FREQUENCY DISTRIBUTIONS AND MEASURES OF CENTRAL TENDENCY

2·1. Frequency Distributions. When observations, discrete or continuous, are available on a single characteristic of a large number of individuals, often it becomes necessary to condense the data as far as possible without losing any information of interest. Let us consider the marks in Statistics obtained by 250 candidates selected at random from among those appearing in a certain examination.

TABLE 1 : MARKS IN STATISTICS OF 250 CANDIDATES

32	47	41	51	41	30	39	18	48	53
54	32	31	46	15	37	32	56	42	48
38	26	50	40	38	42	35	22	62	51
44	21	45	31	37	41	44	18	37	47
68	41	30	52	52	60	42	38	38	34
41	53	48	21	28	49	42	36	41	29
30	33	37	35	29	37	38	40	32	49
43	32	24	38	38	22	41	50	17	46
46	50	26	15	23	42	25	52	38	46
41	38	40	37	40	48	45	30	28	31
40	33	42	36	51	42	56	44	35	38
31	51	45	41	50	53	50	32	45	48
40	43	40	34	34	44	38	58	49	28
40	45	19	24	34	47	37	33	37	36
36	32	61	30	44	43	50	31	38	45
46	40	32	34	44	54	35	39	31	48
48	50	43	55	43	39	41	48	53	34
32	31	42	34	34	32	33	24	43	39
40	50	27	47	34	44	34	33	47	42
17	42	57	35	38	17	33	46	36	23
48	50	31	58	33	44	26	29	31	37
47	55	57	37	41	54	42	45	47	43
37	52	47	46	44	50	44	38	42	19
52	45	23	41	47	33	42	24	48	39
48	44	60	38	38	44	38	43	40	48

This representation of the data does not furnish any useful information and is rather confusing to mind. A better way may be to express the figures in an ascending or descending order of magnitude, commonly termed as *array*. But this does not reduce the bulk of the data. A much better representation is given on the next page.

A bar (|) called *tally mark* is put against the number when it occurs. Having occurred four times, the fifth occurrence is represented by putting a cross tally (/) on the first four tallies. This technique facilitates the counting of the tally marks at the end.

The representation of the data as above is known as *frequency distribution.* Marks are called the *variable* (x) and the *'number of students'* against the marks is known as the *frequency* (f) of the variable. The word *'frequency'* is derived from 'how frequently' a variable occurs. For example, in the above case the frequency of 31 is 10 as there are ten students getting 31 marks. This representation, though better than an array' , does not condense the data much and it is quite cumbersome to go through this huge mass of data.

TABLE 2

Marks	*No. of Students Tally Marks*	*Total Frequency*	*Marks*	*No. of Students Tally Marks*	*Total Frequency*
15	II	= 2	40	𝍸 𝍸 I	=11
17	III	= 3	41	𝍸 𝍸	=10
18	II	= 2	42	𝍸 𝍸 III	=13
19	II	= 2	43	𝍸 III	= 8
21	II	= 2	44	𝍸 𝍸 II	= 12
22	II	= 2	45	𝍸 II	= 7
23	III	= 3	46	𝍸 II	= 7
24	IIII	= 4	47	𝍸 III	= 8
25	I	= 1	48	𝍸 𝍸 II	= 12
26	III	= 3	49	III	= 3
27	I	= 1	50	𝍸 𝍸	= 10
28	III	= 3	51	IIII	= 4
29	II	= 2	52	𝍸	= 5
30	𝍸	= 5	53	IIII	= 4
31	𝍸 𝍸	= 10	54	III	= 3
32	𝍸 𝍸	= 10	55	II	= 2
33	𝍸 III	= 8	56	II	= 2
34	𝍸 𝍸 I	= 11	57	II	= 2
35	𝍸	= 5	58	II	= 2
36	𝍸	= 5	60	III	= 3
37	𝍸 𝍸 II	= 12	61	I	= 1
38	𝍸 𝍸 𝍸 II	= 17	62	I	= 1
39	𝍸 I	= 6	68	I	= 1

If the identity of the individuals about whom a particular information is taken is not relevant, nor the order in which the observations arise, then the first real step of condensation is to divide the observed range of variable into a suitable number of *class-intervals* and to record the number of observations in each class. For example, in the above case, the data may be expressed as shown in Table 3.

Such a table showing the distribution of the frequencies in the different classes is called a *frequency table* and the manner in which the class frequencies are distributed over the class intervals is called the *grouped frequency distribution* of the variable.

Remark. The classes of the type 15—19, 20—24, 25—29 etc., in which both the upper and lower limits are included are called *'inclusive classes'*. For example the class 20—24, includes

TABLE 3 : FREQUENCY TABLE

Marks (x)	*No. of students.* (f)
15 — 19	9
20 — 24	11
25 — 29	10
30 — 34	44
35 — 39	45
40 — 44	54
45 — 49	37
50 — 54	26
55 — 59	8
60 — 64	5
65 — 69	1
Total	250

all the values from 20 to 24, both inclusive and the classification is termed as *inclusive type classification.*

In spite of great importance of classification in statistical analysis, no hard and fast rules can be laid down for it. The following points may be kept in mind for classification:

(i) The classes should be clearly defined and should not lead to any ambiguity.

(ii) The classes should be exhaustive, *i.e.*, each of the given values should be included in one of the classes.

(iii) The classes should be mutually exclusive and non-overlapping.

(iv) The classes should be of equal width. The principle, however, cannot be rigidly followed. If the classes are of varying width, the different class frequencies will not be comparable. Comparable figures can be obtained by dividing the value of the frequencies by the corresponding widths of the class intervals. The ratios thus obtained are called *'frequency densities'*.

(v) Indeterminate classes, *e.g.*, the open-end classes, less than *'a'* or greater than *'b'* should be avoided as far as possible since they create difficulty in analysis and interpretation.

(vi) The number of classes should neither be too large nor too small. It should preferably lie between 5 and 15. However, the number of classes may be more than 15 depending upon the total frequency and the details required, but it is desirable that it is not less than 5 since in that case the classification may not reveal the essential characteristics of the population. The following formula due to Struges may be used to determine an approximate number k of classes:

$$k = 1 + 3{\cdot}322 \log_{10} N,$$

where N is the total frequency.

The Magnitude of the Class Interval

Having fixed the number of classes, divide the range (the difference between the greatest and the smallest observation) by it and the nearest integer to this value gives the magnitude of the class interval. Broad class intervals (*i.e.*, less number of classes) will yield only rough estimates while for high degree of accuracy small class intervals (*i.e.*, large number of classes) are desirable.

Class Limits

The class limits should be chosen in such a way that the mid-value of the class interval and actual average of the observations in that class interval are as near to each other as possible. If this is not the case then the classification gives a distorted picture of the characteristics of the data. If possible, class limits should be located at the points which are multiple of 0, 2, 5, 10,... etc., so that the midpoints of the classes are the common figures, *viz.*, 0, 2, 5, 10..., etc., the figures capable of easy and simple analysis.

2·1·1. Continuous Frequency Distribution. If we deal with a continuous variable, it is not possible to arrange the data in the class intervals of above type. Let us consider the distribution of age in years. If class intervals are 15—19, 20—24 then the persons with ages between 19 and 20 years are not taken into consideration. In such a case we form the class intervals as shown below.

Age in years
Below 5
5 or more but less than 10
10 or more but less than 15
15 or more but less than 20
20 or more but less than 25
and so on.

Here all the persons with any fraction of age are included in one group or the other. For practical purpose we re-write the above classes as

0 — 5
5 — 10
10 — 15
15 — 20
20 — 25

This form of frequency distribution is known as *continuous frequency distribution.*

It should be clearly understood that in the above classes, the upper limits of each class are excluded from the respective classes. Such classes in which the upper limits are excluded from the respective classes and are included in the immediate next class are known as *'exclusive classes'* and the classification is termed as *'exclusive type classification'*.

2·2. Graphic Representation of a Frequency Distribution. It is often useful to represent a frequency distribution by means of a diagram which makes the unwieldy data intelligible and conveys to the eye the general run of the observations. Diagrammatic representation also facilitates the comparison of two or more frequency distributions. We consider below some important types of graphic representation.

2·2·1. Histogram. In drawing the histogram of a given continuous frequency distribution we first mark off along the x–axis all the class intervals on a suitable scale. On each class interval erect rectangles with heights proportional to the frequency of the corresponding class interval so that the area of the rectangle is proportional to the frequency of the class. If, however, the classes are of unequal width then the height of the rectangle will be proportional to the ratio of the frequencies to the width of the classes. The diagram of continuous rectangles so obtained is called *histrogram.*

Remarks. 1. To draw the histogram for an ungrouped frequency distribution of a variable we shall have to assume that the frequency corresponding to the variate value x is spread over the interval $x - h/2$ to $x + h/2$, where h is the jump from one value to the next.

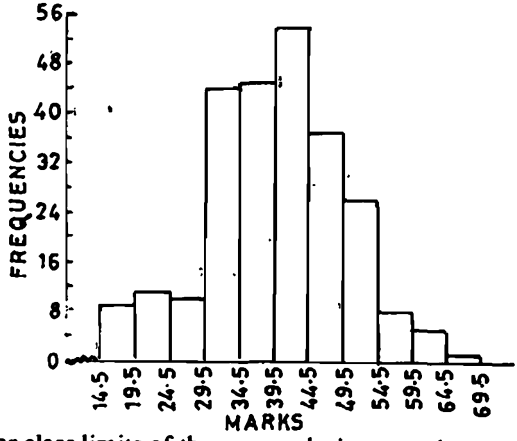
2. If the grouped frequency distribution is not continuous, first it is to be converted into continuous distribution and then the histrogam is drawn.

3. Although the height of each rectangle is proportional to the frequency of the corresponding class, the height of a fraction of the rectangle is not proportional to the frequency of the corresponding fraction of the class, so that histogram cannot be directly used to read frequency over a fraction of a class interval.

4. The histogram of the distribution of marks of 250 students in *Table* 3 (page 2·2) is obtained as follows.

Since the grouped frequency distribution is not continuous, we first convert it into a continuous distribution as follows:

Marks	*No. of Students*
14·5–19·5	9
19·5–24·5	11
24·5–29·5	10
29·5–34·5	44
34·5–39·5	45
39·5–44·5	54
44·5–49·5	37
49·5–54·5	26
54·5–59·5	8
59·5–64·5	5
64·5–69·5	1



Remark. The upper and lower class limits of the *new exclusive type classes* are known as *class boundaries.*

If d is the gap between the upper limit of any class and the lower limit of the succeeding class, the class boundaries for any class are then given by :

$$\text{Upper class boundary} = \text{Upper class limit} + \frac{d}{2}$$

$$\text{Lower class boundary} = \text{Lower class limit} - \frac{d}{2}$$

2·2·2. **Frequency Polygon.** For an ungrouped distribution, the frequency polygon is obtained by plotting points with abscissa as the variate values and the ordinate as the corresponding frequencies and joining the plotted points by means of straight lines. For a grouped frequency distribution, the abscissa of points are mid-values of the class intervals. For equal class intervals the frequency polygon can be obtained by joining the middle points of the upper sides of the adjacent rectangles of the histogram by means of straight lines. If the class intervals are of small width the polygon can be approximated by a smooth curve. The frequency curve can be obtained by drawing a smooth freehand curve through the vertices of the frequency polygon.

2·3. Averages or Measures of Central Tendency or Measures of Location. According to Professor Bowley, averages are "statistical constants which enable us to comprehend in a single effort the significance of the whole." They give us an idea about the concentration of the values in the central part of the distribution. Plainly speaking, an average of a statistical series is the value of the variable which is representative of the entire distribution. The following are the five measures of central tendency that are in common use:

(i) Arithmetic Mean or simply Mean , (ii) Median,

(iii) Mode, (iv) Geometric Mean, and (v) Harmonic Mean.

2·4. Requisites for an Ideal Measure of Central Tendency. According to Professor Yule, the following are the characteristics to be satisfied by an ideal measure of central tendency :

(i) It should be rigidly defined.

(ii) It should be readily comprehensible and easy to calculate.

(iii) It should be based on all the observations.

(iv) It should be suitable for further mathematical treatment. By this we mean that if we are given the averages and sizes of a number of series, we should be able to calculate the average of the composite series obtained on combining the given series.

(v) It should be affected as little as possible by fluctuations of sampling.

In addition to the above criteria, we may add the following (which is not due to Prof. Yule) :

(vi) It should not be affected much by extreme values.

2·5. Arithmetic Mean. Arithmetic mean of a set of observations is their sum divided by the number of observations, *e.g.*, the arithmetic mean $\bar{x}$ of n observations $x_1, x_2, \ldots, x_n$ is given by

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n \right) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

In case of frequency distribution $x_i \mid f_i$, $i = 1, 2, \ldots, n$, where f_i is the frequency of the variable x_i;

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{f_1 + f_2 + \ldots f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N} \sum_{i=1}^{n} f_i x_i, \quad \left[\sum_{i=1}^{n} f_i = N \right] \qquad \ldots(2\cdot1)$$

In case of grouped or continuous frequency distribution, x is taken as the mid-value of the corresponding class.

Remark. The symbol Σ is the letter capital sigma of the Greek alphabet and is used in mathematics to denote the sum of values.

Example 2·1. *(a) Find the arithmetic mean of the following frequency distribution:*

x :	1	2	3	4	5	6	7
f :	5	9	12	17	14	10	6

(b) Calculate the arithmetic mean of the marks from the following table :

Marks :	0–10	10–20	20–30	30–40	40–50	50–60
No. of students :	12	18	27	20	17	6

Solution. *(a)*

x	f	fx
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
	73	299

$$\therefore \qquad \bar{x} = \frac{1}{N} \Sigma f x = \frac{299}{73} = 4{\cdot}09$$

(b)

Marks	*No. of students* (f)	*Mid - point* (x)	fx
0–10	12	5	60
10–20	18	15	270
20–30	27	25	675
30–40	20	35	700
40–50	17	45	765
50–60	6	55	330
Total	100		2,800

Arithmetic mean or $\bar{x} = \frac{1}{N} \Sigma f x = \frac{1}{100} \times 2{,}800 = 28$

It may be noted that if the values of x or (and) f are large, the calculation of mean by formula *(2·1)* is quite time-consuming and tedious. The arithmetic is reduced to a great extent by taking the deviations of the given values from any arbitrary point 'A', as explained below.

Let $d_i = x_i - A$, then $f_i d_i = f_i (x_i - A) = f_i x_i - A f_i$

Summing both sides over i from 1 to n, we get

$$\sum_{i=1}^{n} f_i d_i = \sum_{i=1}^{n} f_i x_i - A \sum_{i=1}^{n} f_i = \sum_{i=1}^{n} f_i x_i - A \,.\, N.$$

$$\Rightarrow \quad \frac{1}{N}\sum_{i=1}^{n} f_i d_i = \frac{1}{N}\sum_{i=1}^{n} f_i x_i - A = \bar{x} - A,$$

where $\bar{x}$ is the arithmetic mean of the distribution.

$$\therefore \quad \bar{x} = A + \frac{1}{N}\sum_{i=1}^{n} f_i d_i \qquad ...(2·2)$$

This formula is much more convenient to apply than formula (2·1).

Any number can serve the purpose of arbitrary point 'A' but, usually, the value of x corresponding to the middle part of the distribution will be much more convenient.

In case of grouped or continuous frequency distribution, the arithmetic is reduced to a still greater extent by taking

$$d_i = \frac{x_i - A}{h},$$

where A is an arbitrary point and h is the common magnitude of class interval. In this case, we have

$$h\,d_i = x_i - A,$$

and proceeding exactly similarly as above, we get

$$\bar{x} = A + \frac{h}{N}\sum_{i=1}^{n} f_i d_i \qquad ...(2·3)$$

Example 2·2. *Calculate the mean for the following frequency distribution.*

Class-interval :	*0–8*	*8–16*	*16–24*	*24–32*	*32–40*	*40–48*
Frequency :	*8*	*7*	*16*	*24*	*15*	*7*

Solution.

Class-interval	*mid-value* (x)	*Frequency* (f)	$d = (x-A)/h$	fd
0–8	4	8	–3	–24
8–16	12	7	–2	–14
16–24	20	16	–1	–16
24–32	28	24	0	0
32–40	36	15	1	15
40–48	44	7	2	14
		77		–25

Here we take $A = 28$ and $h = 8$.

$$\therefore \quad \bar{x} = A + \frac{h\Sigma fd}{N} = 28 + \frac{8 \times (-25)}{77} = 28 - \frac{200}{77} = 25·404$$

2·5·1. Properties of Arithmetic Mean

Property 1. *Algebraic sum of the deviations of a set of values from their arithmetic mean is zero. If* $x_i | f_i$, $i = 1, 2, ..., n$ *is the frequency distribution, then*

$$\sum_{i=1}^{n} f_i (x_i - \bar{x}) = 0, \ \bar{x} \text{ being the mean of distribution.}$$

Proof. $\sum_i f_i (x_i - \bar{x}) = \sum_i f_i x_i - \bar{x} \sum f_i = \sum_i f_i x_i - \bar{x} . N$

Also $\bar{x} = \dfrac{\sum_i f_i x_i}{N} \quad \Rightarrow \quad \sum_i f_i x_i = N \bar{x}$

Hence $\sum_{i=1}^{n} f_i (x_i - \bar{x}) = N . \bar{x} - \bar{x} . N = 0$

Property 2. *The sum of the squares of the deviations of a set of values is minimum when taken about mean.*

Proof. For the frequency distribution $x_i \mid f_i$, $i = 1, 2, ..., n$, let

$$Z = \sum_{i=1}^{n} f_i (x_i - A)^2,$$

be the sum of the squares of the deviations of given values from any arbitrary point 'A'. We have to prove that Z is minimum when $A = \bar{x}$.

