bar{x} - a \quad \text{and} \quad \bar{v} = \bar{y} - b.$$

$$\therefore \text{var}(u) = \sigma_u^2 = \frac{1}{n}\Sigma_{i=1}^{n}(u_i - \bar{u})^2 = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - a - \bar{x} + a)^2 = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^2 = \sigma_x^2.$$

Similarly $\text{var}(v) = \sigma_v^2 = \sigma_y^2$.

$$\begin{aligned}
\text{Cov}(u.v) &= \frac{1}{n}\Sigma_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v}) \\
&= \frac{1}{n}\Sigma_{i=1}^{n}(x_i - a - \bar{x} + a)(y_i - b - \bar{y} + b) \\
&= \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \text{Cov}(x, y)
\end{aligned}$$

$$\therefore \quad r_{uv} = \frac{\text{Cov}(u,v)}{\sigma_u \sigma_v} = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} = r_{xy}.$$

Property 3. let (x, y) and (u, v) be such that $u = ax + b$ and $v = cy + d$, where a, b, c, d are constants; then

$$r_{uv} = \frac{ac}{|a||c|} r_{xy}$$

$$= \begin{cases} r_{xy} & \text{when } a \text{ and } c \text{ have the same sign} \\ -r_{xy} & \text{when } a \text{ and } c \text{ have opposite sign} \end{cases}$$

Proof: Since $u = ax + b$ and $v = cy + d$, then $\bar{u} = a\bar{x} + b$ and $\bar{v} = c\bar{y} + d$.

$$\begin{aligned}
\text{var}(u) &= \sigma_u^2 = \frac{1}{n}\Sigma_{i=1}^{n}(u_i - \bar{u})^2 = \frac{1}{n}\Sigma_{i=1}^{n}(ax_i + b - a\bar{x} - b)^2 \\
&= a^2 \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^2 = a^2\sigma_x^2.
\end{aligned}$$

$\therefore \sigma_u = |a|\ \sigma_x$. Similarly, $\sigma_v = |c|\ \sigma_y$.

Now,

$$\begin{aligned} \text{Cov}(u,v) &= \frac{1}{n}\Sigma_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v}) && (1.10)\\ &= \frac{1}{n}\Sigma_{i=1}^{n}(ax_i+b-a\bar{x}-b)(cy_i+d-c\bar{y}-d) && (1.11)\\ &= \frac{1}{n}\ ac\ \Sigma_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) && (1.12)\\ &= ac\ \text{Cov}(x,y). && (1.13) \end{aligned}$$

$$\therefore r_{uv} = \frac{\text{Cov}(u,v)}{\sigma_u\sigma_v} = ac\ \frac{\text{Cov}(x,y)}{|a|\ \sigma_x|c|\ \sigma_y} = r_{uv} = \frac{ac}{|a||c|}r_{xy}$$

$$= \begin{cases} r_{xy} & \text{when } a \text{ and } c \text{ have the same sign} \\ -r_{xy} & \text{when } a \text{ and } c \text{ have opposite signs.} \end{cases}$$

Property 4. The value of r lies between -1 and 1, i.e., $-1 \leq r \leq 1$.

Proof: Let u_i and v_i be the two sets of two variables such that

$$u_i = \frac{x_i-\bar{x}}{\sigma_x} \text{ and } v_i = \frac{y_i-\bar{y}}{\sigma_y}$$

where the symbols on the r.h.s have usual meaning.

$$\therefore \Sigma u_i^2 = \Sigma\frac{(x_i-\bar{x})}{\sigma_x^2} = \frac{1}{\sigma_x^2}\Sigma(x_i-\bar{x})^2 = \frac{1}{\sigma_x^2}n\sigma_x^2 = n.$$

Similarly, $\Sigma v_i^2 = n$.

Again,

$$\begin{aligned} \Sigma u_i v_i &= \Sigma\frac{(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y} = \frac{n}{\sigma_x\sigma_y}\Sigma\frac{(x_i-\bar{x})(y_i-\bar{y})}{n}\\ &= \frac{n\ \text{Cov}(x,y)}{\sigma_x\sigma_y} = nr_{xy}. \end{aligned}$$

Now $(u_i \pm v_i)^2$ cannot be negative.

$$\begin{aligned} \therefore \Sigma(u_i \pm v_i)^2 &\geq 0\\ \implies \Sigma(u_i^2+v_i^2 \pm 2u_iv_i) &\geq 0\\ \implies \Sigma u_i^2 + \Sigma v_i^2 \pm 2\Sigma u_iv_i &\geq 0\\ \implies n+n+\pm 2nr_{xy} &\geq 0\\ \implies 2n(1\pm r_{xy}) &\geq 0 \implies (1\pm r_{xy}) \geq 0. \end{aligned}$$

Hence, $-1 \leq r_{xy} \leq 1$.

Property 5. Let (x, y) represent bivariate data for the two variables x and y. Then, $\text{var}(x \pm y) = \sigma_x^2 + \sigma_y^2 + \pm 2 r_{xy} \sigma_x \sigma_y$.

Proof: By definition $\text{var}(x \pm y) = \frac{1}{n}\Sigma_{i=1}^n [(x_i \pm y_i) - (\bar{x} \pm \bar{y})]^2$

$$\begin{aligned}
&= \frac{1}{n}\Sigma_{i=1}^n [(x_i - \bar{x}) \pm (y_i - \bar{y})]^2 \\
&= \frac{1}{n}\Sigma_{i=1}^n [(x_i - \bar{x})^2 + (y_i - \bar{y})^2 \pm 2(x_i - \bar{x})(y_i - \bar{y})] \\
&= \frac{1}{n}\Sigma_{i=1}^n (x_i - \bar{x})^2 + \frac{1}{n}\Sigma_{i=1}^n (y_i - \bar{y})^2 \pm 2\frac{1}{n}\Sigma_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\
&= \sigma_x^2 + \sigma_y^2 \pm 2 \text{ Cov}(x, y) \\
&= \sigma_x^2 + \sigma_y^2 \pm 2 \ r_{xy} \sigma_x \sigma_y
\end{aligned}$$

Notes:

1. The standard error of correlation coefficient is given by $\frac{1-r_{xy}^2}{\sqrt{n}}$.
2. If two variables x and y are uncorrelated, then $r_{xy} = 0$.
3. Probable error $= 0.67485 \times \frac{1-r_{xy}^2}{\sqrt{n}}$.

Example. Find the correlation coefficient of the following data:

x :	65	63	67	64	68	62	70	66
y :	68	66	68	65	69	66	68	65

Solution: Since the correlation coefficient is unaffected by change of origin, let us change the origin of x and y to 65 and 687, respectively.

Then, we write $u = x - 65$ and $v = y - 67$.

x	y	u	v	u^2	v^2	uv
65	68	0 ‘	1	0	1	0
63	66	-2	-1	4	1	2
67	68	2	1	4	1	2
64	65	-1	-2	1	4	2
68	69	3	2	9	4	6
62	66	-3	-1	9	1	3
70	68	5	1	25	1	5
66	65	1	-2	4	4	-2
		$\Sigma u = 5$	$\Sigma v = -1$	$\Sigma u^2 = 53$	$\Sigma v^2 = 17$	$\Sigma uv = 18$

$$\therefore \sigma_u^2 = \frac{1}{n}\Sigma u^2 - (\bar{u})^2 = \frac{1}{8}\ 53 - \left(\frac{5}{8}\right)^2 = \frac{399}{64} \tag{1.14}$$

$$\sigma_v^2 = \frac{1}{n}\Sigma v^2 - (\bar{v})^2 = \frac{1}{8}\ - \left(\frac{-1}{8}\right)^2 = \frac{135}{64}. \tag{1.15}$$

$$\mathrm{Cov}(u,v) = \frac{1}{n}\Sigma uv - \bar{u}\bar{v} = \frac{1}{8}\ 18 - \left(\frac{5}{8}\right)\left(\frac{-1}{8}\right) = \frac{149}{64} \tag{1.16}$$

$$r_{xy} = r_{uv} = \frac{\mathrm{Cov}(u,v)}{\sigma_u \sigma_v} = \frac{\frac{149}{64}}{\sqrt{\frac{399}{64}}\sqrt{\frac{135}{64}}} = \frac{149}{\sqrt{399 \times 136}} = 0.64. \tag{1.17}$$

hence the required correlation coefficient is 0.64.