Applying the principle of maxima and minima from differential calculus, Z will be minimum for variations in A if

$$\frac{\partial Z}{\partial A} = 0 \text{ and } \frac{\partial^2 Z}{\partial A^2} > 0$$

Now $\dfrac{\partial Z}{\partial A} = -2 \sum_i f_i (x_i - A) = 0 \quad \Rightarrow \quad \sum_i f_i (x_i - A) = 0$

$\Rightarrow \quad \Sigma f_i x_i - A \Sigma f_i = 0 \quad or \quad A = \dfrac{\Sigma f_i x_i}{N} = \bar{x}$

Again $\dfrac{\partial^2 Z}{\partial A^2} = -2 \sum_i f_i (-1) = 2 \sum_i f_i = 2N > 0$

Hence Z is minimum at the point $A = \bar{x}$. This establishes the result.

Property 3. *(Mean of the composite series). If $\bar{x}_i$, $(i = 1, 2, ..., k)$ are the means of k-component series of sizes n_i, $(i = 1, 2, ..., k)$ respectively, then the mean $\bar{x}$ of the composite series obtained on combining the component series is given by the formula:*

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \ ... \ + n_k \bar{x}_k}{n_1 + n_2 + \ ... \ + n_k} = \sum_i n_i \bar{x}_i \ / \ \sum_i n_i \qquad ...(2·4)$$

Proof. Let $x_{11}, x_{12}, ... x_{1n_1}$ be n_1 members of the first series; $x_{21}, x_{22}, ...$ x_{2n_2} be n_2 members of the second series, $x_{k1}, x_{k2}, ..., x_{kn_k}$ be n_k members of the kth series. Then, by def.,

$$\left.\begin{aligned} \bar{x}_1 &= \frac{1}{n_1}(x_{11} + x_{12} + \ldots + x_{1n_1}) \\ \bar{x}_2 &= \frac{1}{n_2}(x_{21} + x_{22} + \ldots + x_{2n_2}) \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \\ \bar{x}_k &= \frac{1}{n_k}(x_{k1} + x_{k2} + \ldots + x_{kn_k}) \end{aligned}\right\} \quad \ldots(*)$$

The mean $\bar{x}$ of composite series of size $n_1 + n_2 + \ldots + n_k$ is given by

$$\bar{x} = \frac{(x_{11} + x_{12} + \ldots + x_{1n_1}) + (x_{21} + x_{22} + \ldots + x_{2n_2}) + \ldots + (x_{k1} + x_{k2} + \ldots + x_{kn_k})}{n_1 + n_2 + \ldots + n_k}$$

$$= \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots + n_k\bar{x}_k}{n_1 + n_2 + \ldots n_k}, \qquad \text{[From (*)]}$$

Thus, $\bar{x} = \sum_i n_i\bar{x}_i \;/\; (\sum_i n_i)$

Example 2·3. *The average salary of male employees in a firm was Rs.520 and that of females was Rs.420. The mean salary of all the employees was Rs.500. Find the percentage of male and female employees.*

Solution. Let n_1 and n_2 denote respectively the number of male and female employees in the concern and $\bar{x}_1$ and $\bar{x}_2$ denote respectively their average salary (in rupees). Let $\bar{x}$ denote the avarage salary of all the workers in the firm.

We are given that :

$$\bar{x}_1 = 520, \quad \bar{x}_2 = 420 \quad \text{and} \quad \bar{x} = 500$$

Also we know

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$\Rightarrow \quad 500\,(n_1 + n_2) = 520\,n_1 + 420\,n_2$$

$$\Rightarrow \quad (520 - 500)\,n_1 = (500 - 420)\,n_2$$

$$\Rightarrow \quad 20\,n_1 = 80\,n_2$$

$$\Rightarrow \quad \frac{n_1}{n_2} = \frac{4}{1}$$

Hence the percentage of male employees in the firm

$$= \frac{4}{4+1} \times 100 = 80$$

and percentage of female employees in the firm

$$= \frac{1}{4+1} \times 100 = 20$$

2·5·2. Merits and Demerits of Arithmetic Mean

Merits. *(i)* It is rigidly defined .

(ii) It is easy to understand and easy to calculate.

(iii) It is based upon all the observations.

(iv) It is amenable to algebraic treatment. The mean of the composite series in terms of the means and sizes of the component series is given by

$$\bar{x} = \sum_{i=1}^{k} n_i \bar{x}_i / \left(\sum_{i}^{k} n_i \right)$$

(v) Of all the averages, arithmetic mean is affected least by fluctuations of sampling. This property is sometimes described by saying that arithmetic mean is a *stable* average.

Thus, we see that arithmetic mean satisfies all the properties laid down by Prof. Yule for an ideal average.

Demerits. *(i)* It cannot be determined by inspection nor it can be located graphically.

(ii) Arithmetic mean cannot be used if we are dealing with qualitative characteristics which cannot be measured quantitively; such as, intelligence, honesty, beauty, etc. In such cases median (discussed later) is the only average to be used.

(iii) Arithmetic mean cannot be obtained if a single observation is missing or lost or is illegible unless we drop it out and compute the arithmetic mean of the remaining values.

(iv) Arithmetic mean is affected very much by extreme values. In case of extreme items, arithmetic mean gives a distorted picture of the distribution and no longer remains representative of the distribution.

(v) Arithmetic mean may lead to wrong conclusions if the details of the data from which it is computed are not given. Let us consider the following marks obtained by two students *A* and *B* in three tests, *viz.*, terminal test, half-yearly examination and annual examination respectively.

Marks in : →	*I Test*	*II Test*	*III Test*	*Average marks*
A	*50%*	*60%*	*70%*	*60%*
B	*70%*	*60%*	*50%*	*60%*

Thus average marks obtained by each of the two students at the end of the year are 60%. If we are given the average marks alone we conclude that the level of intelligence of both the students at the end of the year is same. This is a fallacious conclusion since we find from the data that student *A* has improved consistently while student *B* has deteriorated consistently.

(vi) Arithmetic mean cannot be calculated if the extreme class is open, *e.g.*, below 10 or above 90. Morever, even if a single observation is missing mean cannot be calculated.

(vii) In extremely asymmertrical (skewed) distribution, usually arithmetic mean is not a suitable measure of location.

2·5·3. Weighted Mean. In calculating arithmetic mean we suppose that all the items in the distribution have equal importance. But in practice this may not be so. If some items in a distribution are more important than others, then this

point must be borne in mind, in order that average computed is representative of the distribution. In such cases, proper weightage is to be given to various items — the weights attached to each item being proportional to the importance of the item in the distribution. For example, if we want to have an idea of the change in cost of living of a certain group of people, then the simple mean of the prices of the commodities consumed by them will not do, since all the commodities are not equally important, *e.g.*, wheat, rice and pulses are more important than cigarettes, tea, confectionery, etc.

Let w_i be the weight attached to the item x_i, $i = 1, 2, \ldots, n$. Then we define :

Weighted arithmetic mean or weighted mean $= \sum_i w_i x_i / \sum_i w_i$ ... (2·5)

It may be observed that the formula for weighted mean is the same as the formula for simple mean with f_i, ($i = 1, 2, \ldots, n$), *the frequencies replaced by* w_i, ($i = 1, 2, \ldots, n$), *the weights.*

Weighted mean gives the result equal to the simple mean if the weights assigned to each of the variate values are equal. It results in higher value than the simple mean if smaller weights are given to smaller items and larger weights to larger items. If the weights attached to larger items are smaller and those attached to smaller items are larger, then the weighted mean results in smaller value than the simple mean.

Example 2·4. *Find the simple and weighted arithmetic mean of the first n natural numbers, the weights being the corresponding numbers.*

Solution. The first *natural* numbers are 1, 2, 3, ..., n.

X	w	wX
1	1	1^2
2	2	2^2
3	3	3^2
⋮	⋮	⋮
n	n	n^2

We know that

$$1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2}$$

$$1^2 + 2^2 + 3^2 + \ldots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

Simple A.M. is

$$\overline{X} = \frac{\sum X}{n} = \frac{1 + 2 + 3 + \ldots + n}{n} = \frac{n+1}{2}$$

Weighted A.M. is

$$\overline{X}_w = \frac{\sum wX}{\sum w} = \frac{1^2 + 2^2 + \ldots + n^2}{1 + 2 + \ldots + n}$$

$$= \frac{n(n+1)(2n+1)}{6} \cdot \frac{2}{n(n+1)}$$

2·6. Median. Median of a distribution is the value of the variable which divides it into two equal parts. It is the value which exceeds and is exceeded by the same number of observations, *i.e.*, it is the value such that the number of observations above it is equal to the number of observations below it. The median is thus a *positional* average.

In case of ungrouped data, if the number of observations is odd then median is the middle value after the values have been arranged in ascending or descending order of magnitude. In case of even number of observations, there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms. For example, the median of the values 25, 20, 15, 35, 18, *i.e.*, 15, 18, 20, 25, 35 is 20 and the median of 8, 20, 50, 25, 15, 30, *i.e.*, of 8, 15, 20, 25, 30, 50 is $\frac{1}{2}(20 + 25) = 22{\cdot}5$.

Remark. In case of even number of observations, in fact any value lying between the two middle values can be taken as median but conventionally we take it to be the mean of the middle terms.

In case of discrete frequency distribution median is obtained by considering the cumulative frequencies. The steps for calculating median are given below:

(i) Find $N/2$, where $N = \Sigma_i f_i$.

(ii) See the (less than) cumulative frequency (*c.f.*) just greater than $N/2$.

(iii) The corresponding value of x is median.

Example 2·5. *Obtain the median for the following frequency distribution:*

x :	1	2	3	4	5	6	7	8	9
f :	8	10	11	16	20	25	15	9	6

Solution.

x	f	*c.f.*
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
	120	

Hence $N = 120 \Rightarrow N/2 = 60$

Cumulative frequency (*c.f.*) just greater than $N/2$, is 65 and the value of x corresponding to 65 is 5. Therefore, median is 5.

In the case of continuous frequency distribution, the class corresponding to the *c.f.* just greater than $N/2$ is called the *median class* and the value of median is obtained by the following formula :

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - c\right) \qquad ...(2·6)$$

where l is the lower limit of the median class,
f is the frequency of the median class,
h is the magnitude of the median class,
'c' is the *c.f.* of the class *preceding* the median class,
and $N = \Sigma f$.

Derivation of the Median Formula (2·6). Let us consider the following continuous frequency distribution, ($x_1 < x_2 < \ldots < x_{n+1}$) :

Class interval : $x_1 - x_2$, $x_2 - x_3$, ... $x_k - x_{k+1}$, ... $x_n - x_{n+1}$

Frequency : f_1 f_2 ..... f_k ..... f_n

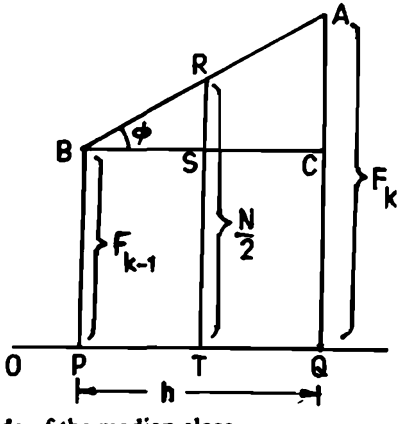
The cumulaltive frequency distribution is given by :

Class interval : $x_1 - x_2$, $x_2 - x_3$, ... $x_k - x_{k+1}$, ... $x_n - x_{n+1}$

Frequency : F_1 F_2 ..... F_k ..... F_n

where $F_i = f_1 + f_2 + \ldots + f_i$. The class $x_k - x_{k+1}$ is the median class if and only if $F_{k-1} < N/2 \leq F_k$.

Now, if we assume that the variate values are uniformly distributed over the median-class which implies that the ogive is a straight line in the median-class, then we get from the Fig. 1,



$$\tan\phi = \frac{RS}{BS} = \frac{AC}{BC}$$

i.e. $$\frac{RT - TS}{BS} = \frac{AQ - CQ}{BC}$$

or $$\frac{RT - BP}{BS} = \frac{AQ - BP}{PQ}$$

or $$\frac{N/2 - F_{k-1}}{BS} = \frac{F_k - F_{k-1}}{PQ}$$

$$= \frac{f_k}{h}$$

where f_k is the frequency and h the magnitude of the median class.

$$\therefore \qquad BS = \frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right)$$

Hence

$$\text{Median} = OT = OP + PT = OP + BS$$

$$= l + \frac{h}{f_k}\left(\frac{N}{2} - F_{k-1}\right)$$

which is the required formula.

Remark. The median formula (2·6) can be used only for continuous classes without any gaps, *i.e.*, for '*exclusive type*' classification. If we are given a frequency

distribution in which classes are of *'inclusive type'* with gaps, then it must be converted into a continuous *'exclusive type'* frequency distribution without any gaps before applying (2·6). This will affect the value of l in (2·6). As an illustration see Example 2·7.

Example 2·6. *Find the median wage of the following distribution :*

Wages (in Rs.) :	*20—30*	*30—40*	*40—50*	*50—60*	*60—70*
No. of labourers :	*3*	*5*	*20*	*10*	*5*

[Gorakhpur Univ. B. Sc. 1989]

Solution.

Wages (in Rs.)	*No. of labourers*	*c.f.*
20—30	3	3
30—40	5	8
40—50	20	28
50—60	10	38
60—70	5	43

Here $N/2 = 43/2 = 21{\cdot}5$. Cumulative frequency just greater than 21·5 is 28 and the corresponding class is 40–50. Thus median class is 40–50. Hence using (2·6), we get

$$\text{Median} = 40 + \frac{10}{20}(21{\cdot}5 - 8) = 40 + 6{\cdot}75 = 46{\cdot}75$$

Thus median wage is Rs. 46·75.

Example 2·7. *In a factory employing 3,000 persons, 5 per cent earn less than Rs. 3 per hour, 580 earn from Rs. 3·01 to Rs. 4·50 per hour, 30 percent earn from Rs. 4·51 to Rs. 6·00 per hour, 500 earn from Rs. 6·01 to Rs. 7·50 per hour, 20 percent earn from Rs. 7·51 to Rs. 9·00 per hour, and the rest earn Rs. 9·01 or more per hour. What is the median wage?* **[Utkal Univ. B.Sc.1992]**

Solution. The given information can be expressed in tabular form as follows.

CALCULATIONS FOR MEDIAN WAGE

Earnings (in Rs.)	*Percentage of workers*	*No. of workers (f)*	*Less than c.f.*	*Class boundaries*
less than 3	5%	$\frac{5}{100} \times 3000 = 150$	150	Below 3·005
3·01—4·50	—	580	730	3·005—4·505
4·51—6·00	30 %	$\frac{30}{100} \times 3000 = 900$	1630	4·505—6·005
6·01—7·50	—	500	2130	6·005—7·505
7·51—9·00	20 %	$\frac{20}{100} \times 3000 = 600$	2730	7·505—9·005
9·01 and above		$3000 - 2730 = 270$	$3000 = N$	9·005 and above

$N/2 = 1500$. The *c.f.* just greater than 1500 is 1630. The corresponding class 4·51–6·00, whose class boundaries are 4·505–6·005, is the median class. Using the median formula, we get :

$$\text{Median} = l + \frac{h}{f}\left(\frac{N}{2} - C\right) = 4{\cdot}505 + \frac{1{\cdot}5}{900}(1500 - 730)$$
$$= 4{\cdot}505 + 1{\cdot}283 \approx 5{\cdot}79$$

Hence median wage is Rs. 5·79.

Example 2·8. *An incomplete frequency distribution is given as follows:*

Variable	*Frequency*	*Variable*	*Frequency*
10—20	*12*	*50—60*	*?*
20—30	*30*	*60—70*	*25*
30—40	*?*	*70—80*	*18*
40—50	*65*	*Total*	*229*

Given that the median value is 46, determine the missing frequencies using the median formula. **[Delhi Univ. B. Sc., Oct. 1992]**

Solution. Let the frequency of the class 30—40 be f_1 and that of 50—60 be f_2.

Then $f_1 + f_2 = 229 - (12 + 30 + 65 + 25 + 18) = 79$.

Since median is given to be 46, the class 40—50 is the median class.

Hence using median formula (2·6), we get

$$46 = 40 + \frac{114{\cdot}5 - (12 + 30 + f_1)}{65} \times 10$$

$$46 - 40 = \frac{72.5 - f_1}{65} \times 10 \text{ or } 6 = \frac{72{\cdot}5 - f_1}{6{\cdot}5}$$

$$f_1 = 72{\cdot}5 - 39 = 33{\cdot}5 \approx 34$$

[Since frequency is never fractional]

$\therefore$ $f_2 = 79 - 34 = 45$ [Since $f_1 + f_2 = 79$]

2·6·1. Merits and Demerits of Median

Merits. *(i)* It is rigidly defined.

(ii) It is easily understood and is easy to calculate. In some cases it can be located merely by inspection.

(iii) It is not at all affected by extreme values.

(iv) It can be calculated for distributions with open-end classes.

Demerits. *(i)* In case of even number of observations median cannot be determined exactly. We merely estimate it by taking the mean of two middle terms.

(ii) It is not based on all the observations. For example, the median of 10, 25, 50, 60 and 65 is 50. We can replace the observations 10 and 25 by any two values which are smaller than 50 and the observations 60 and 65 by any two values greater than 50 without affecting the value of median. This property is sometimes described

by saying that median is *insensitive*.

(iii) It is not amenable to algebraic treatment.

(iv) As compared with mean, it is affected much by fluctuations of sampling.

Uses. *(i)* Median is the only average to be used while dealing with qualitative data which cannot be measured quantitatively but still can be arranged in ascending or descending order of magnitude, *e.g.*, to find the average intelligence or average honesty among a group of people.

(ii) It is to be used for determining the typical value in problems concerning wages, distribution of wealth, etc.

2·7. Mode. Let us cosider the following statements :

(i) The average height of an Indian (male) is 5'–6".

(ii) The average size of the shoes sold in a shop is 7.

(iii) An average student in a hostel spends Rs.150 p.m.

In all the above cases, the average referred to is mode. Mode is the value which occurs most frequently in a set of observastions and around which the other items of the set cluster densely. In other words, mode is the value of the variable which is predominant in the series. Thus in the case of discrete frequency distribution mode is the value of x corresponding to maximum frequency. For example, in the following frequency distribution :

x :	1	2	3	4	5	6	7	8
f :	4	9	16	25	22	15	7	3

the value of x corresponding to the maximum frequency, *viz.*, 25 is 4. Hence mode is 4.

But in any one (or more) of the following cases :

(i) if the maximum frequency is repeated,

(ii) if the maximum frequency occurs in the very beginning or at the end of the distribution, and

(iii) if there are irregularities in the distribution,

the value of mode is determined by the *method of grouping*, which is illustrated below by an example.

Example 2·9. *Find the mode of the following frequency distribution :*

Size (x) :	1	2	3	4	5	6	7	8	9	10	11	12
Frequency (f) :	3	8	15	23	35	40	32	28	20	45	14	6

Solution. Here we see that the distribution is not regular since the frequencies are increasing steadily up to 40 and then decrease but the frequency 45 after 20 does not seem to be consistent with the distribution. Here we cannot say that since maximum frequency is 45, mode is 10. Here we shall locate mode by the method of grouping as explained below :

The frequencies in column *(i)* are the original frequencies. Column *(ii)* is obtained by combining the frequencies two by two. If we leave the first frequency and combine the remaining frequencies two by two we get column *(iii)*. Combining

Size	*Frequency*					
(x)	*(i)*	*(ii)*	*(iii)*	*(iv)*	*(v)*	*(vi)*
1	3					
2	8	11				
3	15		23			
4	23	38		26		
5	35		58		46	
6	40	75				73
7	32		72			
8	28	60		98		
9	20		48			
10	45	65		80		
11	14		59		107	
12	6	20		65		100
					93	
						79

the frequencies two by two after leaving the first two frequencies results in a repetition of column *(ii)*. Hence, we proceed to combine the frequencies three by three, thus getting column *(iv)*. The combination of frequencies three by three after leaving the first frequency results in column *(v)* and after leaving the first two frequencies results in column *(vi)*.

The maximum frequency in each column is given in black type. To find mode we form the following table :

ANALYSIS TABLE

Column Number (1)	*Maximum Frequency (2)*	*Value or combination of values of x giving max. frequency in (2) (3)*
(i)	45	10
(ii)	75	5, 6
(iii)	72	6, 7
(iv)	98	4, 5, 6,
(v)	107	5, 6, 7
(vi)	100	6, 7, 8

On examining the values in column (3) above, we find that the value 6 is repeated the maximum number of times and hence the value of mode is 6 and not 10 which is an irregular item.

In case of continuous frequency distribution, mode is given by the formula :

$$Mode = l + \frac{h(f_1 - f_0)}{(f_1 - f_0) - (f_2 - f_1)} = l + \frac{h(f_1 - f_0)}{2f_1 - f_0 - f_2} \qquad ...(2·7)$$

where l is the lower limit, h the magnitude and f_1 the frequency of the modal class, f_0 and f_2 are the frequencies of the classes preceding and succeeding the modal class respectively.

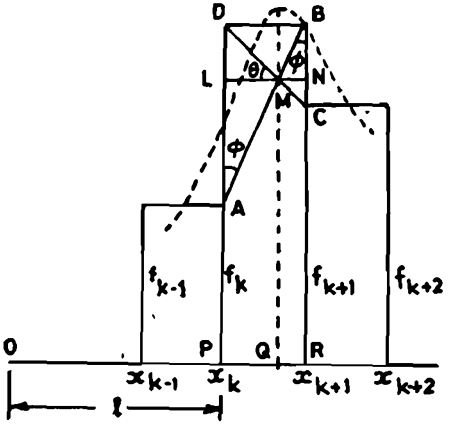
Derivation of the Mode Formula (2·7). Let us consider the continuous frequency distribution :

Class : $x_1 - x_2$, $x_2 - x_3$, ... , $x_k - x_{k+1}$, ... , $x_n - x_{n+1}$

Frequency : f_1 f_2 ... ... f_k ... ... f_n.

If f_k is the maximum of all the frequencies, then the modal class is $(x_k - x_{k+1})$.

Let us further consider a portion of the histogram, namely, the rectangles erected on the modal class and the two adjacent classes. The mode is the value of x for which the frequency curve has a maxima. Let the modal point be Q.



From the figure, we have

$$\tan\theta = \frac{LD}{LM} = \frac{NC}{MN}$$

and $$\tan\phi = \frac{LM}{AL} = \frac{MN}{NB}$$

$$\therefore \quad \frac{LM}{MN} = \frac{LD}{NC} = \frac{AL}{NB} = \frac{AL + LD}{NB + NC} = \frac{AD}{BC}$$

i.e., $$\frac{LM}{LN - LM} = \frac{PD - AP}{BR - CR}$$

or $$\frac{LM}{h - LM} = \frac{f_k - f_{k-1}}{f_k - f_{k+1}},$$ where 'h' is the magnitude of the modal class. Thus solving for LM, we get

$$LM = \frac{h(f_k - f_{k-1})}{(f_k - f_{k+1}) + (f_k - f_{k-1})} = \frac{h(f_k - f_{k-1})}{2f_k - f_{k-1} - f_{k+1}}$$

Hence $\text{Mode} = OQ = OP + PQ = OP + LM$

$$= l + \frac{h(f_k - f_{k-1})}{2f_k - f_{k-1} - f_{k+1}}$$

Example 2·10. *Find the mode for the following distribution :*

Class - interval :	*0–10*	*10–20*	*20–30*	*30–40*	*40–50*	*50–60*	*60–70*	*70–80*
Frequency :	*5*	*8*	*7*	*12*	*28*	*20*	*10*	*10*

Solution. Here maximum frequency is 28. Thus the class 40–50 is the modal class. Using (2·7), the value of mode is given by

$$\text{Mode} = 40 + \frac{10(28 - 12)}{(2 \times 28 - 12 - 20)} = 40 + 6{\cdot}666 = 46{\cdot}67 \quad \text{(approx.)}$$

Example 2·11. *The Median and Mode of the following wage distribution are known to be Rs. 33·50 and Rs. 34 respectively. Find the values of f_3, f_4 and f_5.*

Wages : (in Rs.)	*0–10*	*10–20*	*20–30*	*30–40*	*40–50*
Frequency :	*4*	*16*	f_3	f_4	f_5

Wages :	*50–60*	*60–70*	*Total*
Frequency :	*6*	*4*	*230*

[Gujarat Univ. B.Sc., 1991]

Solution.

CALCULATIONS FOR MODE AND MEDIAN

Wages (in Rs.)	*Frequency (f)*	*Less than c.f.*
0—10	4	4
10—20	16	20
20—30	f_3	$20 + f_3$
30—40	f_4	$20 + f_3 + f_4$
40—50	f_5	$20 + f_3 + f_4 + f_5$
50—60	6	$26 + f_3 + f_4 + f_5$
60—70	4	$30 + f_3 + f_4 + f_5$
Total	$230 = 30 + f_3 + f_4 + f_5$	

From the above table, we get

$$\Sigma f = 30 + f_3 + f_4 + f_5 = 230$$

$$\Rightarrow \quad f_3 + f_4 + f_5 = 230 - 30 = 200 \qquad ...(i)$$

Since median is 33·5, which lies in the class 30–40, 30–40 is the median class. Using the median formula, we get

$$Md = l + \frac{h}{f}\left(\frac{N}{2} - C\right)$$

$$\Rightarrow \quad 33{\cdot}5 = 30 + \frac{10}{f_4}\left[115 - (20 + f_3)\right]$$

$$\Rightarrow \quad \frac{33{\cdot}5 - 30}{10} = \frac{95 - f_3}{f_4}$$

$$\Rightarrow \quad 0{\cdot}35 f_4 = 95 - f_3 \quad \Rightarrow \quad f_3 = 95 - 0{\cdot}35 f_4 \qquad ...(ii)$$

Mode being 34, the modal class is also 30–40. Using mode formula we get :

$$34 = 30 + \frac{10(f_4 - f_3)}{2f_4 - f_3 - f_5}$$

$$\Rightarrow \quad \frac{34 - 30}{10} = \frac{f_4 + 0{\cdot}35 f_4 - 95}{2f_4 - (200 - f_4)} \qquad \text{[Using } (i) \text{ and } (ii)\text{]}$$

$$\Rightarrow \quad 0{\cdot}4 = \frac{1{\cdot}35 f_4 - 95}{3f_4 - 200}$$

$$\Rightarrow \quad 1{\cdot}2 f_4 - 80 = 1{\cdot}35 f_4 - 95$$

$$\Rightarrow \quad f_4 = \frac{95 - 80}{1{\cdot}35 - 1{\cdot}20} = \frac{15}{0{\cdot}15} = 100 \qquad ...(iii)$$

Substituting in *(ii)* we get :

$$f_3 = 95 - 0{\cdot}35 \times 100 = 60$$

Substituting the values of f_3 and f_4 in *(i)* we get :

$$f_5 = 200 - f_3 - f_4 = 200 - 60 - 100 = 40$$

Hence $f_3 = 60, f_4 = 100$ and $f_5 = 40$.

Remarks. 1. In case of irregularities in the distribution, or the maximum frequency being repeated or the maximum frequency occurring in the very beginning or at the end of the distribution, the modal class is determined by the method of grouping and the mode is obtained by using (2·7).

Sometimes mode is estimated from the mean and the median. For a symmetrical distribution, mean, median and mode coincide. If the distribution is *moderately asymmetrical,* the mean, median and mode obey the following empirical relationship (due to Karl Pearson) :

$$\text{Mean} - \text{Median} = \tfrac{1}{3}(\text{Mean} - \text{Mode})$$

$$\Rightarrow \quad \text{Mode} = 3\,\text{Median} - 2\,\text{Mean} \qquad ...(2{\cdot}8)$$

2. If the method of grouping gives the modal class which does not correspond to the maximum frequency, *i.e.,* the frequency of modal class is not the maximum frequency, then in some situations we may get, $2f_k - f_{k-1} - f_{k+1} = 0$. In such cases, the value of mode can be obtained by the formula :

$$\text{Mode} = l + \frac{h(f_k - f_{k-1})}{|f_k - f_{k-1}| + |f_k - f_{k+1}|}$$

2·7·1. Merits and Demerits of Mode

Merits. *(i)* Mode is readily comprehensible and easy to calculate. Like median, mode can be located in some cases merely by inspection.

(ii) Mode is not at all affected by extreme values.

(iii) Mode can be conveniently located even if the frequency distribution has class-intervals of unequal magnitude provided the modal class and the classes preceding and succeeding it are of the same magnitude. Open–end classes also do not pose any problem in the location of mode.

Demerits. *(i)* Mode is ill-defined. It is not always possible to find a clearly defined mode. In some cases, we may come across distributions with two modes. Such distributions are called *bi-modal*. If a distribution has more than two modes, it is said to be *multimodal*.

(ii) It is not based upon all the observations.

(iii) It is not capable of further mathematical treatment.

(iv) As compared with mean, mode is affected to a greater extent by fluctuations of sampling.

Uses. Mode is the average to be used to find the ideal size, *e.g.*, in business forecasting, in the manufacture of ready-made garments, shoes, etc.

2·8. Geometric Mean. Geometric mean of a set of n observations is the nth root of their product. Thus the geometric mean G, of n observations x_i, $i = 1, 2, ..., n$ is

$$G = (x_1 . x_2 . .... . x_n)^{1/n} \qquad ...(2·9)$$

The computation is facilitated by the use of logarithms. Taking logarithm of both sides, we get

$$\log G = \frac{1}{n}(\log x_1 + \log x_2 + ... + \log x_n) = \frac{1}{n}\sum_{i=1}^{n} \log x_i$$

$$\therefore \qquad G = \text{Antilog}\left[\frac{1}{n}\sum_{i=1}^{n} \log x_i\right] \qquad ...(2·9a)$$

In case of frequency distribution x_i | f_i, ($i = 1, 2, ..., n$) geometric mean, G is given by

$$G = \left[x_1^{f_1} . x_2^{f_2} ..... . x_n^{f_n} \right]^{\frac{1}{N}}, \text{ where } N = \sum_{i=1}^{n} f_i \qquad ...(2·10)$$

Taking logarithms of both sides, we get

$$\log G = \frac{1}{N}(f_1 \log x_1 + f_2 \log x_2 + ... + f_n \log x_n)$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \log x_i \qquad ...(2·10a)$$

Thus we see that logarithm of G is the arithmetic mean of the logarithms of the given values. From (2·10*a*), we get

$$G = \text{Antilog}\left(\frac{1}{N}\sum_{i=1}^{n} f_i \log x_i\right) \qquad ...(2\cdot10b)$$

In the case of grouped or continuous frequency distribution, x is taken to be the value corresponding to the mid-point of the class-intervals.

2·8·1. Merits and Demerits of Geometric Mean

Merits. *(i)* It is rigidly defined.

(ii) It is based upon all the observations.

(iii) It is suitable for further mathematical treatment. If n_1 and n_2 are the sizes, G_1 and G_2 the geometric means of two series respectively, the geometric mean G, of the combined series is given by

$$\log G = \frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2} \qquad ...(2\cdot11)$$

Proof. Let x_{1i} $(i = 1, 2, ..., n_1)$ and x_{2j} $(j = 1, 2, ..., n_2)$ be n_1 and n_2 items of two series respectively. Then by def.,

$$G_1 = (x_{11} \cdot x_{12} \ldots x_{1n1})^{1/n_1} \quad \Rightarrow \quad \log G_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \log x_{1i}$$

$$G_2 = (x_{21} \cdot x_{22} \ldots x_{2n2})^{1/n_2} \quad \Rightarrow \quad \log G_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \log x_{2j}$$

The geometric mean G of the combined series is given by

$$G = (x_{11} \cdot x_{12} \ldots x_{1n1} \cdot x_{21} \cdot x_{22} \ldots x_{2n2})^{1/(n_1+n_2)}$$

$$\therefore \quad \log G = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1} \log x_{1i} + \sum_{j=1}^{n_2} \log x_{2j}\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1 \log G_1 + n_2 \log G_2\right]$$

The result can be easily generalised to more than two series.

(iv) It is not affected much by fluctuations of sampling.

(v) It gives comparatively more weight to small items.

Demerits. *(i)* Because of its abstract mathematical character, geometric mean is not easy to understand and to calculate for a non-mathematics person.

(ii) If any one of the observations is zero, geometric mean becomes zero and if any one of the observations is negative, geometric mean becomes imaginary regardless of the magnitude of the other items.

Uses. Geometric mean is used –

(i) To find the rate of population growth and the rate of interest.

(ii) In the construction of index numbers.

Example 2·12. *Show that in finding the arithmetic mean of a set of readings on thermometer it does not matter whether we measure temperature in Centigrade or Fahrenheit, but that in finding the geometric mean it does matter which scale we use.* **[Patna Univ. B.Sc., 1991]**

Solution. Let $C_1, C_2, ..., C_n$ be the n readings on the Centigrade thermometer. Then their arithmetic mean $\overline{C}$ is given by :

$$\overline{C} = \frac{1}{n}(C_1 + C_2 + \ldots + C_n)$$

If F and C be the readings in Fahrenheit and Centigrade respectively then we have the relation :

$$\frac{F - 32}{180} = \frac{C}{100} \quad \Rightarrow \quad F = 32 + \frac{9}{5}C.$$

Thus the Fahrenheit equivalents of $C_1, C_2, ..., C_n$ are

$$32 + \frac{9}{5}C_1,\ 32 + \frac{9}{5}C_2,\ \ldots,\ 32 + \frac{9}{5}C_n,$$

respectively.

Hence the arithmetic mean of the readings in Fahrenheit is

$$\begin{aligned}\overline{F} &= \frac{1}{n}\left\{\left(32 + \frac{9}{5}C_1\right) + \left(32 + \frac{9}{5}C_2\right) + \ldots + \left(32 + \frac{9}{5}C_n\right)\right\} \\ &= \frac{1}{n}\left\{32n + \frac{9}{5}(C_1 + C_2 + \ldots + C_n)\right\} \\ &= 32 + \frac{9}{5}\left(\frac{C_1 + C_2 + \ldots + C_n}{n}\right) \\ &= 32 + \frac{9}{5}\overline{C}.\end{aligned}$$

which is the Fahrenheit equivalent of $\overline{C}$.

Hence in finding the arithmetic mean of a set of n readings on a thermometer, it is immaterial whether we measure temperature in Centigrade or Fahrenheit.