Example. Find the correlation coefficient of the following data:

$x:$	23.3	17.5	17.8	20.7	18.1	20.9	22.9	20.8
$y:$	4.2	3.8	4.6	3.2	5.2	4.7	1.1	5.6

Solution: We know that the correlation coefficient is unaffected by the change of origin and scale. Therefore, we assume

$$u = \frac{x - 20.7}{1} \quad \text{and} \quad v = \frac{y - 4.4}{1}.$$

x	y	u	v	u^2	v^2	uv
23.5	4.2	26 ‘	-2	676	4	-52
17.5	3.8	-32	–6	1024	36	192
17.8	4.6	-29	2	841	4	-58
20.7	3.2	-0	-12	0	144	0
18.1	5.2	-26	8	676	64	-208
20.9	4.7	2	3	4	9	6
22.9	4.4	22	0	484	0	0
20.8	5.6	1	12	1	144	12
		$\Sigma u = -36$	$\Sigma v = 5$	$\Sigma u^2 = 3076$	$\Sigma v^2 = 405$	$\Sigma uv = -108$

We know

$$r_{xy} = r_{uv} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{[n\Sigma u^2 - (\Sigma u)^2]\,[n\Sigma v^2 - (\Sigma v)^2]}} \tag{1.18}$$

$$= \frac{8 \times (-108) - (-36) \times 5}{\sqrt{[8 \times 3706 - (-36)^2]\,[8 \times 405 - 5^2]}} \tag{1.19}$$

$$= -\frac{684}{6547.34} = -0.0716. \tag{1.20}$$

$$\begin{aligned}\text{Probability of error is given by P.E.} &= 0.6745 \times \frac{1-r^2}{\sqrt{n}} \\ &= 0.6745 \times \frac{1-(-0.072)^2}{\sqrt{8}} \\ &= 0.6745 \times \frac{0.9948}{2 \times 1.414} \\ &= 0.2372.\end{aligned}$$

Example. While calculating the correlation coefficient between variables x and y, the following results are found:

$$\Sigma_{i=1}^{25} x_i = 125, \Sigma_{i=1}^{25} = 100, \Sigma_{i=1}^{25} x_i^2 = 650, \Sigma_{i=1}^{25} y_i^2 = 460 \text{ and } \Sigma_{i=1}^{25} x_i y_i = 508.$$

Later it was found that at the time of checking two pairs of observations (x, y) were copied wrongly as $(6, 14)$ and $(8, 6)$ while the correct values were $(8, 12)$ and $(6, 8)$ respectively. Determine the correlation coefficient between x and y.

Solution: Now

$$\begin{aligned}\text{Corrected } \Sigma x_i &= 125 - (6+8) + (8+6) = 125 \\ \text{Corrected } \Sigma y_i &= 10 - (14+6) + (12+8) = 100 \\ \text{Corrected } \Sigma x_i^2 &= 650 - (6^2+8^2) + (8^2+6^2) = 650 \\ \text{Corrected } \Sigma y_i^2 &= 460 - (14^2+6^2) + (12^2+8^2) = 436 \\ \text{Corrected } \Sigma x_i y_i &= 508 - (6 \times 14 + 8 \times 6) + (8 \times 12 + 6 \times 8) = 520\end{aligned}$$

$$\begin{aligned}\therefore \text{ Corrected Cov}(x, y) &= \frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y} \\ &= \frac{1}{25} \times 520 - \frac{125}{25}\frac{100}{25} = \frac{104}{5} - 20 = \frac{4}{5}.\end{aligned}$$

$$\begin{aligned}\therefore \text{ Corrected } \sigma_x^2 &= \frac{1}{n}\Sigma x_i^2 - (\bar{x})^2 \\ &= \frac{1}{25} \times 650 - \left(\frac{125}{25}\right)^2 = 26 - 25 = 1.\end{aligned}$$

and

$$\begin{aligned}\therefore \text{ Corrected } \sigma_y^2 &= \frac{1}{n}\Sigma y_i^2 - (\bar{y})^2 \\ &= \frac{1}{25} \times 436 - \left(\frac{100}{25}\right)^2 = \frac{436}{25} - 16 = \frac{36}{25}.\end{aligned}$$

$\therefore$ Corrected Correlation coefficient is

$$r_{xy} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} = \frac{\frac{4}{5}}{\sqrt{1}\sqrt{\frac{36}{25}}} = \frac{4}{6} = \frac{2}{3}.$$

Example. If $\text{var}(x+y) = 81$, $\text{var}(x)$=36 and $\text{var}(y)$=25, then find the correlation coefficient between x and y.

Solution: We know that $\text{var}(x+y) = \text{var}(x) + \text{var}(y) + 2\ \text{Cov}(x,y)$

$$= \text{var}(x) + \text{var}(y) + 2\ r_{xy}\ \sigma_x\sigma_y$$

or, $81 = 36 + 25 + 2.6.5.r_{xy}$

or, $r_{xy} = \frac{81-61}{60} = \frac{20}{60} = \frac{1}{3}$.

Example. If $\Sigma xy = 60$, $\sigma_y = 2.5, \Sigma x^2 = 90$ and $r_{xy} = 0.8$, them find the number of items where $\Sigma x = \Sigma y = 0$.

Solution: Now, $\text{Cov}(x,y) = \frac{1}{n}\Sigma(x-\bar{)}(y-\bar{y}) = \frac{1}{n}\Sigma xy = \frac{60}{n}$, where n is the number of terms and

$$\sigma_x^2 = \frac{1}{n}\Sigma x^2 - (\bar{x})^2 = \frac{90}{n}.$$

$$\text{We know,} \quad r_{xy} = \frac{\text{Cov}(x,y)}{\sigma_x\sigma_y} = \frac{\frac{60}{n}}{\sqrt{\frac{90}{n}} \times 2.5}$$

$$\text{or,} \quad 0.8 = \frac{60}{\sqrt{90} \times \sqrt{n} \times 2.5}$$

$$\text{or,} \quad \sqrt{90} \times \sqrt{n} \times 2 = 60$$

$$\text{or,} \quad 90n = 30 \times 30$$

$$\text{or,} \quad n = 10.$$

Example. If $u - 7x = 5$ and $v - 5y = 11$ and the correlation coefficient of x and y is 0.23, then find the correlation coefficient of u and v.

Solution: From the given relations, we get $u = 7x + 5$ and $v = 5y + 11$ which are linear functions of x and y then $r_{xy} = r_{uv}$, since the coefficients of x and y have same sign.

$$\therefore r_{uv} = 0.23.$$

Example. Two variables x and y have n pair of values. The variance of x, y and $x - y$ are given by σ_x^2, σ_y^2 and σ_{x-y}^2 respectively. Prove that correlation coefficient r_{xy} between x and y is given by

$$r_{xy} = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\ \sigma_x\sigma_y}.$$

Solution: Let $u_i = x_i - y_i, \quad i = 1, 2, ..., n$, then $\bar{u} = \bar{x} - \bar{y}$

and

$$\begin{aligned}
\sigma_u^2 &= \frac{1}{n}\Sigma(u_i-\bar{u})^2 = \frac{1}{n}\Sigma[(x_i-y_i)-(\bar{x}-\bar{y})]^2 \\
&= \frac{1}{n}\Sigma[(x_i-\bar{x})-(y_i-\bar{y})]^2 \\
&= \frac{1}{n}\Sigma[(x_i-\bar{x})^2+(y_i-\bar{y})^2-2(x_i-\bar{x})(y_i-\bar{y})] \\
&= \frac{1}{n}\Sigma(x_i-\bar{x})^2+\frac{1}{n}\Sigma(y_i-\bar{y})^2-2\frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y}) \\
&= \sigma_x^2+\sigma_y^2-2\ \mathrm{Cov}(x,y) \\
\text{or,}\quad \sigma_{x-y}^2 &= \sigma_x^2+\sigma_y^2-2r_{xy}\sigma_x\sigma_y \\
\text{or,}\quad 2r_{xy}\sigma_x\sigma_y &= \sigma_x^2+\sigma_y^2-\sigma_{x-y}^2 \\
\text{or,}\quad r_{xy} &= \frac{\sigma_x^2+\sigma_y^2-\sigma_{x-y}^2}{2\ \sigma_x\sigma_y}.
\end{aligned}$$

Regression Analysis

The word *regression* refers to the method of finding the most suitable equation for *predicting* or *estimating* one variable for a given value of other. It also refers to the method of finding the *error* in such prediction.