Geometric mean G, of n readings in Centigrade is

$$G = (C_1 . C_2 \ldots C_n)^{1/n}$$

Geometric mean G_1, (say), of Fahrenheit equivalents of $C_1, C_2, ..., C_n$ is

$$G_1 = \left\{\left(32 + \frac{9}{5}C_1\right)\left(32 + \frac{9}{5}C_2\right)\ldots\left(32 + \frac{9}{5}C_n\right)\right\}^{1/n}$$

which is not equal to Fahrenheit equivalent of G, *viz.*,

$$\left\{\frac{9}{5}(C_1 . C_2 \ \ldots \ C_n)^{1/n} + 32\right\}$$

Hence in finding the geometric mean of the n readings on a thermometer, the scale (Centrigrade or Fahrenheit) is important.

2·9. Harmonic Mean. Harmonic mean of a number of observations is the reciprocal of the arithmetic mean of the reciprocals of the given values. Thus, harmonic mean H, of n observations x_i, $i = 1, 2, ..., n$ is

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}(1/x_i)} \qquad ...(2·12)$$

In case of frequency distribution $x_i \mid f_i$, $(i = 1, 2, ..., n)$,

$$H = \frac{1}{\frac{1}{N}\sum_{i=1}^{n}(f_i/x_i)}, \quad \left[N = \sum_{i=1}^{n} f_i\right] \qquad ...(2·12a)$$

2·9·1. Merits and Demerits of Harmonic Mean

Merits. Harmonic mean is rigidly defined, based upon all the observations and is suitable for further mathematical treatment. Like geometric mean, it is not affected much by fluctuations of sampling. It gives greater importance to small items and is useful only when small items have to be given a greater weightage.

Demerits. Harmonic mean is not easily understood and is difficult to compute.

Example 2·13. *A cyclist pedals from his house to his college at a speed of 10 m.p.h. and back from the college to his house at 15 m.p.h. Find the average speed.*

Solution. Let the distance from the house to the college be x miles. In going from house to college, the distance (x miles) is covered in $\frac{x}{10}$ hours, while in coming from college to house, the distance is covered in $\frac{x}{15}$ hours. Thus a total distance of $2x$ miles is covered in $\left(\frac{x}{10} + \frac{x}{15}\right)$ hours.

$$\text{Hence average speed} = \frac{\text{Total distance travelled}}{\text{Total time taken}} = \frac{2x}{\left(\frac{x}{10} + \frac{x}{15}\right)}$$

$$= \frac{2}{\left(\frac{1}{10} + \frac{1}{15}\right)} = 12 \text{ m.p.h.}$$

Remark. 1. In this case the average speed is given by the harmonic mean of 10 and 15 and not by the airthmetic mean.

Rather, we have the following general result :

If equal distances are covered (travelled) per unit of time with speeds equal to $V_1, V_2, ..., V_n$, say, then the average speed is given by the harmonic mean of $V_1, V_2, ..., V_n$, *i.e.*,

$$\text{Average speed} = \frac{n}{\left(\frac{1}{V_1} + \frac{1}{V_2} + ... + \frac{1}{V_n}\right)} = \frac{n}{\Sigma\left(\frac{1}{V}\right)}$$

Proof is left as an exercise to the reader.

Hint. Speed $= \frac{\text{Distance}}{\text{Time}} \Rightarrow$ Time $= \frac{\text{Distance}}{\text{Speed}}$

Average Speed $= \frac{\text{Total distance travelled}}{\text{Total time taken}}$

2. Weighted Harmonic Mean. Instead of fixed (constant) distance being travelled with varying speed, let us now suppose that different distances, say, $S_1, S_2, \ldots, S_n$, are travelled with different speeds, say, $V_1, V_2, \ldots, V_n$ respectively. In that case, the average speed is given by the weighted harmonic mean of the speeds, the weights being the corresponding distances travelled, *i.e.*,

$$\text{Average speed} = \frac{S_1 + S_2 + \ldots + S_n}{\left(\frac{S_1}{V_1} + \frac{S_2}{V_2} + \ldots + \frac{S_n}{V_n}\right)} = \frac{\Sigma S}{\Sigma\left(\frac{S}{V}\right)}$$

Example 2·14. *You can take a trip which entails travelling 900 km. by train at an average speed of 60 km. per hour, 3000 km. by boat at an average of 25 km. p.h., 400 km. by plane at 350 km. per hour and finally 15 km. by taxi at 25 km. per hour. What is your average speed for the entire distance ?*

Solution. Since different distances are covered with varying speeds , the required average speed for the entire distance is given by the weighted harmonic mean of the speeds (in km.p.h.), the weights being the corresponding distances covered (in kms.).

COMPUTATION OF WEIGHTED H. M.

Speed (km. / hr.) X	*Distance (in km.)* W	W / X
60	900	15·00
25	3000	120·00
350	400	1·43
25	15	0·60
Total	$\Sigma W = 4315$	$\Sigma (W/X) = 137·03$

Average speed

$$= \frac{\Sigma W}{\Sigma (W/X)}$$

$$= \frac{4315}{137·03}$$

$= 31·489$ *km.p.h.*

2·10. Selection of an Average. From the preceding discussion it is evident that no single average is suitable for all practical purposes. Each one of the average has its own merits and demerits and thus its own particular field of importance and utility. We cannot use the averages indiscriminately. A judicious selection of the average depending on the nature of the data and the purpose of the enquiry is essential for sound statistical analysis. Since arithmetic mean satisfies all the properties of an ideal average as laid down by Prof. Yule, is familiar to a layman and further has wide applications in statistical theory at large, it may be regarded as the best of all the averages.

2·11. Partition Values. These are the values which divide the series into a number of equal parts.

The three points which divide the series into four equal parts are called *quartiles*. The first, second and third points are known as the first, second and third quartiles respectively. The first quartile, Q_1, is the value which exceed 25% of the observations and is exceeded by 75% of the observations. The second quartile, Q_2, coincides with median. The third quartile, Q_3, is the point which has 75% observations before it and 25% observations after it.

The nine points which divide the series into ten equal parts are called *deciles* whereas *percentiles* are the ninety-nine points which divide the series into hundred equal parts. For example, D_7, the seventh decile, has 70% observations before it and P_{47}, the forty-seventh percentile, is the point which exceed 47% of the observations. The methods of computing the partition values are the same as those of locating the median in the case of both discrete and continuous distributions.

Example 2·15. *Eight coins were tossed together and the number of heads resulting was noted. The operation was repeated 256 times and the frequencies (f) that were obtained for different values of x, the number of heads, are shown in the following table. Calculate median, quartiles, 4th decile and 27th precentile.*

x:	0	1	2	3	4	5	6	7	8
f:	1	9	26	59	72	52	29	7	1

Solution.

x:	0	1	2	3	4	5	6	7	8
f:	1	9	26	59	72	52	29	7	1
c.f. :	1	10	36	95	167	219	248	255	256

Median : Here $N/2 = 256/2 = 128$. Cumulative frequency (*c.f.*) just greater than 128 is 167. Thus, median = 4.

Q_1 : Here $N/4 = 64$. *c.f.* just greater than 64 is 95. Hence, $Q_1 = 3$.

Q_3 : Here $3N/4 = 192$ and *c.f.* just greater than 192 is 219. Thus $Q_3 = 5$.

D_4 : $\frac{4N}{10} = 4 \times 25{\cdot}6 = 102{\cdot}4$ and *c.f.* just greater than 102·4 is 167. Hence $D_4 = 4$.

P_{27} : $\frac{27N}{100} = 27 \times 2{\cdot}56 = 69{\cdot}12$ and *c.f.* just greater than 69·12 is 95. Hence $P_{27} = 3$.

2·11·1. Graphical Location of the Partition Values. The partition values, *viz.*, quartiles, deciles and percentiles, can be conveniently located with the help of a curve called the 'cumulative frequency curve' or 'Ogive'. The procedure is illustrated below.

First form the cumulative frequency table. Take the class intervals (or the variate values) along the x–axis and plot the corresponding cumulative frequencies along the y–axis against the *upper limit* of the class interval (or against the variate value in the case of discrete frequency distribution). The curve obtained on joining

the points so obtained by means of free hand drawing is called the *cumulative frequency curve* or *ogive*. The graphical location of partition values from this curve is explained below by means of an example.

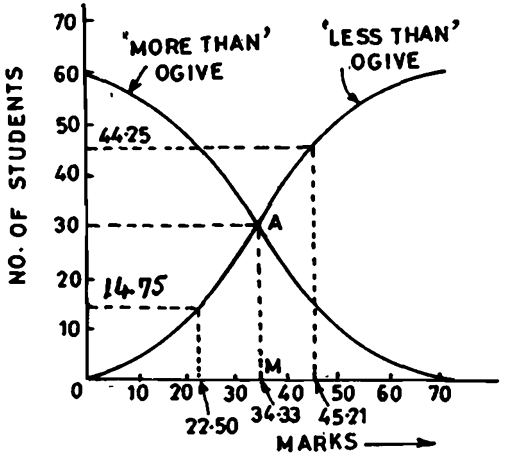
Example 2·16. *Draw the cumulative frequency curve for the following distribution showing the number of marks of 59 students in Statistics.*

Marks-group :	0—10	10—20	20—30	30—40	40—50	50—60	60—70
No. of Students :	4	8	11	15	12	6	3

Solution.

Marks-group	*No. of Students*	*Less than c.f.*	*More than c.f.*
0—10	4	4	59
10—20	8	12	55
20—30	11	23	47
30—40	15	38	36
40—50	12	50	21
50—60	6	56	9
60—70	3	59	3

Taking the marks-group along x-axis and *c.f.* along y-axis, we plot the cumulative frequencies, *viz.*, 4, 12, 23, ..., 59 against the upper limits of the corresponding classes, *viz.*, 10, 20, ..., 70 respectively. The smooth curve obtained on joining these points is called *ogive* or more particularly *'less than' ogive.*



If we plot the 'more than' cumulative frequencies, *viz.*, 59, 55,..., 3 against the lower limits of the corresponding classes, *viz.*, 0, 10, ..., 60 and join the points by a smooth curve, we get cumulative frequency curve which is also known as *ogive* or more particularly *'more than' ogive.*

To locate graphically the value of median, mark a point corresponding to $N/2$ along y-axis. At this point draw a line parallel to x-axis meeting the ogive at the point 'A' (say). From 'A' draw a line prependicular to x-axis meeting it in 'M' (say). Then abscissa of 'M' gives the value of median.

To locate the values of Q_1 (or Q_3), we mark the points along y-axis corresponding to $N/4$ (or $3N/4$) and proceed exactly similarly.

In the above example, we get from ogive

Median= 34·33, Q_1 = 22·50, and Q_3 = 45·21.

Remarks. 1. The median can also be located as follows :

From the point of intersection of 'less than' ogive and 'more than' ogive, draw perpendicular to OX. The abscissa of the point so obtained gives median.

2. Other partition values, *viz.*, deciles and percentiles, can be similarly located from 'ogive'.

EXERCISE

1. *(a)* What are grouped and ungrouped frequency distributions? What are their uses? What are the considerations that one has to bear in mind while forming the frequency distribution?

(b) Explain the method of constructing Histogram and Frequency Polygon. Which, out of these two, is better representative of frequencies of *(i)* a particular group, and *(ii)* whole group.

2. What are the principles governing the choice of :

(i) Number of class intervals,

(ii) The length of the class interval,

(iii) The mid-point of the class interval.

3. Write short notes on :

(i) Frequency distribution,

(ii) Histogram, frequency polygon and frequency curve,

(iii) Ogive.

4. *(a)* What are the properties of a good average? Examine these properties with reference to the Arithmetic Mean, the Geometric Mean and the Harmonic Mean, and give an example of situations in which each of them can be the appropriate measure for the average.

(b) Compare mean, median and mode as measures of location of a distribution.

(c) The mean is the most common measure of central tendency of the data. It satisfies almost all the requirements of a good average. The median is also an average, but it does not statisfy all the requirements of a good average. However, it carries certain merits and hence is useful in particular fields. Critically examine both the averages.

(d) Describe the different measures of central tendency of a frequency distribution, mentioning their merits and demerits.

5. Define *(i)* arithmetic mean, *(ii)* geometric mean and *(iii)* harmonic mean of grouped and ungrouped data. Compare and contrast the merits and demerits of them. Show that the geometric mean is capable of further mathematical treatment.

6. *(a)* When is an average a meaningful statistics? What are the requisites of a statisfactory average? In this light compare the relative merits and demerits of three well-known averages.

(b) What are the chief measures of central tendency? Discuss their merits.

7. Show that *(i)* Sum of deviations about arithmetic mean is zero.

(ii) Sum of absolute deviations about median is least.

(iii) Sum of the squares of deviations about arithmetic mean is least.

8. The following numbers give the weights of 55 students of a class. Prepare a suitable frequency table.

42	74	40	60	82	115	41	61	75	83	63
53	110	76	84	50	67	65	78	77	56	95
68	69	104	80	79	79	54	73	59	81	100
66	49	77	90	84	76	42	64	69	70	80
72	50	79	52	103	96	51	86	78	94	71

(i) Draw the histogram and frequency polygon of the above data.

(ii) For the above weights, prepare a cumulative frequency table and draw the less than ogive.

9. *(a)* What are the points to be borne in mind in the formation of frequency table?

Choosing appropriate class-intervals, form a frequency table for the following data:

10·2	0·5	5·2	6·1	3·1	6·7	8·9	7·2	8·9
5·4	3·6	9·2	6·1	7·3	2·0	1·3	6·4	8·0
4·3	4·7	12·4	8·6	13·1	3·2	9·5	7·6	4·0
5·1	8·1	1·1	11·5	3·1	6·8	7·0	8·2	2·0
3·1	6·5	11·2	12·0	5·1	10·9	11·2	8·5	2·3
3·4	5·2	10·7	4·9	6·2				

(b) What are the considerations one has to bear in mind while forming a frequency distribution?

A sample consists of 34 observations recorded correct to the nearest integer, ranging in value from 201 to 337. If it is decided to use seven classes of width 20 integers and to begin the first class at 199·5, find the class limits and class marks of the seven classes.

(c) The class marks in a frequency table (of whole numbers) are given to be 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50. Find out the following :

(i) the true classes.

(ii) the true class limits.

(iii) the true upper class limits.

10. *(a)* The following table shows the distribution of the number of students per teacher in 750 colleges :-

Students :	1	4	7	10	13	16	19	22	25	28
Frequency :	7	46	165	195	189	89	28	19	9	3

Draw the histogram for the data and superimpose on it the frequency polygon.

(b) Draw the histogram and frequency curve for the following data.

Monthly wages in Rs.	10–13	13–15	15–17	17–19	19–21	21–23	23–25
No. of workers	6	53	85	56	21	16	8

(c) Draw a histogram for the following data :

Age (in years) :	2–5	5–11	11–12	12–14	14–15	15–16
No. of boys :	6	6	2	5	1	3

11. *(a)* Three people *A, B, C* were given the job of finding the average of 5000 numbers. Each one did his own simplification. *A*'s method : Divide the sets into sets of 1000 each, calculate the average in each set and then calculate the average of these averages. *B*'s method : Divide the set into 2,000 and 3,000 numbers, take average in each set and then take the average of the averages. *C*'s method :500 numbers were unities. He averaged all other numbers and then added one. Are these methods correct?

Ans. Correct, not correct, not correct.

(b) The total sale (in '000 rupees) of a particular item in a shop, on 10 consecutive days, is reported by a clerk as, 35·00, 29·60, 38·00, 30·00, 40·00, 41·00, 42·00, 45·00, 3·60, 3·80. Calculate the average. Later it was found that there was a number 10·00 in the machine and the reports of 4th to 8th days were 10·00 more than the true values and in the last 2 days he put a decimal in the wrong place thus for example 3·60 was really 36·0. Calculate the true mean value.

Ans. 30·8, 32·46.

12. *(a)* Given below is the distribution of 140 candidates obtaining marks *X* or higher in a certain examination (all marks are given in whole numbers) :

X :	10	20	30	40	50	60	70	80	90	100
c.f. :	140	133	118	100	75	45	25	9	2	0

Calculate the mean, median and mode of the distribution.

Hint.

Class	*Frequency (f)*	*Class boundaries*	*Mid value*	*c.f. (less than)*
10—19	140 – 133 = 7	9·5—19·5	14·5	7
20—29	133 – 118 = 15	19·5—29·5	24·5	22
30—39	118 – 100 = 18	29·5—39·5	34·5	40
40—49	100 – 75 = 25	39·5—49·5	44·5	65
50—59	75 – 45 = 30	49·5—59·5	54·5	95
60—69	45 – 25 = 20	59·5—69·5	64·5	115
70—79	25 – 9 = 16	69·5—79·5	74·5	131
80—89	9 – 2 = 7	79·5—89·5	84·5	138
90—99	2 – 0 = 2	89·5—99·5	94·5	140

$$\text{Mean} = 54{\cdot}5 + \frac{10 \times (-53)}{140} = 50{\cdot}714$$

$$\text{Median} = 49{\cdot}5 + \frac{10}{30}\left(\frac{140}{2} - 65\right) = 51{\cdot}167$$

(b) The four parts of a distribution are as follows :

Part	*Frequency*	*Mean*
1	50	61
2	100	70
3	120	80
4	30	83

Find the mean of the distribution. **(Madurai Univ. B.Sc., 1988)**

13. *(a)* Define a 'weighted mean'. If several sets of observations are combined into a single set, show that the mean of the combined set is the weighted mean of several sets.

(b) The weighted geometric mean of three numbers 229, 275 and 125 is 203 The weights for the first and second numbers are 2 and 4 respectively. Find the weight of third. Ans. 3.

14. Define the weighted arithmetic mean of a set of numbers. Show that it is unaffected if all weights are multiplied by some common factor.

The following table shows some data collected for the regions of a country:

Region	*Number of inhabitants (million)*	*Percentage of literates*	*Average annual income per person (Rs.)*
A	10	52	850
B	5	68	620
C	18	39	730

Obtain the overall figures for the three regions taken together. Prove the formulae you use. **[Calcutta Univ. B.A.(Hons.), 1991]**

15. Draw the Ogives and hence estimate the median.

Class	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79
Frequency	8	32	142	216	240	206	143	13

16. The following data relate to the ages of a group of workers in a factory.

Ages	*No. of workers*	*Ages*	*No. of workers*
20—25	35	40—45	90
25—30	45	45—50	74
30—35	70	50—55	51
35—40	105	55—60	30

Draw the percentage cumulative curve and find from the graph the number of workers between the ages 28—48.

17. *(a)* The mean of marks obtained in an examination by a group of 100 students was found to be 49·96. The mean of the marks obtained in the same examination by another group of 200 students was 52·32. Find the mean of the marks obtained by both the groups of students taken together.

(b) A distribution consists of three components with frequencies 300, 200 and 600 having their means 16, 8 and 4 respectively. Find the mean of the combined distribution.

(c) The mean marks got by 300 students in the subject of Statistics are 45. The mean of the top 100 of them was found to be 70 and the mean of the last 100 was known to be 20. What is the mean of the remaining 100 students?

(d) The mean weight of 150 students in a certain class is 60 kilograms. The mean weight of boys in the class is 70 kilograms and that of the girls is 55 kilograms. Find the number of boys and number of girls in the class.

Ans. *(a)* 51·53, *(b)* 8, *(c)* 45, *(d)* Boys = 50, Girls = 100.

18. From the following data, calculate the percentage of workers getting wages

(a) more than Rs. 44, *(b)* between Rs. 22 and Rs. 58, *(c)* Find Q_1 and Q_3.

Wages (Rs.)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
No. of workers	20	45	85	160	70	55	35	30

Hint. Assuming that frequencies are uniformly distributed over the entire interval,

(a) Number of persons with wages more than Rs. 44 is

$$\left(\frac{50-44}{10}\times 70\right)+55+35+30=162$$

Hence the percentage of workers getting over Rs. 44 is

$$=\frac{162}{500}\times 100=32{\cdot}4\%$$

(b) Percentage of workers getting wages between Rs. 22 and Rs. 58 is

$$\left[\left(\frac{30-22}{10}\times 85\right)+160+70+\left(\frac{58-50}{10}\times 55\right)\right]\times 100\div 500=68{\cdot}4\%$$

19. For the two frequency distributions give below the mean calculated from the first was 25·4 and that from the second was 32·5. Find the values of x and y.

Class	*Distribution I* *Frequency*	*Distribution II* *Frequency*
10—20	20	4
20—30	15	8
30—40	10	4
40—50	x	$2x$
50—60	y	y

Ans. $x=3,\ y=2$

20. A number of particular articles has been classified according to their weights. After drying for two weeks the same articles have again been weighted and similarly classified. It is known that the median weight in the first weighing was 20·83 oz. while in the second weighing it was 17·35 oz. Some frequencies a and b in the first weighing and x and y in the second are missing. It is known that $a = \frac{1}{3}x$ and $b = \frac{1}{2}y$. Find out the values of the missing frequencies.

Class	*Frequencies*		*Class*	*Frequencies*	
	1st weighing	*IInd weighing*		*1st weighing*	*IInd weighing*
0—5	a	x	15—20	52	50
5—10	b	y	20—25	75	30
10—15	11	40	25—30	22	28

Hint. We have $x = 3a$, $y = 2b$,

N_1 = Total frequency in 1st weighing $= 160 + a + b$.

N_2 = Total frequency in 2nd weighing $= 148 + x + y = 148 + 3a + 2b$.

Using Median formula, we shall get

$$20{\cdot}83 = 20 + \frac{5}{75}\left[\frac{N_1}{2} - (63 + a + b)\right]$$

$$\Rightarrow \quad 15\,(20{\cdot}83 - 20) = \frac{160 + a + b}{2} - (63 + a + b)$$

$$\Rightarrow \quad 12{\cdot}45 = 17 - \frac{a + b}{2}$$

$$\Rightarrow \quad a + b = 2\,(17 - 12{\cdot}45) = 9{\cdot}10 \simeq 9 \qquad \ldots(*)$$

Since a and b, being frequencies are integral valued, $a + b$ is also integral valued. Now the median of 2nd weighing gives:

$$17{\cdot}35 = 15 + \frac{5}{50}\left[\frac{148 + 3a + 2b}{2} - (40 + x + y)\right]$$

$$\Rightarrow \quad 10 \times 2{\cdot}35 = 74 + \frac{3a + 2b}{2} - 40 - 3a - 2b$$

$$\Rightarrow \quad \frac{3a + 2b}{2} = 34 - 23{\cdot}5 = 10{\cdot}5$$

$$\Rightarrow \quad 3a + 2b = 21 \qquad \ldots(**)$$

Multiplying (*) by 3, we get

$$3a + 3b = 27 \qquad \ldots(***)$$

Subtracting (**) from (***), we get $b = 6$. Substituting in (*), we get $a = 9 - 6 = 3$.

$$\therefore \quad a = 3,\ b = 6;\ x = 3a = 9,\ y = 2b = 12.$$

21. From the following table showing the wage distribution in a certain factory, determine :

(a) the mean wage,

(b) the median wage,

(c) the modal wage,

(d) the wage limits for the middle 50% of the wage earners,

(e) the percentages of workers who earned between Rs.75 and Rs.125.

(f) the percentage who earned more than Rs.150 per week, and

(g) the percentage who earned less than Rs.100 per week.

Weekly wages (Rs.)	*No. of employees*	*Weekly wages (Rs.)*	*No. of employees*
20–40	8	120–140	35
40–60	12	140–160	18
60–80	20	160–180	7
80–100	30	180–200	5
100–120	40		

Ans. *(a)* $\overline{X} = 108{\cdot}5$, *(b)* Med. = 108·75, *(c)* Mo = 118·3, *(d)* 81·25 , 129·3 *(e)* 48, *(f)* 12, *(g)* 40.

22. *(a)* Explain how the ogives are drawn for any frequency distribution. Point out the method of finding out the values of median, mode, quartiles, deciles and percentiles graphically. Also, write down the formula for the computation of each of them for any frequency distribution.

(b) The following table gives the frequency distribution of marks in a class of 65 students.

Marks	*No. of Students*	*Marks*	*No. of students*
0–4	10	14–18	5
4–8	12	18–20	3
8–12	18	20–25	4
12–14	7	25 and over	6
Total			65

Calculate : *(i)* Upper and lower quartiles.

(ii) No. of students who secured marks more than 17.

(iii) No. of students who secured marks between 10 and 15.

(c) The following table shows the age distribution of heads of families in a certain country during the year 1957. Find the median, the third quartile and the second decile of the distribution. Check your results by the graphical method.

Age of head of family years	Under 25	25–29	30–34	35–44	45–54	55–64	65–74	above 74
Number (million)	2·3	4·1	5·3	10·6	9·7	6·8	4·4	1·8

Total 45

Ans. Md = 45·2 yrs.; Q_3 = 57·5 yrs.; D_2 = 32·5 yrs.

23. The following data represent travel expenses (other than transportation) for 7 trips made during November by a salesman for a small firm :

Trip	*Days*	*Expense (Rs.)*	*Expense per day (Rs.)*
1	0·5	13·50	27
2	2·0	12·00	6
3	3·5	17·50	5
4	1·0	9·00	9
5	9·0	27·00	3
6	0·5	9·00	18
7	8·5	17·00	2
Total	25·0	105·00	70

An auditor criticised these expenses as excessive, asserting that the average expense per day is Rs. 10 (Rs. 70 divided by 7). The salesman replied that the average is only Rs. 4·20 (Rs. 105 divided by 25) and that in any event the median is the appropriate measure and is only Rs. 3. The auditor rejoined that the arithmetic mean is the appropriate measure, but that the median is Rs. 6.

You are required to :

(a) Explain the proper interpretation of each of the four averages mentioned.

(b) Which average seems appropriate to you ?

24. *(a)* Define Geometric and Harmonic means and explain their uses in statistical analysis.

You take a trip which entails travelling 900 miles by train at an average speed of 60 m.p.h., 300 miles by boat at an average of 25 m.p.h., 400 miles by plane at 350 m.p.h. and finally 15 miles by taxi at 25 m.p.h. What is your speed for the entire distance?

(b) A train runs 25 miles at a speed of 30 m.p.h., another 50 miles at a speed of 40 m.p.h., then due to repairs of the track travels for 6 minutes at a speed of 10 m.p.h. and finally covers the remaining distance of 24 miles at a speed of 24 m.p.h. What is the average speed in m.p.h.?

(c) A man motors from A to B. A large part of the distance is uphill and he gets a mileage of only 10 per gallon of gasoline. On the return trip he makes 15 miles per gallon. Find the harmonic mean of his mileage. Verify the fact that this is the proper average to be used by assuming that the distance from A to B is 60 miles.

(d) Calculate the average speed of a car running at the rate of 15 km.p.h. during the first 30 kms., at 20 km.p.h. during the second 30 kms. and at 25 kmp.h. during the third 30 kms.

25. The following table shows the distribution of 100 families according to their expenditure per week. Number of families corresponding to expenditure groups Rs. (10—20) and Rs.(30—40) are missing from the table. The median and

mode are given to be Rs.25 and 24 Calculate the missing frequencies and then arithmetic mean of the data :

Expenditure :	0—10	10—20	20—30	30—40	40—50
No. of families :	14	?	27	?	15

Hint.

Expenditure	*No. of Families*	*Cumulative frequencies*
0—10	14	14
10—20	f_1	$14+f_1$
20—30	27	$41+f_1$
30—40	f_2	$41+f_1+f_2$
40—50	15	$56+f_1+f_2$

$$\therefore \qquad 25 = 20 + \frac{\frac{56+f_1+f_2}{2} - (14+f_1)}{27} \times 10$$

and
$$24 = 20 + \frac{27 - f_1}{2 \times 27 - f_1 - f_2} \times 10$$

Simplying these equations, we get

$$f_1 - f_2 = 1$$

and
$$3f_1 - 2f_2 = 27.$$

Ans. 25, 24

26. *(a)* The numbers 3·2, 5·8, 7·9 and 4·5 have frequencies x, $(x+2)$, $(x-3)$ and $(x+6)$ respectively. If their arithmetic mean is 4·876, find the value of x.

(b) If $M_{g.x}$ is the geometric mean of N x's amd $M_{g.y}$ is the geometric mean of N y's, then the geometric mean M_g of the $2N$ values is given by

$$M_g^2 = M_{g.x} M_{g.y}.$$ **(Nagpur Univ. B.Sc., 1990)**

(c) The weighted geometric mean of the three numbers 229, 275 and 125 is 203. The weights for the first and the second numbers are 2 and 4 respectively. Find the weight of the third. **Ans.** 3.

27. The geometric mean of 10 observations on a certain variable was calculated as 16·2. It was later discovered that one of the observations was wrongly recorded as 12·9; in fact it was 21·9. Apply appropriate correction and calculate the correct geometric mean.

Hint. Correct value of the geometic mean, G' is given by

$$G' = \left(\frac{(16{\cdot}2)^{10} \times 21{\cdot}9}{12{\cdot}9} \right)^{1/10} = 17{\cdot}68$$

28. A variate takes the values $a, ar, ar^2, ..., ar^{n-1}$ each with frequency unity. If A, G and H are respectively the arithmetic mean, geometric mean and harmonic mean, show that

$$A = \frac{a(1-r^n)}{n(1-r)},\ G = ar^{(n-1)/2},\ H = \frac{an(1-r)r^{n-1}}{(1-r^n)}$$

Prove that $G^2 = AH$. Prove also that $A > G > H$ unless $r = 1$, when all the three means coincide.

29. If $\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{x}_2 = \frac{1}{n}\sum_{i=2}^{n+1} x_i$ and $\bar{x}_3 = \frac{1}{n}\sum_{i=3}^{n+2} x_i$

then show that

(*a*) $\bar{x}_2 = \bar{x}_1 + \frac{1}{n}(x_{n+1} - x_1)$, and (*b*) $\bar{x}_3 = \bar{x}_2 + \frac{1}{n}(x_{n+2} - x_2)$

30. A distribution $x_1, x_2, \ldots, x_n$ with frequencies $f_1, f_2 \ldots, f_n$ transformed into the distribution $X_1, X_2, \ldots, X_n$ with the same corresponding frequencies by the relation $X_r = ax_r + b$, where a and b are constants. Show that the mean, median and mode of the new distribution are given in terms of those of the first distribution by the same transformation. **[Kanpur Univ. B.Sc., 1992]**

Use the method indicated above to find the mean of the following distribution: x (duration of telephone conversation in seconds)

49·5, 149·5, 249·5, 349·5, 449·5, 549·5, 649·5, 749·5, 849·5, 949·5

f (respective frequency)

6	28	88	180	247	260	132	48	11	5

31. If $\bar{x}_w$ is the weighted mean of x_i's with weights w_i, prove that

$$\left(\sum_{i=1}^{n} w_i\right)\left(\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2\right) = \sum_{i=1}^{n}\sum_{j>i}^{n} w_i w_j (x_i - x_j)^2, \text{ where } \sum_{i=1}^{n} w_i \neq 0.$$

(Allahabad Univ. B.Sc., 1992)

Hint. $\left[\sum_{i=1}^{n}\sum_{j>i} w_i w_j (x_i - x_j)^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=i}^{n} w_i w_j (x_i - x_j)^2\right.$

$$\left. = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=i}^{n} w_i w_j \left\{\left(x_i - \bar{x}_w\right) - \left(x_j - \bar{x}_w\right)\right\}^2\right]$$

32. In a frequency table, the upper boundary of each class interval has a constant ratio to the lower boundary. Show that the geometric mean G may be expressed by the formula :

$$\log G = x_0 + \frac{c}{N}\sum_i f_i (i-1)$$

where x_0 is the logarithm of the mid-value of the first interval and c is the logarithm of the ratio between upper and lower boundaries.

[Delhi Univ. B.Sc. (Stat. Hons.), 1990, 1986]

33. Find the minimum value of :

(i) $f(x) = (x-6)^2 + (x+3)^2 + (x-8)^2 + (x+4)^2 + (x-3)^2$

(ii) $g(x) = |x-6| + |x+3| + |x-8| + |x+4| + |x-3|$.

[Delhi Univ. B.Sc.(Stat. Hons.), 1991]

Hint. The sum of squares of deviations is minimum when taken from arithmetic mean and the sum of absolute deviations is minimum when taken from median.

34. If A, G and H be the arithmetic mean, geometric mean and harmonic mean respectively of two positive numbers a and b, then prove that :

(i) $A \geq G \geq H$.

When does the equality sign hold?

(ii) $G^2 = AH$.

35 Calculate simple and weighted arithmetic averages from the following data and comment on them :

Designation	*Monthly salary (in Rs.)*	*Strength of the cadre*
Class I Officers	1,500	10
Class II Officers	800	20
Subordinate staff	500	70
Clerical staff	250	100
Lower staff	100	150

Ans. $\bar{X}$ = Rs. 630, $\bar{X}_w$ = Rs. 302·86. Latter is more representative.

36. Treating the number of letters in each word in the following passage as the variable x, prepare the frequency distribution table and obtain its mean, median, mode.

"The reliability of data must always be examined before any attempt is made to base conclusions upon them. This is true of all data, but particularly so of numerical data, which do not carry their quality written large on them. It is a waste of time to apply the refined theoretical methods of Statistics to data which are suspect from the beginning."

Ans. Mean = 4·565, Median = 4, Mode = 3.

OBJECTIVE TYPE QUESTIONS

I. Match the correct parts to make a valid statement :

(a) Arithmetic Mean	*(i)* $l + [f_2/(f_1 + f_2)] \times i$
(b) Geometric Mean	*(ii)* $(x_1 . x_2 . \ldots . x_n)^{1/n}$
(c) Harmonic Mean	*(iii)* $\Sigma fX/\Sigma f$
(d) Median	*(iv)* $l + \dfrac{N/2 - c.f.}{f} \times i$

(*e*) Mode

(*v*) $\left[\frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}\right)\right]^{-1}$

(*vi*) $l=\frac{f_1-f_o}{2f_1-f_o-f_2}\times i$

II. Which measure of location will be suitable to compare:

(*i*) heights of students in two classes;

(*ii*) size of agricultural holdings;

(*iii*) average sales for various years;

(*iv*) intelligence of students;

(*v*) per capita income in several countries;

(*vi*) sale of shirts with collar size; 16″, $15\frac{1}{2}$″, 15″, 14″, 13″, 15″;

(*vii*) marks obtained 10, 8, 12, 4, 7, 11, and X ($X < 5$).

Ans. (*i*) Mean, (*ii*) Mode, (*iii*) Mean, (*iv*) Median, (*v*) Mean, (*vi*) Mode, (*vii*) Median.

III. Which of the following are true for all sets of data?

(*i*) Arithmetic Mean ≤ median ≤ mode,

(*ii*) Arithmetic mean ≥ median ≥ mode,

(*iii*) Arithmetic mean = median = mode

(*iv*) None of these.

IV. Which of the following are true in respect of any distribution?

(*i*) The percentile points are in the ascending order.

(*ii*) The percentile points are equispaced.

(*iii*) The median is the mid-point of the range and the distribution.

(*iv*) A unique median value exists for each and every distribution.

V. Find out the missing figures :

(*a*) Mean = ? (3 Median – Mode).

(*b*) Mean – Mode = ? (Mean – Median).