Let us suppose that the variables are x and y where x is independent and y is depends on x.

Linear regression: If the dependence can be expressed in the form $y = a + bx$, then the regression that is studied is known as *linear regression*, because the above equation represents straight line.

Curvillinear regression: If the dependence is given by an equation representing a curve then the regression is known as curvillinear regression, *e.g.*, $y = ax^2 + bx + c$. We shall discuss here linear regression only.

Normal equation: The equations

$$\begin{aligned}
\Sigma y &= na + b\Sigma x \\
\Sigma xy &= a\Sigma x + b\Sigma x^2.
\end{aligned}$$

are called *normal equation* for the regression equation $y = a + bx$.

Regression equation of y on x

The regression equation of y on x is the equation of the best fitting straight line in the form $y = a + bx$, obtained by the method of least square.
Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be a set of n pair of observations and let us fit a straight line in the form

$$y = a + bx$$

to these data . Applying method of last squares, the constants a and b are obtained by solving the normal equations, *i.e.,*

$$\begin{aligned} \Sigma y = &\quad na + b\Sigma x \\ \Sigma xy = &\quad a\Sigma x + b\Sigma x^2. \end{aligned} \tag{1.21}$$

We now solve the normal equations in a and b. Multiplying Σx with first equation of (1.21) and second equation by n and then subtracting, we get

$$\begin{aligned} \Sigma x\Sigma y - n\Sigma xy &= b(\Sigma x)^2 - nb\Sigma x^2 \\ \text{or,} \qquad b &= \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{\frac{1}{n}\Sigma xy - \frac{\Sigma x}{n}\frac{\Sigma y}{n}}{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} \\ &= \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\frac{1}{n}\Sigma x^2 - (\bar{x})^2} \\ &= \frac{\mu_{11}}{\sigma_x^2}. \end{aligned}$$

where $\mu_{11} = \text{Cov}(x, y) = \frac{1}{n}\Sigma xy - \bar{x}\bar{y}$.

Putting the value of b in (1.21), we obtain

$$\Sigma y = na + \frac{\mu_{11}}{\sigma_x^2}\,\Sigma x \tag{1.22}$$

$$\text{or,} \quad \frac{1}{n}\Sigma y = a + \frac{\mu_{11}}{\sigma_x^2}\frac{\Sigma x}{n} \tag{1.23}$$

$$\text{or,} \quad \bar{y} = a + \frac{\mu_{11}}{\sigma_x^2}\bar{x} \tag{1.24}$$

$$\text{or,} \quad a = \bar{y} - \frac{\mu_{11}}{\sigma_x^2}\bar{x} \tag{1.25}$$

Substituting these values a and b in the regression equation,

$$y = \left(\bar{y} - \frac{\mu_{11}}{\sigma_x^2}\bar{x}\right) + \frac{\mu_{11}}{\sigma_x^2}x$$

$$\text{or,} \quad (y - \bar{y}) = \frac{\mu_{11}}{\sigma_x^2}(x - \bar{x})$$

which is the equation of the line of regression of y on x.

Regression coefficient of y on x

The coefficient b i.e., $\frac{\mu_{11}}{\sigma_x^2}$ or $\frac{\text{Cov}(x,y)}{\sigma_x^2}$ is called the *regression coefficient of y on x* and is denoted by b_{yx}.

The regression equation of y on x is, therefore, written as

$$y - \bar{y} = b_{yx}(x - \bar{x}).$$

Regression equation of x on y

The best fitting straight line of bivariate distribution representing a regression equation of the form

$$x = c + dy$$

where y is the independent variable and x is the dependent variable, known as the *line of regression of x on y.*

Proceeding exactly in the same manner as before, we obtain the *regression equation of x on y* as

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

where $b_{xy} = \frac{\mu_{11}}{\sigma_y^2}$ or $\frac{\text{Cov}(x,y)}{\sigma_y^2}$ which is known as the *regression coefficient of x on y.*

Properties of regression coefficient

Property 1. Regression coefficients are unaffected by the change of origin.

Proof. Let $u_i = x_i - a$ and $v_i = y_i - b$.

$$\begin{aligned} \text{Now} \qquad b_{yx} &= \frac{\text{Cov}(x,y)}{\sigma_x^2} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \\ &= \frac{\Sigma\{(u_i + a) - (\bar{u} + a)\}\{(v_i + b) - (\bar{v} + b)\}}{\Sigma[(u_i + a) - (\bar{u} + a)]^2} \\ &= \frac{\Sigma(u_i - \bar{u})(v_i - \bar{v})}{\Sigma(u_i - \bar{u})^2} = \frac{\text{Cov}(u,v)}{\sigma_u^2} = b_{uv}. \end{aligned}$$

which is the regression coefficient of v on u. It can be similarly proved that $b_{xy} = b_{uv}$.

Property 2. Regression coefficient is affected by the change of scale.

Proof. Let $u_i = \frac{x_i - a}{c}$ and $v_i = \frac{y_i - b}{d}$.

$$\begin{aligned} \therefore \qquad & x_i = a + cu_i \quad \text{and} \quad y_i = b + dv_i \\ \therefore \qquad & \bar{x} = a + c\bar{u} \quad \text{and} \quad \bar{y} = b + d\bar{v} \end{aligned}$$

$$\begin{aligned} \text{Hence} \quad b_{yx} &= \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \\ &= \frac{\Sigma\{(a + cu_i) - (a + c\bar{u})\}\{(b + dv_i) - (b + d\bar{v})\}}{\Sigma[(a + cu_i) - (a + c\bar{u})]^2} \\ &= \frac{c.d\ \Sigma(u_i - \bar{u})(v_i - \bar{v})}{c^2\Sigma(u_i - \bar{u})^2} \\ &= \frac{d}{c}\, b_{vu}. \end{aligned}$$

It can be similarly be shown that $b_{xy} = \frac{c}{d}\, b_{uv}$.

Relation between regression coefficient and between regression coefficient and correlation coefficient

1. It is known that $r = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}$ where r is the correlation coefficient

$$\text{and} \quad b_{yx} = \frac{\text{Cov}(x,y)}{\sigma_x^2} = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} \frac{\sigma_y}{\sigma_x}$$
$$= r \frac{\sigma_y}{\sigma_x}.$$

Similarly, $b_{xy} = r\frac{\sigma_x}{\sigma_y}$. Hence $b_{yx} b_{xy} = r^2$.

In other words, r is the geometric mean of the regression coefficients.

2. Both the regression coefficients must have the same algebraic signs. If b_{yx} and b_{xy} are positive the r is positive and if b_{yx} and b_{xy} are negative, then r is negative.

3. Since $-1 \leq r \leq 1$, both the regression coefficients cannot be greater than 1.

4. Arithmetic mean of two regression coefficients is either equal to or greater than the correlation coefficient

$$\text{i.e.,} \quad \frac{b_{yx} + b_{xy}}{2} \geq r.$$

5. The regression lines are usually different. But since they always pass through $(\bar{x}, \bar{y})$, therefore, they become identical if their slopes become equal, *i.e.,* if $b_{yx} = \frac{1}{b_{xy}}$ or if $b_{xy} b_{yx} = 1$. In such a case

$$\left(r\frac{\sigma_y}{\sigma_x}\right)\left(r\frac{\sigma_x}{\sigma_y}\right) = 1 \implies r^2 = 1 \implies r = \pm 1.$$

6. If $r = +1$, then both regression equation take the form

$$(y - \bar{y}) = \frac{\sigma_y}{\sigma_x}(x - \bar{x}).$$

7. If $r = -1$, then both regression equation take the form

$$(y - \bar{y}) = -\frac{\sigma_y}{\sigma_x}(x - \bar{x}).$$

8. Correlation is said to be of high degree if $\frac{3}{4} \leq |r| \leq 1$, of moderate degree if $\frac{1}{4} \leq |r| < \frac{3}{4}$ and of low degree if $0 \leq |r| < \frac{1}{4}$.

9. The acute angle between two regression line is given by

$$tan\theta = \left|\frac{1 - r^2}{b_{xy} + b_{yx}}\right|.$$

$\therefore$ Two lines coincide, iff $\theta = 0$. *i.e.,* iff $r = \pm 1$.

Example. Given the following bivariate data:

$x:$	1	5	3	2	1	1	7	3
$y:$	6	1	0	0	1	2	1	5

Fit the regression line of y on x and that of x on y. Predict y when $x = 10$ and x when $y = 2.5$.

Solution: we are to find the equations

$$\begin{aligned} y - \bar{y} &= b_{yx}(x - \bar{x}) \\ x - \bar{x} &= b_{xy}(y - \bar{y}) \end{aligned}$$

We shall assume $u = x - 3$ and $v = y - 3$ and use the formula

$$b_{yx} = b_{vu} = \frac{n\Sigma uv - \Sigma u \Sigma v}{n\Sigma u^2 - (\Sigma u)^2}$$

$$\text{and} \quad b_{xy} = b_{uv} = \frac{n\Sigma uv - \Sigma u \Sigma v}{n\Sigma v^2 - (\Sigma v)^2}$$

x	y	u	v	u^2	v^2	uv
1	6	-2 ‘	-3	4	9	-6
5	1	2	-2	4	4	-4
3	0	-0	-3	0	9	0
2	0	-1	-3	1	9	3
1	1	-2	-2	4	4	4
1	2	-2	-1	4	1	2
7	1	4	-2	16	4	-8
3	5	0	2	0	4	0
		$\Sigma u = -1$	$\Sigma v = -8$	$\Sigma u^2 = 33$	$\Sigma v^2 = 44$	$\Sigma uv = -9$

$$\therefore \quad \bar{x} = \bar{u} + 3 = -\frac{1}{8} + 3 = 2.875 \tag{1.26}$$

$$\bar{y} = \bar{v} + 3 = -\frac{8}{8} + 3 = 2 \tag{1.27}$$

$$b_{yx} = \frac{8 \times (-9) - (-1) \times (-8)}{8 \times 33 - (-1^2)} = \frac{-72 - 8}{264 - 1} = -0.304 \tag{1.28}$$

$$b_{xy} = \frac{8 \times (-9) - (-1) \times (-8)}{8 \times 44 - (-8^2)} = \frac{-80}{352 - 64} = -0.278. \tag{1.29}$$

The regression line of y on x is

$$(y-2) = -0.304\ (x-2.875),\ \text{or}\ y = -0.304x + 2.874.$$

Value of y when $x = 10$ is $y = -0.304 \times 10 + 2.874 = -0.166$.

The regression line of x on y is

$$(x-2.875) = -0.278(y-2),\quad \text{or}\quad x = -0.278x + 3.431.$$

Value of x when $y = 2.5$ is $x = -0.278 \times 2.5 + 3.431 = 2.736$.

Example. Find the equation of regression line x on y for the following bivariate data:

x :	1	1.5	2	2.5	3	3.5	4
y :	5.3	5.7	6.3	7.2	8.2	8.7	8.4

Solution: For simplifying the calculations, let us make a change of origin and scale for both the variables as follows:

$$u = \frac{x-2.5}{0.5},\ v = \frac{y-7.0}{0.1}.$$

x	y	u	v	v^2	uv
1	5.3	-3 ‘	-17	289	51
1.5	5.7	-2	-13	169	26
2	6.3	-1	-7	49	7
2.5	7.2	0	2	4	0
3	8.2	1	12	144	12
3.5	8.7	2	17	289	34
4	8.4	3	14	196	42
17.5	49.8	$\Sigma u = 0$	$\Sigma v = 8$	$\Sigma u^2 = 1140$	$\Sigma uv = 172$

We know that

$$\begin{aligned} b_{xy} &= \frac{c}{d}\,\frac{n\Sigma uv - \Sigma u \Sigma v}{n\Sigma v^2 - (\Sigma v)^2} \quad \text{where}\ \ c = 0.5\ \ \text{and}\ \ d = 0.1 \\ &= \frac{0.5}{0.1} \times \frac{172 \times 7 - 0 \times 8}{7 \times 1140 - (8)^2} = \frac{5 \times 1204}{7916} = 0.76. \\ \bar{x} &= \frac{17.5}{7} = 2.5 \ \text{and}\ \ \bar{y} = \frac{49.8}{7} = 7.11. \end{aligned}$$

$\therefore$ The regression line of x on y is

$$\begin{aligned} x - 2.5 &= 0.76\ (y - 7.11) \\ x &= 0.76y - 2.90. \end{aligned}$$

Example. Let the line of regression concerning two variables x and y be given by $y = 32 - x$ and $x = 13 - 0.25y$. Obtain the values of the means and correlation coefficient.

Solution: Since the regression lines intersects at $(\bar{x}, \bar{y})$, the means will be obtained by solving the two equations. Solving $y = 32 - x$ and $x = 13 - 0.25y$, we get $x = 6.7$ and $y = 25.3$. So $\bar{x} = 6.7$ and $\bar{y} = 25.3$.

Now $y = 32 - x$ is the regression equation of y on x,

$$\therefore \quad b_{yx} = -1$$

and $x = 13 - 0.25y$ being the regression equation of x on y,

$$\therefore \quad b_{xy} = -0.25$$

$$\therefore \quad r^2 = b_{yx} \times b_{xy} = (-1) \times (-0.25) = 0.25$$

$$\therefore \quad r = \pm\sqrt{0.25} = \pm 0.5.$$

Bur, since both regression coefficients are negative (note that both must have same sign), the correlation coefficient must be negative, *i.e.,* $r = -0.5$.

Example. For the variables x and y, the equations of the regression lines are $4x - 5y + 33 = 0$ and $20x - 9y = 107$. Identify the regression line of y on x and that of x on y. What is the correlation coefficient? If the variance of x is 9 find the standard deviation of y. Also find $\bar{x}, \bar{y}$. What is the estimate value of y at $x = 10$? If this estimate be y_0, find the estimated value of x when $y = y_0$.

Solution: Let the regression line of y on x be the $4x - 5y + 33 = 0$, then

$$5y = 4x + 33 \quad \text{or} \quad y = \frac{4}{5}x + \frac{33}{5}.$$

$\therefore$ The regression coefficient of y on x is given by $b_{yx} = \frac{4}{5}$.

Let the regression line of x on y be the $20x - 9y = 107$, then

$$20x = 9x + 107 \quad \text{or} \quad x = \frac{9}{20}y + \frac{107}{20}.$$

$\therefore$ The regression coefficient of x on y is given by $b_{xy} = \frac{9}{20}$.

$$\therefore \quad r^2 = b_{yx} \times b_{xy} = \frac{4}{5} \times \frac{9}{20} = \frac{9}{25}$$

$$\therefore r = \pm\ \frac{3}{5} = \pm\ 0.6.$$

Since $b_{xy} and b_{yx}$ are both positive, then $r = 0.6$. So our hypothesis is correct.