(*c*) Median = Mode + ? (Mean – Mode).

(*d*) Mode = Mean – ? (Mean – Median).

Ans. (*a*) 1/2, (*b*) 3, (*c*) 2/3, (*d*) 3.

VI. Fill in blanks :

(*i*) Harmonic mean of a number of observations is .....

(*ii*) The geometric mean of 2, 4, 16, and 32 is .....

(*iii*) The strength of 7 colleges in a city are 385; 1,748; 1,343; 1,935; 786; 2,874 and 2,108. Then the median strength is .....

(*iv*) The geometric mean of a set of values lies between arithmetic mean and...

(*v*) The mean and median of 100 items are 50 and 52 respectively. The value of the largest item is 100. It was later found that it is actually 110. Therefore, the true mean is ... and the true median is .....

(vi) The algebraic sum of the deviations of 20 observations measured from 30 is 2. Therefore, mean of these observations is .....

(vii) The relationship between A.M., G.M. and H.M. is .....

(viii) The mean of 20 observations is 15. On checking it was found that two observations were wrongly copied as 3 and 6. If wrong observations are replaced by correct values 8 and 4, then the correct mean is .....

(ix) Median = ... Quartile.

(x) Median is the average suited for ..... classes.

(xi) A distribution with two modes is called ..... and with more than two modes is called .....

(xii) ..... is not affected by extreme observations.

Ans. *(ii)* 8 ; *(iii)* 1,748 ; *(iv)* H.M. ; *(v)* 50·1, 52 ; *(vi)* 30·1 ; *(vii)* A.M. ≥ G.M. ≥ H.M. ; *(viii)* 15·15 ; *(ix)* Second ; *(x)* Open end ; *(xi)* Bimodal, multimodal ; *(xii)* Median or mode.

VII. For the questions given below, give correct answers.

(i) The algebraic sum of the deviations of a set of n values from their arithmetic mean is

(a) n, *(b)* 0, *(c)* 1, *(d)* none of these.

(ii) The most stable measure of central tendency is

(a) the mean, *(b)* the median, *(c)* the mode, *(d)* none of these.

(iii) 10 is the mean of a set of 7 observations and 5 is the mean of a set of 3 observations. The mean of a combined set is given by

(a) 15, *(b)* 10, *(c)* 8·5, *(d)* 7·5, *(e)* none of these.

(iv) The mean of the distribution, in which the value of x are 1, 2, ..., n, the frequency of each being unity is:

(a) $n(n+1)/2$, *(b)* $n/2$, *(c)* $(n+1)/2$, *(d)* none of these.

(v) The arithmetic mean of the numbers 1, 2, 3, ..., n is

(a) $\frac{n(n+1)(2n+1)}{6}$, *(b)* $\frac{n(n+1)^2}{4}$, *(c)* $\frac{n(n+1)}{2}$, *(d)* none of these.

(ii) The most stable measure of central tendency

(vi) The point of intersection of the 'less than' and the 'greater than' ogive corresponds to

(a) the mean, *(b)* the median, *(c)* the geometric mean, *(d)* none of these.

(vii) When x_i and y_i are two variables ($i = 1, 2, ..., n$) with G.M.'s G_1 and G_2 respectively then the geometric mean of $\left(\frac{x_i}{y_i}\right)$ is

(a) $\frac{G_1}{G_2}$, *(b)* $\text{antilog}\left(\frac{G_1}{G_2}\right)$, *(c)* $n(\log G_1 - \log G_2)$;

(d) $\text{Antilog}\left(\frac{\log G_1 - \log G_2}{2n}\right)$

Ans. *(i)* (b) ; *(ii)* (a) ; *(iii)* (c) ; *(iv)* (c) ; *(v)* (d) ; *(vi)* (b) ; *(vii)* (a).

VIII. State which of the following statements are True and which are False. In case of false statements give the correct statement.

(i) The harmonic mean of n numbers is the reciprocal of the Arithmetic mean of the reciprocals of the numbers.

(ii) For the wholesale manufacturers interested in the type which is usually in demand, median is the most suitable average.

(iii) The algebraic sum of the deviations of a series of individual observations from their mean is always zero.

(iv) Geometric mean is the appropriate average when emphasis in on the rate of change rather than the amount of change.

(v) Harmonic mean becomes zero when one of the items is zero.

(vi) Mean lies between median and mode.

(vii) Cumulative frequency is not-decreasing.

(viii) Geometric mean is the arithmetic mean of harmonic mean and arithmetic mean.

(ix) Mean, median mode have the same unit.

(x) One quintal of wheat was purchased at 0·8 kg. per rupee and another quintal at 1·2 kg. per rupee. The average rate per rupee is 1kg.

(xi) One limitation of the median is that it cannot be calculated from a frequency distribution with open end classes.

(xii) The arithmetic mean of a frequency distribution is always located in the class which has the greatest number of frequencies.

(xiii) In a moderately asymmetrical distribution, the mean, median and mode are the same.

(xiv) It is really immaterial in which class an item falling at the boundary between two classes is listed.

(xv) The median is not affected by extreme items.

(xvi) The median is the point about which the sum of squared deviations is minimum.

(xvii) In construction of the frequency distribution, the selection of the class interval is arbitrary.

(xviii) Usual attendance of B.Sc. class is 35 per day. So for 100 working days total attendance is 3,500.

(xix) A car travels 100 miles at a speed of 40 m.p.h. and another 400 miles at a speed of 30 m.p.h. So the average speed for the whole journey is either 35 m.p.h. or 33 m.p.h.

(xx) In calculating the mean for grouped data, the assumption is made that the mean of the items in each class is equal to the mid-value of the class.

(xxi) The geometric mean of a group of numbers is less than the arithmetic mean in all cases, except in the special case in which the numbers are all the same.

(xxii) The geometric mean equals the antilog of the arithmetic mean of the logs of the values.

(xxiii) The median may be considered more typical than the mean because the median is not affected by the size of the extremes.

(xxiv) The Harmonic Mean of a series of fractions is the same as the reciprocal of the arithmetic mean of the series.

(xxv) In a frequency distribution the true value of mode cannot be calculated exactly.

IX. In each of the following cases, explain whether the description applies to mean, median or both.

(i) it can be calculated from a frequency distribution with open-end classes.

(ii) the values of all items are taken into consideration in the calculation.

(iii) the values of extreme items do not intluence the average.

(iv) In a distribution with a single peak and moderate skewness to the right it is closer to the concentration of the distribution.

Ans. *(i)* median, *(ii)* mean, *(iii)* median, *(iv)* median.

X. Be brief in your answer :

(a) The production in an industrial unit was 10,000 units during 1981 and in 1980 the production was 25,000 units. Hence the production has declined by 150 percent. Comment.

(b) A man travels by a car for 4 days. He travelled for 10 hours each day. He drove on the first day at the rate of 45 km per hour, second day at 40 km. per hour, third day at the rate of 38 km. per hour and the fourth day at the rate of 37 km. per hour.

Which average, harmonic mean or arithmetic mean or median will give us his average speed? Why?

(c) It is seen from records that a country does not export more than 5 % of its total production. Hence export trade is not vital to the economy of that country. Is the conclusion right?

(d) A survey revealed that the children of engineers, doctors and lawyers have high intelligence quotients. It further revealed that the grandfathers of these children were also highly intelligent. Hence the inference is that intelligence is hereditary. Do you agree?

XI. Do you agree with the following interpretations made on the basis of the facts given. Explain briefly your answer.

(a) The number of deaths in military in the recent war was 10 out of 1,000 while the number of deaths in Hyderabad in the same period was 18 per 1,000. Hence it is safe to join military service than to live in the city of Hyderabad.

(b) The examination result in a college X was 70% in the year 1991. In the same year and at the same examination only 500 out of 750 students were successful in college Y. Hence the teaching standard in college X was better.

(c) The average daily production in a small-scale factory in January 1991 was 4,000 candles and 3,800 candles in February 1981. So the workers were more efficient in January.

(d) The increase in the price of a commodity was 25%. Then the price decreased by 20% and again increased by 10%. So the resultant increase in the price was $25 - 20 + 10 = 15$ %.

(e) The rate of tomato in the first week of January was 2 kg. for a rupee and in the 2nd week was 4 kg. for a rupee. Hence the average price of tomato is $\frac{1}{2}(2 + 4) = 3$ kg. for a rupee.

XII. *(a)* The mean mark of 100 students was given to be 40. It was found later that a mark 53 was read as 83. What is the corrected mean mark?

(b) The mean salary paid to 1,000 employees of an establishment was found to be Rs. 108·40. Later on, after disbursement of salary it was discovered that the salary of two employees was wrongly entered as Rs. 297 and Rs. 165. Their correct salaries were Rs. 197 and Rs. 185. Find the correct arithmetic mean.

(c) Twelve persons gambled on a certain night. Seven of them lost at an average rate of Rs. 10·50 while the remaining five gained at an average of Rs. 13·00. Is the information given above correct? If not, why?

CHAPTER THREE

Measures of Dispersion, Skewness and Kurtosis

3·1. Dispersion. Averages or the measures of central tendency give us an idea of the concentration of the observations about the central part of the distribution. If we know the average alone we cannot form a complete idea about the distribution as will be cear from the following example.

Consider the series *(i)* 7, 8, 10, 11, *(ii)* 3, 6, 9, 12, 15, *(iii)* 1, 5, 9, 13, 17. In all these cases we see that n, the number of observations is 5 and the mean is 9. If we are given that the mean of 5 observations is 9, we cannot form an idea as to whether it is the average of first series or second series or third series or of any other series of 5 observations whose sum is 45. Thus we see that the measures of central tendency are inadequate to give us a complete idea of the distribution. They must be supported and supplemented by some other measures. One such measure is *Dispersion.*

Literal meaning of dispersion is *'scatteredness'*. We study dispersion to have an idea about the homogeneity or heterogeneity of the distribution. In the above case we say that series *(i)* is more homogeneous (less dispersed) than the series *(ii)* or *(iii)* or we say that series *(iii)* is more heterogeneous (more scattered) than the series *(i)* or *(ii).*

3·2. Characteristics for an Ideal Measure of Dispersion. The desiderata for an ideal measure of dispersion are the same as those for an ideal measure of central tendency, *viz.,*

(i) It should he rigidly defined.

(ii) It should be easy to calculate and easy to understand.

(iii) It should be based on all the observations.

(iv) It should be amenable to further mathematical treatment.

(v) It should be affected as little as possible by fluctuations of sampling.

3·3. Measures of Dispersion. The following are the measures of dispersion:

(i) Range,

(ii) Quartile deviation or Semi-interquartile range,

(iii) Mean deviation, and

(iv) Standard deviation.

3·4. Range. The range is the difference between two extreme obsertations of the distribution. If A and B are the greatest and smallest observations respectively in a distribution, then its range is $A - B$.

Range is the simplest but a crude measure of dispersion. Since it is based on two extreme observations which themselves are subject to chance fluctuations, it is not at all a reliable measure of dispersion.

3·5. Quartile Deviation. Quartile deviation or semi-interquartile range

Q is given by

$$Q = \tfrac{1}{2}(Q_3 - Q_1), \qquad ...(3·1)$$

where Q_1 and Q_3 are the first and third quartiles of the distribution respectively.

Quartile deviation is definitely a better measure than the range as it makes use of 50% of the data. But since it ignores the other 50% of the data, it cannot be regarded as a reliable measure.

3·6. Mean Deviation. If $x_i \mid f_i$, $i = 1, 2, ..., n$ is the frequency distribution, then mean deviation from the average A, (usually mean, median or mode), is given by

$$\text{Mean deviation} = \frac{1}{N} \sum_i f_i \mid x_i - A \mid, \ \Sigma f_i = N \qquad ...(3·2)$$

where $\mid x_i - A \mid$ represents the modulus or the absolute value of the deviation $(x_i - A)$, when the –ive sign is ignored.

Since mean deviation is based on all the observations, it is a better measure of dispersion than range or quartile deviation. But the step of ignoring the signs of the deviations $(x_i - A)$ creates artificiality and renders it useless for further mathematical treatment.

It may be pointed out here that mean deviation is least when taken from median. (The proof is given for continuous variable in Chapter 5)

3·7. Standard Deviation and Root Mean Square Deviation. Standard deviation, usually denoted by the Greek letter small sigma (σ), is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. For the frequency distribution $x_i \mid f_i$, $i = 1, 2, ..., n$,

$$\sigma = \sqrt{\frac{1}{N} \sum_i f_i (x_i - \bar{x})^2} \qquad ...(3·3)$$

where $\bar{x}$ is the arithmetic mean of the distribution and $\sum_i f_i = N$.

The step of squaring the deviations $(x_i - \bar{x})$ overcomes the drawback of ignoring the signs in mean deviation. Standard deviation is also suitable for further mathematical treatment (§ 3·7·3). Moreover of all the measures, standard deviation is affected least by fluctuations of sampling.

Thus we see that standard deviation satisfies almost all the properties laid down for an ideal measure of dispersion except for the general nature of extracting the square root which is not readily comprehensible for a non-mathematical person. It may also be pointed out that standard deviation gives greater weight to extreme values and as such has not found favour with economists or businessmen who are more interested in the results of the modal class. Taking into consideration the pros and cons and also the wide applications of standard deviation in statistical theory, we may regard standard deviation as the best and the most powerful measure of dispersion.

The square of standard deviation is called the *variance* and is given by

$$\sigma^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 \qquad ...(3·3a)$$

Root mean square deviation, denoted by 's' is given by

$$s = \sqrt{\frac{1}{N} \sum_i f_i (x_i - A)^2} \qquad ...(3·4)$$

where A is any arbitrary number. s^2 is called mean square deviation.

3·7·1. Relation between σ and *s*. By definition, we have

$$s^2 = \frac{1}{N} \sum_i f_i (x_i - A)^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x} + \bar{x} - A)^2$$

$$= \frac{1}{N} \sum_i f_i \left[(x_i - \bar{x})^2 + (\bar{x} - A)^2 + 2 (\bar{x} - A) (x_i - \bar{x}) \right]$$

$$= \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 + (\bar{x} - A)^2 \frac{1}{N} \sum_i f_i + 2 (\bar{x} - A) \sum_i f_i (x_i - \bar{x}),$$

$(\bar{x} - A)$, being constant is taken outside the summation sign. But $\sum_i f_i (x_i - \bar{x}) = 0$, being the algebraic sum of the deviations of the given values from their mean. Thus

$$s^2 = \sigma^2 + (\bar{x} - A)^2 = \sigma^2 + d^2, \text{ where } d = \bar{x} - A$$

Obviously s^2 will be least when $d = 0$, *i.e.*, $\bar{x} = A$. Hence mean square deviation and consequently root mean square deviation is least when the deviations are taken from $A = \bar{x}$, *i.e.*, standard deviation is the least value of root mean square deviation.

The same result could be obtained alternatively as follows:

Mean square deviation is given by

$$s^2 = \frac{1}{N} \sum_i f_i (x_i - A)^2$$

It has been shown in § 2·5·1 Property 2 that $\sum_i f_i (x_i - A)^2$ is minimum when $A = \bar{x}$. Thus mean square deviation is minimum when $A = \bar{x}$ and its minimum value is

$$(s^2)\min = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = \sigma^2$$

Hence variance is the minimum value of mean square deviation or standard deviation is the minimum value of root mean square deviation.

3·7·2. Different Formulae For Calculating Variance. By definition, we have

$$\sigma^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2$$

More precisely we write it as σ_x^2, *i.e.*, variance of x. Thus

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 \qquad ...(3·5)$$

If $\bar{x}$ is not a whole number but comes out to be in fractions, the calculation of σ_x^2 by using (3·5) is very cumbersome and time consuming. In order to overcome this difficulty, we shall develop different forms of the formula (3·5) which reduce the arithmetic to a great extent and are very useful for computational work. In the following sequence the summation is extended over i from 1 to n.

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = \frac{1}{N} \sum_i f_i (x_i^2 + \bar{x}^2 - 2 x_i \bar{x})$$

$$= \frac{1}{N} \sum_i f_i x_i^2 + \bar{x}^2 \frac{1}{N} \sum_i f_i - 2\bar{x} \cdot \frac{1}{N} \sum_i f_i x_i$$

$$= \frac{1}{N} \sum_i f_i x_i^2 + \bar{x}^2 - 2\bar{x}^2 = \frac{1}{N} \sum_i f_i x_i^2 - \bar{x}^2 \qquad ...(3·6)$$

$$\Rightarrow \qquad \sigma_x^2 = \frac{1}{N} \sum_i f_i x_i^2 - \left(\frac{1}{N} \sum_i f_i x_i \right)^2 \qquad ...(3·6a)$$

If the values of x and f are large the calculation of fx, fx^2 is quite tedious. In that case we take the deviations from any arbitrary point 'A'. Generally the point in the middle of the distribution is much convenient though the formula is true in general. We have

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^2 = \frac{1}{N} \sum_i f_i (x_i - A + A - \bar{x})^2$$

$$= \frac{1}{N} \sum_i f_i (d_i + A - \bar{x})^2, \text{ where } d_i = x_i - A.$$

$$\sigma_x^2 = \frac{1}{N} \sum_i f_i [\, d_i^2 + (A - \bar{x})^2 + 2 (A - \bar{x})\, d_i \,]$$

$$= \frac{1}{N} \sum_i f_i d_i^2 + (A - \bar{x})^2 + 2 (A - \bar{x}) \cdot \frac{1}{N} \sum_i f_i d_i$$

We know that if $d_i = x_i - A$ then $\bar{x} = A + \frac{1}{N} \sum_i f_i d_i$

$$\therefore \qquad A - \bar{x} = -\frac{1}{N} \sum_i f_i d_i$$

Hence

$$\sigma_x^2 = \frac{1}{N} \Sigma f_i d_i^2 + \left(-\frac{1}{N} \sum_i f_i d_i \right)^2 + 2 \left(-\frac{1}{N} \sum_i f_i d_i \right) \left(\frac{1}{N} \sum_i f_i d_i \right)$$

$$= \frac{1}{N} \sum_i f_i d_i^2 - \left(\frac{1}{N} \sum_i f_i d_i \right)^2 \qquad ...(3·7)$$

$$\Rightarrow \qquad \sigma_x^2 = \sigma_d^2 \qquad \text{[On comparison with (3·6}a\text{)]}$$

Hence variance and consequently standard deviation is independent of change of origin.

If we take $d_i = (x_i - A)/h$ so that $(x_i - A) = hd_i$, then

$$\sigma_x^2 = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^2 = \frac{1}{N}\sum_i f_i (x_i - A + A - \bar{x})^2$$

$$= \frac{1}{N}\sum_i f_i (hd_i + A - \bar{x})^2$$

$$= h^2 \frac{1}{N}\sum_i f_i d_i^2 + (A - \bar{x})^2 + 2(A - \bar{x}) \cdot h \cdot \frac{1}{N}\sum_i f_i d_i$$

Using $\bar{x} = A + h\dfrac{\sum f_i d_i}{N}$, we get

$$\sigma_x^2 = h^2\left[\frac{1}{N}\sum_i f_i d_i^2 - \left(\frac{1}{N}\sum_i f_i d_i\right)^2\right] = h^2 \sigma_d^2, \qquad ...(3·8)$$

which shows that *variance is not independent of change of scale.*

Aliter. If $d_i = \dfrac{x_i - A}{h}$, then

$$x_i = A + hd_i \quad \text{and} \quad \bar{x} = A + h \cdot \frac{1}{N}\sum_i f_i d_i = A + h\bar{d}$$

Obviously $$x_i - \bar{x} = h(d_i - \bar{d})$$

$$\therefore \quad \sigma_x^2 = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^2 = h^2 \cdot \frac{1}{N}\sum_i f_i (d_i - \bar{d})^2 = h^2 \sigma_d^2$$

Hence variance is independent of change of origin but not of scale.

Example 3·1. *Calculate the mean and standard deviation for the following table giving the age distribution of 542 members.*

Age in years :	20—30	30—40	40—50	50—60	60—70	70—80	80—90
No. of members :	3	61	132	153	140	51	2

Solution. Here we take $d = \dfrac{x - A}{h} = \dfrac{x - 55}{10}$

Age group	*Mid-value* (x)	*Frequency* (f)	$d = \dfrac{x - 55}{10}$	fd	fd^2
20 — 30	25	3	–3	–9	27
30 — 40	35	61	–2	–122	244
40 — 50	45	132	–1	–132	132
50 — 60	55	153	0	0	0
60 — 70	65	140	1	140	140
70 — 80	75	51	2	102	204
80 — 90	85	2	3	6	18
		$N = \Sigma f = 542$		$\Sigma fd = -15$	$\Sigma fd^2 = 765$

$$\bar{x} = A + h\frac{\Sigma fd}{N} = 55 + \frac{10 \times (-15)}{542} = 55 - 0·28 = 54·72 \text{ years.}$$

$$\sigma^2 = h^2\left[\frac{1}{N}\Sigma fd^2 - \left(\frac{1}{N}\Sigma fd\right)^2\right] = 100\left[\frac{765}{542} - (0·28)^2\right]$$

$= 100 \times 1{\cdot}333 = 133{\cdot}3$

$\therefore$ σ (standard deviation) = 11·55 years

Example 3·2. *Prove that for any discrete distribution standard deviation is not less than mean deviation from mean.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

Solution. Let $x_i \mid f_i$, $i = 1, 2, 3, ..., n$ be any discrete distribution. Then we have to prove that

$$\text{S.D.} \not< \text{Mean deviation from mean}$$

$$\Rightarrow \quad (\text{S.D.})^2 \not< (\text{Mean deviation from mean})^2$$

$$\Rightarrow \quad (\text{S.D.})^2 \geq (\text{M. D. from mean})^2$$

$$\Rightarrow \quad \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^2 \geq \left(\frac{1}{N}\sum_{i=1}^{n} f_i |x_i - \bar{x}|\right)^2$$

If we put $|x_i - \bar{x}| = z_i$, then we have to prove that

$$\frac{1}{N}\sum_{i=1}^{n} f_i z_i^2 \geq \left(\frac{1}{N}\sum_{i=1}^{n} f_i z_i\right)^2$$

i.e., $$\frac{1}{N}\sum_{i=1}^{n} f_i z_i^2 - \left(\frac{1}{N}\sum_{i=1}^{n} f_i z_i\right)^2 \geq 0$$

i.e., $$\frac{1}{N}\sum_{i=1}^{n} f_i (z_i - \bar{z})^2 \geq 0$$

i.e., $$\sigma_z^2 \geq 0,$$

which is always true. Hence the result.

Example 3·3. *Find the mean deviation from the mean and standard deviation of A.P. $a, a + d, a + 2d, ..., a + 2nd$ and verify that the latter is greater than the former.* **[Delhi Univ. B.Sc. (Stat. Hons.), 1990]**

Solution. We know that the mean of a series in A.P. is the mean of its first and last term. Hence the mean of the given series is

$$\bar{x} = \frac{1}{2}(a + a + 2nd) = a + nd$$

x	$\lvert x - \bar{x} \rvert$	$(x - \bar{x})^2$
a	nd	n^2d^2
$a + d$	$(n-1)\,d$	$(n-1)^2\,d^2$
$a + 2d$	$(n-2)\,d$	$(n-2)^2\,d^2$
⋮	⋮	⋮
$a + (n-2)\,d$	$2d$	$2^2 . d^2$
$a + (n-1)\,d$	d	$1^2 . d^2$
$a + nd$	0	0
$a + (n+1)\,d$	d	$1^2 . d^2$
$a + (n+2)\,d$	$2d$	$2^2 . d^2$

⋮	⋮	⋮
$a + (2n-2)\,d$	$(n-2)\,d$	$(n-2)^2 d^2$
$a + (2n-2)\,d$	$(n-1)\,d$	$(n-2)^2 d^2$
$a + 2nd$	nd	$n^2 d^2$

$$\text{Mean deviation from mean} = \frac{1}{2n+1}\Sigma\,|x-\bar{x}|$$

$$= \frac{1}{2n+1}\,2.d\,(1+2+3+\ldots+n)$$

$$= \frac{n\,(n+1)d}{(2n+1)}$$

$$\sigma^2 = \frac{1}{2n+1}\Sigma\,(x-\bar{x})^2 = \frac{1}{2n+1}\,2.d^2\,(1^2+2^2+3^2+\ldots+n^2)$$

$$= \frac{1}{2n+1}\,2d^2 \cdot \frac{n\,(n+1)\,(2n+1)}{6} = \frac{n\,(n+1)d^2}{3}$$

Hence standard deviation

$$\sigma = \sqrt{\frac{n\,(n+1)}{3}} \times d$$

Verification.

S.D. > M.D. from mean

if (S.D.)^2 > (M.D. from mean)^2

i.e., if $\dfrac{n\,(n+1)d^2}{3} > \left(\dfrac{n\,(n+1)d}{2n+1}\right)^2$

or if $(2n+1)^2 > 3n\,(n+1)$

or if $n^2 + n + 1 > 0$

or if $(n+\frac{1}{2})^2 + \frac{3}{4} > 0$

which is always ture.

Example 3·4. *Show that in a discrete series if deviations are small compared with mean M so that* $(x/M)^3$ *and higher powers of* (x/M) *are neglected, we have*

(i) $G = M\left(1 - \dfrac{1}{2}\cdot\dfrac{\sigma^2}{M^2}\right)$,

(ii) $M^2 - G^2 = \sigma^2$, *and (iii)* $H = M\left(1 - \dfrac{\sigma^2}{M^2}\right)$,

where M is the arithemetic mean, G, the grometric mean, H, the harmonic mean and σ *is the standard deviation of the distribution.*

Solution. Let $X_i \mid f_i$, $i = 1, 2, \ldots, n$ be the given frequency distribution. Then we are given that $x_i = X_i - M$, *i.e.*, $X_i = x_i + M$ where M is the mean of

the distribution. We have

$$\sum_i f_i x_i = \sum_i f_i (X_i - M) = 0, \qquad ...(1)$$

being the algebraic sum of the deviations of the given values from their mean. Also

$$\sum_i f_i x_i^2 = \sum_i f_i (X_i - M)^2 = \sigma^2 \qquad ...(2)$$

(i) By definition, we have

$$G = (X_1^{f_1} . X_2^{f_2} \dots X_n^{f_n})^{1/N}, \text{ where } N = \Sigma f_i$$

$$\log G = \frac{1}{N} \sum_i f_i \log X_i = \frac{1}{N} \sum_i f_i \log (x_i + M)$$

$$= \frac{1}{N} \sum_i \left\{ f_i \log \left[M \left(1 + \frac{x_i}{M} \right) \right] \right\}$$

$$= \frac{1}{N} \sum_i \left\{ f_i \left[\log M + \log \left(1 + \frac{x_i}{M} \right) \right] \right\}$$

$$= \log M + \frac{1}{N} \sum_i f_i \log \left(1 + \frac{x_i}{M} \right)$$

$$= \log M + \frac{1}{N} \sum_i f_i \left[\frac{x_i}{M} - \frac{1}{2} \frac{x_i^2}{M^2} + \frac{1}{3} \left(\frac{x_i}{M} \right)^3 + \dots \right],$$

the expansion of $\log \left(1 + \frac{x_i}{M} \right)$ in ascending powers of (x_i/M) being valid since $|x_i / M| < 1$. Neglecting $(x_i/M)^3$ and higher powers of (x_i/M), we get

$$\log G = \log M + \frac{1}{NM} \sum_i f_i x_i - \frac{1}{2M^2} \cdot \frac{1}{N} \sum_i f_i x_i^2$$

$$= \log M - \frac{\sigma^2}{2M^2}, \qquad \text{[On using (1) and (2)]}$$

$$= \log \left(M e^{-\sigma^2/2M^2} \right)$$

$$\Rightarrow \qquad G = M e^{-\sigma^2/2M^2} = M \left(1 - \frac{\sigma^2}{2M^2} \right),$$

neglecting higher powers.

Hence $$G = M \left(1 - \frac{1}{2} \cdot \frac{\sigma^2}{M^2} \right). \qquad ...(3)$$

(ii) Squaring both sides in (3), we get

$$G^2 = M^2 \left(1 - \frac{1}{2} \cdot \frac{\sigma^2}{M^2} \right)^2 = M^2 \left(1 - \frac{\sigma^2}{M^2} \right) = M^2 - \sigma^2,$$

neglecting $(\sigma/M)^4$.

$$\therefore \qquad M^2 - G^2 = \sigma^2 \qquad ...(4)$$

(iii) By definition, harmonic mean H is given by

$$\frac{1}{H} = \frac{1}{N}\sum_i (f_i/X_i) = \frac{1}{N}\sum_i [f_i/(x_i + M)]$$

$$= \frac{1}{MN}\sum_i \frac{f_i}{[1 + (x_i/M)]} = \frac{1}{MN}\sum_i f_i\left(1 + \frac{x_i}{M}\right)^{-1}$$

Since $\left|\frac{x_i}{M}\right| < 1$, the expansion of $\left(1 + \frac{x_i}{M}\right)^{-1}$ in ascending powers of (x_i/M) is valid. Neglecting $(x_i/M)^3$ and higher powers of (x_i/M), we get

$$\frac{1}{H} = \frac{1}{MN}\sum_i f_i\left(1 - \frac{x_i}{M} + \frac{x_i^2}{M^2}\right)$$

$$= \frac{1}{M}\left(\frac{1}{N}\sum_i f_i - \frac{1}{MN}\sum_i f_i x_i + \frac{1}{M^2}\frac{1}{N}\sum_i f_i x_i^2\right)$$

$$= \frac{1}{M}\left(1 + \frac{\sigma^2}{M^2}\right) \qquad \text{[On using (1) and (2)]}$$

$$\therefore \qquad H = M\left(1 + \frac{\sigma^2}{M^2}\right)^{-1} = M\left(1 - \frac{\sigma^2}{M^2}\right),$$

higher powers being neglected.

Hence $$H = M\left(1 - \frac{\sigma^2}{M^2}\right) \qquad ...(5)$$

Example 3·5. *For a group of 200 candidates, the mean and standard deviation of scores were found to be 40 and 15 respectively. Later on it was discovered that the scores 43 and 35 were misread as 34 and 53 respectively. Find the corrected mean and standard deviation corresponding to the corrected figures.*

Solution. Let x be the given variable. We are given $n = 200$, $\bar{x} = 40$ and $\sigma = 15$

Now $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \Rightarrow \quad \sum_i x_i = n\bar{x} = 200 \times 40 = 8000$$

Also $$\sigma^2 = \frac{1}{n}\sum_i x_i^2 - \bar{x}^2$$

$$\therefore \qquad \sum_i x_i^2 = n(\sigma^2 + \bar{x}^2) = 200\,(225 + 1600) = 365000$$

$$\text{Corrected } \sum_i x_i = 8000 - 34 - 53 + 43 + 35 = 7991$$

and $$\text{Corrected } \sum_i x_i^2 = 365000 - (34)^2 - (53)^2 + (43)^2 + (35)^2 = 364109$$

Hence, $$\text{Corrected mean} = \frac{7991}{200} = 39{\cdot}955$$

$$\text{Corrected } \sigma^2 = \frac{364109}{200} - (39{\cdot}955)^2 = 1820{\cdot}54 - 1596{\cdot}40 = 224{\cdot}14$$

$\therefore$ Corrected standard deviation = 14·97

3·7·3. Theorem. *(Variance of the combined series). If n_1, n_2 are the sizes; $\bar{x}_1, \bar{x}_2$ the means, and σ_1, σ_2 the standard deviations of two series, then the standard deviation σ of the combined series of size $n_1 + n_2$ is given by*

$$\sigma^2 = \frac{1}{n_1 + n_2}\left[n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2) \right] \qquad ...(3{\cdot}9)$$

where $d_1 = \bar{x}_1 - \bar{x}$, $d_2 = \bar{x}_2 - \bar{x}$

and $\bar{x} = \dfrac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$, *is the mean of the combined series.*

Proof. Let x_{1i}; $i = 1, 2, ..., n_1$ and x_{2j}; $j = 1, 2, ..., n_2$, be the two series then

$$\left.\begin{aligned} \bar{x}_1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i} \\ \bar{x}_2 &= \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j} \end{aligned}\right\} ...(*) \quad \text{and} \quad \left.\begin{aligned} \sigma_1^2 &= \frac{1}{n_1}\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 \\ \sigma_2^2 &= \frac{1}{n_2}\sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2 \end{aligned}\right\} ...(**)$$

The mean $\bar{x}$ of the combined series is given by

$$\bar{x} = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1} x_{1i} + \sum_{j=1}^{n_2} x_{2j} \right] = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} \qquad \text{[From (*)]}$$

The variance σ^2 of the combined series is given by

$$\sigma^2 = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1} (x_{1i} - \bar{x})^2 + \sum_{j=1}^{n_2} (x_{2j} - \bar{x})^2 \right] \qquad ...(3{\cdot}10)$$

Now

$$\sum_{i=1}^{n_1} (x_{1i} - \bar{x})^2 = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1 + \bar{x}_1 - \bar{x})^2$$

$$= \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + n_1 (\bar{x}_1 - \bar{x})^2 + 2 (\bar{x}_1 - \bar{x}) \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1). \qquad (3{\cdot}10a)$$

But $\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1) = 0$, being the algebraic sum of the deviations the values of first series from their mean. Hence from (3·10*a*), on using (**), we get

$$\sum_{i=1}^{n_1} (x_{1i} - \bar{x})^2 = n_1 \sigma_1^2 + n_1 (\bar{x}_1 - \bar{x})^2 = n_1 \sigma_1^2 + n_1 d_1^2 \qquad ...(3{\cdot}10b)$$

where $d_1 = \bar{x}_1 - \bar{x}$.

Similarly, we get

$$\sum_{j=1}^{n_2} (x_{2j} - \bar{x})^2 = \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2 + \bar{x}_2 - \bar{x})^2$$

$$= \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2 + n_2 (\bar{x}_2 - \bar{x})^2 = n_2 \sigma_2^2 + n_2 d_2^2 \qquad ...(3·10c)$$

where $d_2 = \bar{x}_2 - \bar{x}$.