Hence the regression line of y on x and of x on y are given by

$$4x - 5y + 33 = 0 \text{ and } 20x - 9y = 107, \text{ respectively.}$$

Again the variance of $x = 9$. So $\sigma_x = 3$.

$$\text{Now,} \quad b_{yx} = r\frac{\sigma_y}{\sigma_x} \tag{1.30}$$

$$\text{or,} \quad \frac{4}{5} = \frac{3}{5}\frac{\sigma_y}{\sigma_3} \tag{1.31}$$

$$\text{or,} \quad \sigma_y = 4. \tag{1.32}$$

$\therefore$ The standard deviation of y is 4.

We know that the two regression line s intersects at the point $(\bar{x}, \bar{y})$, where $\bar{x}$ and $\bar{y}$ are the mean of x and y respectively.

$\therefore \quad 4\bar{x} = 5\bar{y} + 33 = 0$ and $20\bar{x} - 9\bar{y} - 107 = 0$. Solving, we get $\bar{x} = 13$ and $\bar{y} = 17$.

Also when $x = 10$, $y_0 = \frac{4}{5}x + 6.6 = \frac{4}{5} \times 10 + 6.6 = 8 + 6.6 = 14.6$.

For $y = y_0 = 14.6$, $x_0 = \frac{9}{20}y + \frac{107}{20} = \frac{9}{20} \times 14.6 + \frac{107}{20} = \frac{238.4}{20} = 11.92$.

Example. If $x = 4y + 5$ and $y = Kx + 4$ be two regression lines of x on y and of y on x respectively, find the interval in which K lies.

Solution: Since $x = 4y + 5$ and $y = Kx + 4$ be two regression lines of x on y and of y on x respectively, then the regression coefficients of x on y and y on x are given by

$$b_{xy} = 4 \text{ and } b_{yx} = K$$

Since

$$r_{xy}^2 = b_{xy}\, b_{yx} \quad \therefore \quad r_{xy}^2 = 4K.$$

As $-1 \leq r_{xy} \leq 1$, $0 \leq r_{xy}^2 \leq 1 \qquad \therefore 0 \leq 4K \leq 1$, or $0 \leq K \leq \frac{1}{4}$.

Example. The relationship between travel expenses (y) and the duration of travel (x) is found to be linear. A summary of data for 102 pairs is given below:

$\Sigma x = 510$, $\Sigma y = 7140, \Sigma x^2 = 4150, \Sigma xy = 54900$ and $\Sigma y^2 = 7, 40, 200$.

1. Find the two regression coefficients.
2. Find the two regression line.
3. A given trip has to take seven days. How much money should a salesman be allowed so that he will not run short of money?

Solution: Here $\bar{x} = \frac{1}{n}\Sigma x = \frac{510}{102} = 5$ where $n = 102$ and $\bar{y} = \frac{1}{n}\Sigma y = \frac{7140}{102} = 70$.

$$\begin{aligned}
\text{Cov } (x, y) &= \frac{1}{n}\Sigma xy - \bar{x}\bar{y} = \frac{54900}{102} - 5 \times 70 = \frac{9150}{17} - 350 = 188.24 \\
\sigma_x^2 &= \frac{1}{n}\Sigma x^2 - (\bar{x})^2 = \frac{4150}{102} - 25 = \frac{2075}{51} - 25 = 15.686 \\
\sigma_y^2 &= \frac{1}{n}\Sigma y^2 - (\bar{y})^2 = \frac{740200}{12} - 4900 = 2356.863
\end{aligned}$$

1.

$$b_{xy} = r_{xy}\,\frac{\sigma_x}{\sigma_y} = \frac{\sigma_x}{\sigma_y}\,\frac{\text{Cov}(x, y)}{\sigma_x\sigma_y} = \frac{\text{Cov}(x, y)}{\sigma_y^2} = \frac{188.24}{2356.863} = 0.08.$$

and

$$b_{yx} = \frac{\text{Cov}(x, y)}{\sigma_x^2} = \frac{188.24}{15.686} = 12.$$

2. The regression line of y on x is $y - \bar{y} = b_{yx}\ (x - \bar{x})$ or $y - 70 = 12(x - 5)$ or $y = 12x + 10$.

   The regression line of x on y is $x - \bar{x} = b_{xy}\ (y - \bar{y})$ or $x - 5 = 0.08(y - 70)$ or $x = 0.08y - 0.6$.

3. For $x = 7$, $y = 12x + 10 = 12 \times 7 + 10 = 94$.

Example. If $\text{var}(x) = 4,\ \text{var}(y) = 9$ and $r_{xy} = \frac{2}{3}$, then find $\text{var}(2x - 3y)$.

Solution: Now

$$\begin{aligned}
\text{var}(2x - 3y) &= \text{var}(2x) + \text{var}(3y) - 2\sqrt{\text{var}(2x)}\sqrt{\text{var}(3y)}\ r_{2x,3y} \\
&= 2^2\text{var}(x) + 3^2\text{var}(y) - 2\sqrt{2^2\text{var}(x)}\sqrt{3^2\text{var}(y)}\ r_{xy} \\
&= 4 \times 4 + 9 \times 9 - 2 \times 2 \times 3 \times \sqrt{4} \times \sqrt{9} \times \frac{2}{3} \\
&= 16 + 81 - 48 = 49.
\end{aligned}$$

Example. If x and y are two correlated variables with same variance and the correlation coefficient is r, find the regression coefficient of x on $(x+y)$ and that of $(x+y)$ on x. Hence find the correlation coefficient between x and $(x + y)$.

Solution: Let $\text{var}(x) = \text{var}(y) = \sigma^2$.

We know that $r = \frac{\text{Cov}(x,y)}{\sigma_x\sigma_y} = \frac{\text{Cov}(x,y)}{\sigma\sigma} = \frac{\text{Cov}(x,y)}{\sigma^2}$.

$\therefore \text{Cov}(x, y) = r\sigma^2$.

Let $u = x + y$, then $b_{xu} = \frac{\text{Cov}(x,u)}{\text{var}(u)}$ and $b_{ux} = \frac{\text{Cov}(x,u)}{\text{var}(x)}$.

$$\begin{aligned}
\mathrm{Cov}(x,u) &= \mathrm{Cov}(x,x+y) = \frac{1}{n}\Sigma(x-\bar{x})(x+y-\bar{x}-\bar{y}) && (1.33)\\
&= \frac{1}{n}\Sigma(x-\bar{x})[(x-\bar{x})+(y-\bar{y})] && (1.34)\\
&= \frac{1}{n}\Sigma(x-\bar{x})^2 + \frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y}) && (1.35)\\
&= \mathrm{var}(x) + \mathrm{Cov}(x,y) = \sigma^2 + r\sigma^2 = (1+r)\sigma^2 && (1.36)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{var}(u) &= \mathrm{var}(x+y) = \mathrm{var}(x) + \mathrm{var}(y) + 2\ \mathrm{Cov}(x,y) && (1.37)\\
&= \sigma^2 + \sigma^2 + 2r\sigma^2 && (1.38)\\
&= 2(1+r)\sigma^2 && (1.39)
\end{aligned}$$

$$\therefore b_{xu} = \frac{\mathrm{Cov}(x,u)}{\mathrm{var}(u)} = \frac{(1+r)\sigma^2}{2(1+r)\sigma^2} = \frac{1}{2}$$

$$b_{ux} = \frac{\mathrm{Cov}(x,u)}{\mathrm{var}(x)} = \frac{(1+r)\sigma^2}{\sigma^2} = 1+r$$

and the correlation coefficient between x and u is given by

$$r_{xu} = \sqrt{b_{xu} \times b_{ux}} = \sqrt{\frac{1}{2}(1+r)} = \sqrt{\frac{1+r}{2}}.$$

Example. For two variables x and y, the two regression lines are $x+4y+3=0$ and $4x+9y+5=0$. Identify which one is of y on x. Find the means of x and y. Find the correlation coefficient between x and y. Estimate the value of x when $y = 1.5$.

Solution: Let $x+4y+3=0$ be the regression line of y on x. Then $4x+9y+5=0$ must be the regression line x on y. So if b_{yx} and b_{xy} denote the respective regression coefficients, then we get

$$b_{yx} = -\frac{1}{4} \ \text{ and } \ b_{xy} = -\frac{9}{4}.$$

$$\therefore r^2 = b_{yx} \times b_{xy} = \frac{9}{16}.$$

Since $0 \le r^2 \le 1$, our assumption is correct, *i.e.,* $x+4y+3=0$ be the regression line of y on x.

Now, $\quad r^2 = \frac{9}{16}$ which gives $r = \pm\frac{3}{4}$.

Since b_{xy} and b_{yx} are both negative, the $r = -\frac{3}{4}$.

Solving the two equations $x+4y+3=0$ and $4x+9y+5=0$, we get

$$x = 1, y = -1.$$

$\therefore \bar{x} = 1$ and $\bar{y} = -1$.

For estimate x when $y = -1.5$, we take the regression line of x on y and putting $y = 1.5$, we get $4x + 9.5 = -5 \implies x = -\frac{18.5}{4} = -4.625$.

Therefore, the estimated value of x is -4.625.

Example. Let (x, y) and (u, v) be two bivariate variables such that $2u = x + 9$ and $3v = 2y + 7$. The regression coefficient of x on y is σ. Then find the regression coefficient of u on v.

Solution: Here $2u = x + 9$ and $3v = 2y + 7$.

$$\therefore u = \frac{x}{2} + \frac{9}{2} \text{ and } v = \frac{2}{3}y + \frac{7}{3}.$$

Now,

$$\begin{aligned}
\bar{u} &= \frac{\bar{x}}{2} + \frac{9}{2} \text{ and } \bar{v} = \frac{2}{3}\bar{y} + \frac{7}{3}. \\
\therefore \sigma_u &= \frac{1}{2}\sigma_x \text{ and } \sigma_v = \frac{2}{3}\sigma_y. \\
\therefore r_{uv} &= \frac{\frac{1}{2}\frac{2}{3}}{\left|\frac{1}{2}\right|\left|\frac{2}{3}\right|} r_{xy} \\
\therefore b_{uv} &= \frac{\sigma_u}{\sigma_v} r_{uv} = \frac{\frac{1}{2}\sigma_x}{\frac{2}{3}\sigma_y} r_{xy} = \frac{3}{4}\frac{\sigma_x}{\sigma_y} r_{xy} = \frac{3}{4} \times \sigma = \frac{3}{4}\sigma.
\end{aligned}$$

Example. The variates x and y are normally correlated and u, v are defined by

$$\begin{aligned}
u &= x \cos\alpha + y \sin \alpha \\
v &= y \cos\alpha - x \sin \alpha
\end{aligned}$$

Show that u and v will be correlated if

$$\tan 2\alpha = \frac{2r\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}$$

where r is the correlation coefficient between x and y.

Further show that in the case

$$\sigma_u^2 + \sigma_v^2 = \sigma_x^2 + \sigma_y^2.$$

Solution: Now,

$$\begin{aligned}
\text{Cov}(x,y) &= \frac{1}{n}\Sigma(u_i-\bar{u})(v_i-\bar{v}) \\
&= \frac{1}{n}\Sigma(x_i\cos\alpha + y_i\sin\alpha - \bar{x}\cos\alpha - \bar{y}\sin\alpha) \\
&\quad (y_i\cos\alpha - x_i\sin\alpha - \bar{y}\cos\alpha + \bar{x}\sin\alpha) \\
&= \frac{1}{n}\Sigma[(x_i-\bar{x})\cos\alpha + (y_i-\bar{y})\sin\alpha] \\
&\quad [(y_i-\bar{y})\cos\alpha - (x_i-\bar{x})\sin\alpha] \\
&= (\cos^2\alpha - \sin^2\alpha)\frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y}) \\
&\quad -\cos\alpha\sin\alpha\left[\frac{1}{n}\Sigma(x_i-\bar{x})^2 - \frac{1}{n}\Sigma(y_y-\bar{y})^2\right] \\
&= \cos 2\alpha\ \text{Cov}(x,y) - \frac{1}{2}\sin 2\alpha\ [\sigma_x^2-\sigma_y^2]
\end{aligned}$$

Now u and v will be uncorrelated if $\text{Cov}(u,v)=0$, *i.e.,* if

$$\cos 2\alpha\ \text{Cov}(x,y) - \frac{1}{2}\sin 2\alpha\ [\sigma_x^2-\sigma_y^2] = 0$$

$$\textit{i.e.,}\ \text{if}\ \ \tan 2\alpha = \frac{2r\sigma_x\sigma_y}{\sigma_x^2-\sigma_y^2}.$$

Further

$$\begin{aligned}
\sigma_u^2+\sigma_v^2 &= \frac{1}{n}\Sigma(u_i-\bar{u})^2 + \frac{1}{n}\Sigma(v_i-\bar{v})^2 \\
&= \frac{1}{n}\Sigma(x_i\cos\alpha + y_i\sin\alpha - \bar{x}\cos\alpha - \bar{y}\sin\alpha)^2 \\
&\quad +\frac{1}{n}\Sigma(y_i\cos\alpha - x_i\sin\alpha - \bar{y}\cos\alpha + \bar{x}\sin\alpha)^2 \\
&= \frac{1}{n}\Sigma[(x_i-\bar{x})\cos\alpha + (y_i-\bar{y})\sin\alpha]^2 \\
&\quad +\frac{1}{n}\Sigma[(y_i-\bar{y})\cos\alpha - (x_i-\bar{x})\sin\alpha]^2 \\
&= \cos^2\alpha\ \frac{1}{n}\Sigma(x_i-\bar{x})^2 + \sin^2\alpha\ \frac{1}{n}\Sigma(y_i-\bar{y})^2 \\
&\quad +2\sin\alpha\cos\alpha\frac{1}{n}\ \Sigma(x_i-\bar{x})(y_i-\bar{y}) \\
&\quad +\sin^2\alpha\ \frac{1}{n}\Sigma(x_i-\bar{x})^2 + \cos^2\alpha\ \frac{1}{n}\Sigma(y_i-\bar{y})^2 \\
&\quad -2\sin\alpha\cos\alpha\frac{1}{n}\ \Sigma(x_i-\bar{x})(y_i-\bar{y}) \\
&= \sigma_x^2\cos^2\alpha + \sigma_y^2\sin^2\alpha + \sigma_x^2\sin^2\alpha + \sigma_y^2\cos^2\alpha \\
&= \sigma_x^2(\cos^2\alpha + \sin^2\alpha) + \sigma_y^2(\cos^2\alpha + \sin^2\alpha) \\
&= \sigma_x^2+\sigma_y^2.
\end{aligned}$$

Example. If θ be the acute angle between two regression lines of the variables x nd y, prove that

$$\tan\theta = \frac{1-r_{xy}^2}{r_{xy}}.\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$$

where r_{xy} is the correlation coefficient between x and y.

Solution: The regression lines are

$$\begin{aligned} y-\bar{y} &= r_{xy}\,\frac{\sigma_y}{\sigma_x}\,(x-\bar{x}) \\ x-\bar{x} &= r_{xy}\,\frac{\sigma_x}{\sigma_y}\,(y-\bar{y}) \end{aligned} \tag{1.40}$$

Let m_1 and m_2 be the slopes of the lines of (1.40) and (1.40). Then

$$m_1 = r_{xy}\,\frac{\sigma_y}{\sigma_x} \quad \text{and} \quad m_2 = \frac{\sigma_y}{r_{xy}\,\sigma_x}.$$

Now,

$$\begin{aligned} \tan\theta &= \frac{m_2 \sim m_1}{1+m_1m_2} = \frac{\frac{\sigma_y}{r_{xy}\ \sigma_x} - r_{xy}\,\frac{\sigma_y}{\sigma_x}}{1+\frac{\sigma_y}{r_{xy}\ \sigma_x}r_{xy}\,\frac{\sigma_y}{\sigma_x}} \\ &= \frac{\frac{\sigma_y}{\sigma_x}\left(\frac{1}{r_{xy}}-r_{xy}\right)}{1+\frac{\sigma_y^2}{\sigma_x^2}} = \frac{1-r_{xy}^2}{r_{xy}}.\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2} \end{aligned}$$

Exercise.

1. Find the correlation coefficient of the following data:

$x:$	1	3	4	6	8	9	11	14
$y:$	1	2	4	4	5	7	8	9

2. Find the covariance and correlation coefficient of the two variables x and y of the following data:

$x:$	50	53	55	57	60	56	62	52
$y:$	53	55	57	60	56	52	64	54

3. The bivariate data (x, y) has the following results:
$\Sigma x = 200, \Sigma y = 250, \Sigma x^2 = 2000, \Sigma y^2 = 2900, \Sigma xy = 2250, n = 25$. Find the correlation coefficient between x and y.

4. If var$(x+y)$=45, var(x)=9 and var(y)=16, then find Cov(x,y).

5. Calculate the correlation coefficient from the following data:
   $n=10, \Sigma x=100, \Sigma y=150, \Sigma(x-10)^2=180, \Sigma(y-15)^2=215$ and $\Sigma(x-10)(y-15)=60$.

6. Find the regression lines of the following data:

x :	60	65	72	64	63	75	77	70
y :	45	48	44	47	51	52	54	50

7. Marks of 5 students in mathematics and statistics are given:

Mathematics :	38	48	43	40	41
Statistics :	31	38	43	33	35

   Find the regression lines when marks of a student in Mathematics is 42, determine the most likely marks in statistics.

8. If x and y are uncorrelated variables and their standard deviations are 3 and 4 respectively. Find the correlation coefficient between $5x+2y$ and $2x-5y$.

9. if (x,y) and (u,v) be the bivariate variables such that $4u=2x+7$ and $6v=2y-15$ and if the regression coefficient of y on x is 3, then find the regression coefficient of v on u.

10. Find the regression lines from the following data:
    $\bar{x}=90, \bar{y}=70, n=10, \Sigma x^2=6360, \Sigma y^2=2860, \Sigma xy=3900$.

11. The regression equation of y on x and x on Y are given by $2x+3y=26$ and $6x+y=31$, respectively. Find the regression coefficient b_{yx} and b_{xy}.

12. For the variables x and y, the equation of regression lines on $3x+12y=19$ and $3y+9x=46$. Identify the regression lines of y on x and x on y. Find the correlation coefficient and ration of standard deviation of x and y. Find the mean of x and y.

13. If $5y-7x=11$ be the regression line of y on x, variance of x is 25 and correlation coefficient between x and y is $\frac{1}{7}$, then find the variance of y.

14. If $4x=3y+11$ and $3=5x+7$ be the two regression lines of y on x and x on y respectively, find the interval in which K lies.

15. If σ_x and σ_y are the standard deviations of two uncorrelated variables x and y, prove that the standard deviation of $ax+by$ is $\sqrt{a^2\sigma_x^2+b^2\sigma_y^2}$.

16. Show that $2x+3y$ and $4x+9y$ are uncorrelated if

$$8\sigma_x^2+30r\sigma_x\sigma_y+27\sigma_y^2=0.$$

17. The regression lines of y on x and x on y are given by $x+3y=0, 3x+2y=0$. If $\sigma_x = 1$, then find the regression line of v on u where $u = x+y$ and $v = x-y$.

18. If a, b and c are positive constants, show that the correlation coefficient between $ax+by$ and cy is
$$\frac{ar\sigma_x + b\sigma_y}{\sqrt{a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\sigma_x\sigma_y}}$$

Answer: 1. 0.977 2. $10.85, 0.752$ 3. 0.625 4. $\text{cov}(x,y)$=10. 5. 0.305 6.$y-49.5 = 0.34(x-68.25), x-68.25 = 1.56(y-49.5)$ 7. $y = 0.79x - 2.82, x = 0.52y + 23.28$; 36 8. -0.2 9. 2 10. $x = 0.13y + 80.9, y = 106x + 64.6$ 11. $b_y x = -\frac{3}{2}, b_{xy} = \frac{1}{6}$. 12. y on x is $3x+12y=19$ and x on y is $3y+9x=46$, $r_{xy} = -1, \bar{x} = 5, \bar{y} = \frac{1}{3}$. 13. 49 14. $0 \le K \le 4$ 17. $5v - 3u = 0$.

Chebyshev's Inequality

Let X be an arbitrary random variable with mean μ and variance σ^2. What is the probability that X is within t of its average μ? If we knew the exact distribution of and pdf of X, then we could compute this probability $P(|X-\mu| \le t) = P(\mu - t \le X \le \mu + t)$.

But there is another way to find a lower bound for this probability. For instance, we may obtain an expression like $P(|X-\mu| \le 2) \ge 0.60$. That is, there is at least a 60% chance for an obtained measurement of this X to be within 2 of its mean.

Theorem 1.1. *Let X be a random variable with mean μ and variance σ^2. For all $t > 0$*
$$P(|X-\mu| > t) \le \frac{\sigma^2}{t^2} \quad \textit{and} \quad P(|X-\mu| \le t) \ge 1 - \frac{\sigma^2}{t^2}.$$

Proof. Consider
$$Y = \begin{cases} t^2, & \text{if } |X-\mu| > t. \\ 0, & \text{otherwise.} \end{cases} \tag{1.41}$$

□

Observe that $Y \le |X-\mu|^2$. Then
$$t^2 \times P(|X-\mu| > t) = E[Y] \le E[|X-\mu|^2] = \text{ var}(X) = \sigma^2$$
where $E[Y]$ denotes the expectation of Y. Thus
$$P(|X-\mu| > t) \le \frac{\sigma^2}{t^2}.$$
$\therefore -P(|X-\mu| > t) \ge -\frac{\sigma^2}{t^2}$ which gives
$$P(|X-\mu| \le t) = 1 - P(|X-\mu| > t) \ge 1 - \frac{\sigma^2}{t^2}.$$

Note. Chebyshev's Inequality is meaningless when $t \leq \sigma$. For instance, when $t = \sigma$ it is simply saying $P(|X - \mu| > t) \leq 1$ or $P(|X - \mu| \leq t) \geq 0$, which are already obvious. So we must use $t > \sigma$ to apply the inequalities.

Generalized form of Chebyshev's ineqiality

Let $g(X)$ be a non-negative function of random variable X. Then for all $K > 0$,

$$P[g(X) \geq K] \leq \frac{E[g(X)]}{K}.$$

Other forms of Chebyshev's ineqiality

If we put $g(X) = (X - \mu)^2$ and $K = K^2\sigma^2$ in the general form, we obtain

$$\begin{aligned} P[(X-\mu)^2 \geq K^2\sigma^2] &\leq \frac{E[(X-\mu)^2]}{K^2\sigma^2} \\ \text{or,} \quad P[|X-\mu| \geq K\sigma] &\leq \frac{\sigma^2}{K^2\sigma^2} \\ \text{or,} \quad P[|X-\mu| \geq K\sigma] &\leq \frac{1}{K^2}. \end{aligned}$$

Example. (a). Let X is Poisson distributed with parameter $\mu = 9$. Give a lower bound for $P(|X - \mu| \leq 5)$.

(b). Let X be normally distributed with $\mu = 100, \sigma = 15$. Give a lower bound for $P(|X - \mu| \leq 20)$.

Solution: (a) Since X is Poisson distributed with $\mu = 9$, so the mean is $\mu = 9$ and variance $= \sigma^2 = 9$.
Then $P(|X - \mu| \leq 5) = P(|X - 9| \leq 5) \geq 1 - \frac{\sigma^2}{5^2} = 1 - \frac{9}{25} = \frac{16}{25} = 0.64$.

(b) Here mean is $\mu = 100$ and $\sigma = 15$.
$\therefore P(|X - \mu| \leq 20) = P(|X - 100| \leq 20) \geq 1 - \frac{\sigma^2}{20^2} = 1 - \frac{15^2}{20^2} = \frac{175}{400} = 0.4375$.

Note: Using a calculator, we obtain $P(|X - 100| \leq 20) \approx 0.817577$. From these examples, we see that the lower bound provided by Chebyshev's Inequality is not very accurate. However, the inequality is very useful when applied to the sample mean $\bar{x}$ from a large random sample.

Example. A random variable has mean 10 and variance 16. Find the lower bound for $P(5 < X < 15)$.

Solution: By Chebyshev's inequality

$$P[|X - \mu| < K\sigma] \geq 1 - \frac{1}{K^2}$$

$$\text{or} \quad P[\mu - K\sigma < X < \mu + K\sigma] \geq 1 - \frac{1}{K^2}$$

In the present case, $\mu = 10$ and $\sigma = 4$.

$\therefore P[10 - 4K < X < \mu + 4K] \geq 1 - \frac{1}{K^2}$.

Substituting $K = \frac{5}{4}$, we get $P(5 < X < 15) \geq 1 - \frac{1}{\frac{25}{16}} = 1 - \frac{16}{25} = \frac{9}{25}$.

Example. If X a random variable with $E(X) = 3$ and $E(X^2) = 13$, find the lower bound for $P(-2 < X < 8)$ using Chebyshev's inequality.

Solution: We have $\text{var}(X) = E(X^2) - [E(X)]^2 = 13 - 9 = 4$.

By Chebyshev's inequality

$$P(\mu - K\sigma < X < \mu + K\sigma) \geq 1 - \frac{1}{K^2}$$

$$\text{or} \quad P(3 - 2K < X < 3 + 2K) \geq 1 - \frac{1}{K^2}.$$

Putting $K = \frac{5}{2}$, we get

$$P(-2 < X < 8) \geq 1 - \frac{4}{25} = \frac{21}{25}.$$

Example. An unbiased coin is tossed 100 times. Show that the probability that the number of heads will lie between 30 and 70 is greater than 0.93.

Solution: Let X be the number of heads. Then X follows Binomial distribution with mean $np = 100 \times \frac{1}{2} = 50$ and standard deviation$= \sqrt{100 \times \frac{1}{2} \times \frac{1}{2}} = 5$.

By Chebyshev's inequality,

$$P(\mu - K\sigma < X < \mu + K\sigma) \geq 1 - \frac{1}{K^2}$$

$$\text{or} \quad P(50 - 5K < X < 50 + 5K) \geq 1 - \frac{1}{K^2}.$$

Putting $K = 4$, we get

$$P(30 < X < 80) \geq 1 - \frac{1}{16} = \frac{15}{16} = 0.9375.$$

$$\therefore \quad P(30 < X < 80) > 0.93.$$

Example. If a die is thrown 3,600 times, show that the probability that the number of sixes lies between 550 and 650 is at least $\frac{4}{5}$.

Solution: Let X be the number of sixes. Clearly, X follows Binomial distribution with mean $n = 3600$ and $p = \frac{1}{6}$.

So $\mu = E(X) = np = 3600 \times \frac{1}{6} = 600$ and $\sigma^2 = \text{var}(X) = np(1-p) = 3600 \times \frac{1}{6} \times \frac{5}{6} = 500$.

Hence, by Chebyshev's inequality,

$$P(|X - 600| < 50) \geq 1 - \frac{\text{var}(X)}{50^2} = 1 - \frac{500}{50^2} = 1 - \frac{1}{5} = \frac{4}{5}.$$

$$\textit{i.e.,} \quad P(550 < X < 650) \geq \frac{4}{5}.$$

Example. Use Chebyshev's inequality to show that for $n \geq 36$, the probability that in n throws of a fair die the number of sixes lies between $\frac{1}{6}n - \sqrt{n}$ and $\frac{1}{6}n + \sqrt{n}$ is at least $\frac{31}{36}$.

Solution: Let X denote the number of sixes in n throws of a fair die.

Then clearly X is binomial (n, p) variate with $p = \frac{1}{6}$.

$\therefore E(X) = np = \frac{n}{6}$ and $\text{var}(X) = np(1-p) = n.\frac{1}{6}.\frac{5}{6} = \frac{5n}{36}$.

Now, by Chebyshev's inequality,

$$\begin{aligned} P\left(\frac{1}{6}n - \sqrt{n} < X < \frac{1}{6}n + \sqrt{n}\right) &= P\left(\left|X - \frac{n}{6}\right| < \sqrt{n}\right) \\ &= 1 - P\left(\left|X - \frac{n}{6}\right| \geq \sqrt{n}\right) \\ &\geq 1 - \frac{5}{36} = \frac{31}{36}. \end{aligned}$$

Example. A random variable X has probability density function $f(x) = 12x^2(1-x)$ for $0 < x < 1$. Compute $P\left(|X - E(X)| \geq 2\sqrt{\text{var}(X)}\right)$ and compare it with the limits determined by Chebyshev's inequality.

Solution: Here,

$$\begin{aligned} E(X) &= \int_0^1 x f(x)dx = \int_0^1 x 12x^2(1-x)dx = \frac{3}{5} \\ E(X^2) &= \int_0^1 12x^4(1-x)dx = \frac{2}{5}. \end{aligned}$$

and $\text{var}(X) = E(X^2) - [E(X)]^2 = \frac{2}{5} - \frac{9}{25} = \frac{1}{25}$.

$$\begin{aligned}
\therefore P\left(|X - E(X)| \geq 2\sqrt{\text{var}(X)}\right) &= P\left(\left|X - \frac{3}{5}\right| \geq \frac{2}{5}\right) \\
&= 1 - P\left(\frac{3}{5} - \frac{2}{5} < X < \frac{3}{5} + \frac{2}{5}\right) \\
&= 1 - P\left(\frac{1}{5} < X < 1\right) \\
&= 1 - 12\int_{\frac{1}{5}}^{1} x^2(1-x)dx = \frac{17}{625}.
\end{aligned}$$

Now, by Chebyshev's inequality

$$P(|X - E(X)| \geq K\sigma) \leq \frac{1}{K^2}$$

and hence $P\left(|X - E(X)| \geq 2\sqrt{\text{var}(X)}\right) \leq \frac{1}{4}$.

Clearly, $\frac{17}{625} < \frac{1}{4}$. Thus, the above result supports the Chebyshev's limits.

Exercise

1. Let X be a random variable such that $E(X) = 2$ and $E(X^2) = 29$; then find the lower bound for $P(-5 < X < 7)$ using Chebyshev's inequality. **Ans.** $\frac{24}{49}$

2. The probability distribution of a discrete random variable X is given by

$$\begin{array}{lcccc} X = i: & -1 & 1 & 3 & 5 \\ P(X = i): & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{2} \end{array}$$

   Find the upper bound for $P(|X - 3| \geq 1)$ by Chebyshev's inequality. **Ans.** $\frac{3}{16}$

3. A random variable X has mean 3 and variance 2. Using Chebyshev's inequality to find the upper bound for

   (i) $P(|X - 3| \geq 2)$ $\quad$ (ii) $P(|X - 3| \geq 1)$. $\quad$ **Ans.(i)1, (ii)$\frac{1}{4}$**

4. A continuous random variable X follows normal distribution with parameters m and σ. Find $P(|X - m| \geq 1.5\sigma)$ and compare it with the value given by Chebyshev's inequality. **Ans.0.1336, 0.444**

5. A coin is tossed 400 times. Show that the probability that the number of heads will be between 150 and 200 is greater than 0.95.

6. If a die is thrown 1800 times, show that the probability that the number of sixes lies between 250 and 350 is at least $\frac{9}{10}$.

7. The probability density function of a continuous variable is given by

$$f(x) = \begin{cases} 6x(1-x), & \text{if } 0 \leq x \leq 1. \\ 0, & \text{otherwise.} \end{cases}$$

Find the lower bound for $P\left(|X - \frac{1}{2}| < 2\right)$ by Chebyshev's inequality. **Ans.** $\frac{79}{80}$