Substituting from (3·10*b*) and (3·10*c*) in (3·10), we get the required formula

$$\sigma^2 = \frac{1}{n_1 + n_2} \left[n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2) \right]$$

This formula can be simplified still further. We have

$$d_1 = \bar{x}_1 - \bar{x} = \bar{x}_1 - \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{n_2 (\bar{x}_1 - \bar{x}_2)}{n_1 + n_2}$$

$$d_2 = \bar{x}_2 - \bar{x} = \bar{x}_2 - \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{n_1 (\bar{x}_2 - \bar{x}_1)}{n_1 + n_2}$$

Hence

$$\sigma^2 = \frac{1}{n_1 + n_2} \left[n_1 \sigma_1^2 + n_2 \sigma_2^2 + \left\{ \frac{n_1 n_2^2 (\bar{x}_1 - \bar{x}_2)^2}{(n_1 + n_2)^2} + \frac{n_2 n_1^2 (\bar{x}_2 - \bar{x}_1)^2}{(n_1 + n_2)^2} \right\} \right]$$

$$= \frac{1}{n_1 + n_2} \left[n_1 \sigma_1^2 + n_2 \sigma_2^2 + \frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)^2 \right] \qquad ...(3·11)$$

Remark. The formula (3·9) can be easily generalised to the case of more than two series. If $n_i, \bar{x}_i$ and σ_i, $i = 1, 2, ..., k$ are the sizes, means and standard deviations respectively of k-component series then the standard deviation σ of the combined series of size $\sum_{i=1}^{k} n_i$ is given by

$$\sigma^2 = \frac{1}{n_1 + n_2 + ... + n_k} \left[n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2) + ... + n_k (\sigma_k^2 + d_k^2) \right]$$

...(3·12)

where $d_i = \bar{x}_i - \bar{x}$; $i = 1, 2, ..., k$

and $$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + ... + n_k \bar{x}_k}{n_1 + n_2 + ... + n_k}$$

Example 3·6. *The first of the two samples has 100 items with mean 15 and standard deviation 3. If the whole group has 250 items with mean 15·6 and standard deviation* $\sqrt{13·44}$*, find the standard deviation of the second group.*

Solution. Here we are given

$$n_1 = 100, \ \bar{x}_1 = 15 \text{ and } \sigma_1 = 3$$

$$n = n_1 + n_2 = 250, \ \bar{x} = 15·6, \text{ and } \sigma = \sqrt{13·44}$$

We want σ_2.

Obviously $n_2 = 250 - 100 = 150$. We have

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \quad \Rightarrow \quad 15{\cdot}6 = \frac{100 \times 15 + 150 \times \bar{x}_2}{250}$$

$$\Rightarrow \quad 150\,\bar{x}_2 = 250 \times 15{\cdot}6 - 1500 = 2400$$

$$\therefore \quad \bar{x}_2 = \frac{2400}{150} = 16$$

Hence $d_1 = \bar{x}_1 - \bar{x} = 15 - 15{\cdot}6 = -0{\cdot}6$

and $d_2 = \bar{x}_2 - \bar{x} = 16 - 15{\cdot}6 = 0{\cdot}4$

The variance σ^2 of the combined group is given by the formula :

$$(n_1 + n_2)\sigma^2 = n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)$$

$$\Rightarrow \quad 250 \times 13{\cdot}44 = 100(9 + 0{\cdot}36) + 150(\sigma_2^2 + 0{\cdot}16)$$

$$\therefore \quad 150\,\sigma_2^2 = 250 \times 13{\cdot}44 - 100 \times 9{\cdot}36 - 150 \times 0{\cdot}16$$

$$= 3360 - 936 - 24 = 2400$$

$$\therefore \quad \sigma_2^2 = \frac{2400}{150} = 16$$

Hence $\sigma_2 = \sqrt{16} = 4$

3·8. Co-efficient of Dispersion. Whenever we want to compare the variability of the two series which differ widely in their averages or which are measured in different units, we do not merely calculate the measures of dispersion but we calculate the co-efficients of dispersion which are pure numbers independent of the units of measurement. The co-efficients of dispersion (C.D.) based on different measures of dispersion are as follows :

1. C.D. based upon range $= \frac{A - B}{A + B}$, where A and B are the greatest and the smallest items in the series.

2. Based upon quartile deviation :

$$\text{C.D.} = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

3. Based upon mean deviation :

$$\text{C.D.} = \frac{\text{Mean deviation}}{\text{Average from which it is calculated}}$$

4. Based upon standard deviation :

$$\text{C.D.} = \frac{\text{S.D.}}{\text{Mean}} = \frac{\sigma}{\bar{x}}$$

3·8·1. Co-efficient of Variation. 100 times the co-efficient of desperison based upon standard deviation is called co-efficient of variation (C.V.),

$$\text{C.V.} = 100 \times \frac{\sigma}{\bar{x}} \qquad \ldots(3{\cdot}13)$$

According to Professor Karl Pearson who suggested this measure, *C.V. is the percentage variation in the mean, standard deviation being considered as the total variation in the mean.*

For comparing the variability of two series, we calculate the co-efficient of variations for each series. The series having greater C.V. is said to be more variable than the other and the series having lesser C.V. is said to be

more consistent (or homogenous) than the other.

Example 3·7. *An analysis of monthly wages paid to the workers of two firms A and B belonging to the same industry gives the following results :*

	Firm A	*Firm B*
Number of workers	*500*	*600*
Average monthly wage	*Rs. 186·00*	*Rs. 175·00*
Variance of distribution of wages	*81*	*100*

(i) Which firm, A or B, has a larger wage bill ?

(ii) In which firm, A or B, is there greater variability in individual wages?

(iii) Calculate (a) the average monthly wage, and (b) the variance of the distribution of wages, of all the workers in the firms A and B taken together.

Solution.

(i) **Firm A :**

No. of wage-earners (say) $n_1 = 500$

Average monthly wages (say) $\bar{x}_1 = \text{Rs.}186$

$$\text{Average monthly wage} = \frac{\text{Total wages paid}}{\text{No. of workers}}$$

Hence total wages paid to the workers $= n_1\bar{x}_1 = 500 \times 186 = \text{Rs. } 93{,}000$

Firm B

No. of wage-earners (say) $n_2 = 600$

Average monthly wages (say) $\bar{x}_2 = \text{Rs.}175$

$\therefore$ Total wages paid to the workers $= n_2\bar{x}_2 = 600 \times 175 = \text{Rs. } 1{,}05{,}000$

Thus we see that the firm *B* has larger wage bill.

(ii) Variance of distribution of wages in firm *A* (say) $\sigma_1^2 = 81$

Variance of distribution of wages in firm *B* (say) $\sigma_2^2 = 100$

$$\text{C.V. of distribution of wages for firm } A = 100 \times \frac{\sigma_1}{\bar{x}_1} = \frac{100 \times 9}{186} = 4{\cdot}84$$

$$\text{C.V. of distribution of wages for firm } B = 100 \times \frac{\sigma_2}{\bar{x}_2} = \frac{100 \times 10}{175} = 5{\cdot}71$$

Since C.V. for firm *B* is greater than C.V. for firm *A*, firm *B* has greater variability in individual wages.

(iii) (a) The average monthly wages (say) $\bar{x}$, of all the workers in the two firms A and *B* taken together is given by

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{500 \times 186 + 600 \times 175}{500 + 600} = \frac{198000}{1100} = \text{R } 180$$

(b) The combined variance σ^2 is given by the formula:

$$\sigma^2 = \frac{1}{n_1 + n_2}\left[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)\right]$$

where $d_1 = \bar{x}_1 - \bar{x}$ and $d_2 = \bar{x}_2 - \bar{x}$

Here $d_1 = 186 - 180 = 6$ and $d_2 = 175 - 180 = -5$

Hence $\sigma^2 = \frac{500(81+36)+600(100+25)}{500+600} = \frac{133500}{1100} = 121{\cdot}36$

EXERCISE 3 (a)

1. (a) Explain with suitable examples the term 'dispersion.' State the relative and absolute measures of dispersion and describe the merits and demerits of standard deviation.

(b) Explain the main difference between mean deviation and standard deviation. Show that standard deviation is independent of change of origin and scale.

(c) Distinguish between absolute and retative measures of dispersion.

2. *(a)* Explain the graphical method of obtaining median and quartile deviation. **(Calicut Univ.B.Sc, April 1989)**

(b) Compute quartile deviation graphically for the following data :

Marks :	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 & over
Number of students :	5	20	14	10	8	5

3. *(a)* Show that for raw data mean deviation is minimum when measured from the median.

(b) Compute a suitable measure of dispersion for the following grouped frequency distribution giving reasons :

Classes	*Frequency*
Less than 20	30
20 – 30	20
30 – 40	15
40 – 50	10
50 – 60	5

(c) Age distribution of hundred life insurance policyholders is as follows:

Age as on nearest birthday	*Number*
17 – 19·5	9
20 – 25·5	16
26 – 35·5	12
36 – 40·5	26
41 – 50·5	14
51 – 55·5	12
56 – 60·5	6
61 – 70·5	5

Calculate mean deviation from median age.

Ans. Median = 38·25, M.D.=10·605

4. Prove that the mean deviation about the mean $\bar{x}$ of the variate x, the frequency of whose ith size x_i is f_i is given by

$$\frac{2}{N}\left[\bar{x}\sum_{x_i<\bar{x}} f_i - \sum_{x_i<\bar{x}} f_i x_i\right].$$

Hint. Mean deviation about mean

$$= \frac{1}{N}\left[\sum_{x_i<\bar{x}} f_i(\bar{x}-x_i) + \sum_{x_i>\bar{x}} f_i(x_i-\bar{x})\right]$$

$$= \frac{1}{N}\left[-\sum_{x_i<\bar{x}} f_i(x_i-\bar{x}) + \sum_{x_i>\bar{x}} f_i(x_i-\bar{x})\right]$$

Since $\Sigma f_i(x_i-\bar{x}) = 0$,

$$\sum_{x_i>\bar{x}} f_i(x_i-\bar{x}) + \sum_{x_i<\bar{x}} f_i(x_i-\bar{x}) = 0$$

$$\therefore M.D. = \frac{1}{N}\left(-\sum_{x_i<\bar{x}} f_i(x_i-\bar{x}) - \sum_{x_i<\bar{x}} f_i(x_i-\bar{x})\right) = -\frac{2}{N}\left(\sum_{x_i<\bar{x}} f_i(x_i-\bar{x})\right)$$

5. What is standard deviation ? Explain its superiority over other measures of dispersion.

6. Calculate the mean and standard deviation of the following distribution:

x :	2.5 — 7.5	7.5 — 12.5	12.5 — 17.5	17.5 — 22.5
f :	12	28	65	121

x :	22.5 — 27.5	27.5 — 32.5	32.5 — 37.5	37.5 — 42.5	42 5 — 47.5
f :	175	198	176	120	66

x :	47.5 — 52.5	52.5 — 57.5	57.5 — 62.5
f :	27	9	3

Ans. Mean = 30.005, Standard Deviation = 0.01

7. Explain clearly the ideas implied in using arbitrary working orgin, and scale for the calculation of the arithmetic mean and standard deviation of a frequency distribution. The values of the arithmetic mean and standard deviation of the following frequency distribution of a continuous variable derived from the analysis in the above manner are40.604 lb. and 7.92 lb. respectively.

x :	-3	-2	-1	0	1	2	3	4	Total
f :	3	15	45	57	50	36	25	9	240

Determine the actual class intervals.

8. *(a).* The arithmetic mean and variance of a set of 10 figures are known to be 17 and 33 respectively. Of the 10 figures, one figure (*i.e.*, 26) was subsequently found inacurate, and was weeded out. What is the resulting *(a)* arithmetic mean and *(b)* standard deviation. **(M.S. Baroda U. B.Sc. 1993)**

(b) The mean and standard deviation of 20 items is found to be 10 and 2 respectively. At the time of checking it was found that one item 8 was incorrect. Calculate the mean and standard deviation if

(i) the wrong item is omitted, and

(ii) it is replaced by 12.

(c) For a frequency distribution of marks in Statistics of 200 candidates (grouped in intervals 0–5, 5–10,..., etc.), the mean and standard deviation were found to be 40 and 15 respectively. Later it was discovered that the score 43 was misread as 53 in obtaining the frequency distribution. Find the corrected mean and standard deviation corresponding to the corrected frequency distribution.

Ans. Mean = 39·95, S.D. = 14·974.

9. *(a)* Complete a table showing the frequencies with which words of different numbers of letters occur in the extract reproduced below (omitting punctuation marks) treating as the variable the number of letters in each word, and obtain the mean, median and co-efficient of variation of the distribution :

"Her eyes were blue : blue as autumn distance blue as the blue we see, between the retreating mouldings of hills and woody slopes on a sunny September morning : a misty and shady blue, that had no beginning or surface, and was looked into rather than at. "

Ans. Mean = 4·35, Median = 4, σ = 2·23 and C.V. = 51·26

(b) Treating the number of letters in each word in the following passage as the variable x, prepare the frequency distribution table and obtain its mean, median, mode and variance.

"The reliability of data must always be examined before any attempt is made to base conclusions upon them. This is true of all data, but particularly so of numerical data, which do not carry their quality written large on them. It is a waste of time to apply the refined theoretical methods of Statistics to data which are suspect from the beginning. "

Ans. Mean = 4·565, Median = 4, Mode = 3, S.D. = 2·673.

10. The mean of 5 observations is 4·4 and variance is 8·24. It three of the five observation are 1, 2 and 6, find the other two.

11. *(a)* Scores of two golfers for 24 rounds were as follows :

Golfer A : 74, 75, 78, 72, 77, 79, 78, 81, 76, 72, 72, 77, 74, 70, 78, 79, 80, 81, 74, 80, 75, 71, 73.

Golfer B : 86, 84, 80, 88, 89, 85, 86, 82, 82, 79, 86, 80, 82, 76, 86, 89, 87, 83, 80, 88, 86, 81, 81, 87

Find which golfer may be considered to be a more consistent player ?

Ans. Golfer B is more consistent player.

(b) The sum and sum of squares corresponding to length X (in cms.) and weight Y (in gms.) of 50 tapioca tubers are given below :

$\Sigma X = 212$, $\Sigma X^2 = 902{\cdot}8$

$\Sigma Y = 261$, $\Sigma Y^2 = 1457{\cdot}6$

Which is more varying, the length or weight.

12. *(a)* Lives of two models of refrigerators turned in for new models in a recent survey are

Life (No. of years)	*Model A*	*Model B*
0 – 2	5	2
2 – 4	16	7

4 – 6	13	12
6 – 8	7	19
8 – 10	5	9
10 – 12	4	1

What is the average life of each model of these refrigerators ? Which model shows more uniformity ?

Ans. C.V. (Model A)=54·9%, C.V. (Model B)=3·62%

(b) Goals scored by two teams A and B in a football season were as follows :

No. of goals scored in a match	*No. of matches* A	B
0	27	17
1	9	9
2	8	6
3	5	5
4	4	3

(Sri Venketeswara U. B.Sc. Sept. 1992)

Find out which team is more consistent.

Ans. Team *A* : C.V. = 122·0, Team *B* : C.V. = 108·3 .

(c) An analysis of monthly wages paid to the workers in two firms, A and B belonging to the same industry, gave the following results :

	Firm A	*Firm B*
No. of wage-earners	986	548
Average monthly wages	Rs. 52·5	Rs. 47·5
Variance of distribution of wages	100	121

(i) Which firm, *A* or *B*, pays out larger amount as monthly wages ?

(ii) In which firm *A* or *B*, is there greater variability in individual wages?

(iii) What are the measures of average monthly wages and the variability in individual wages, of all the workers in the two firms, *A* and *B* taken together.

Ans. *(i)* Firm *B* pays a larger amount as monthly wages.
(ii) There is greater variability in individual wages in firm *B*.
(iii) Combined arithmetic mean = Rs.49·87.
Combined standard deviation = Rs.10·82.

14. *(a)* The following data give the arithmetic averages and standard deviations of three sub-groups. Calculate the arithmetic average and standard deviation of the whole group.

Sub-group	*No. of men*	*Average wages (Rs.)*	*Standard deviation (Rs.)*
A	50	61·0	8·0
B	100	70·0	9·0
C	120	80·5	10·0

Ans. Combined Mean = 73, Combined S.D.=11·9 .

(b) Find the missing information from the following data :

	Group I	*Group II*	*Group III*	*Combined*
Number	50	?	90	200
Standard Deviation	6	7	?	7·746
Mean	113	?	115	116

Ans. $n_2 = 60$, $\bar{x}_2 = 120$ and $\sigma_3 = 8$

15. A collar manufacturer is considering the production of a new style collar to attract young men. The following statistics of neck circumference are avail based on the measurement of a typical group of students :

Mid-value in inches	12·5	13·0	13·5	14·0	14·5	15·0	15·5	16·0
No. of students	4	19	30	63	66	29	18	1

Compute the mean and standard deviation and use the criterion $\bar{x} \pm$ obtain the largest and smallest size of collar he should make in order to meet needs of practically all his customers bearing in mind that the collars are worn on average 3/4 inch larger than neck size. **(Nagpur Univ. B.Sc., 1992)**

Ans. Mean = 14·232, S.D.=0·72, largest size = 17·14", smallest size = 12·83"

16. *(a)* A frequency distribution is divided into two parts. The mean and standard deviation of the first part are m_1 and s_1 and those of the second part are m_2 and s_2 respectively. Obtain the mean and standard deviation for the combined distribution. **[Delhi Univ. B.Sc.(Stat.Hons.), 1986]**

(b) The means of two samples of size 50 and 100 respectively are 54·1 and 50·3 and the standard deviations are 8 and 7. Obtain the mean and standard deviation of the sample of size 150 obtained by combining the two samples.

Ans. Combined mean = 51·57. Combined S.D. = 7·5 approx.

(c) A distribution consists of three components with frequencies 200, 250 and 300 having means 25, 10 and 15 and standard deviations 3, 4 and 5 respectively.

Show that the mean of the combined group is 16 and its standard deviation 7.2 approximately. **(Bangalore Univ. B.Sc. 1992)**

17. In a certain test for which the pass marks is 30, the distribution of marks of passing candidates classified by sex (boys and girls) were as given below :

Marks	*Frequency*	
	Boys	*Girls*
30-34	5	15
35-39	10	20
40-44	15	30
45-49	30	20
50-54	5	5
55-59	5	-
Total	70	90

The overall means and standard deviation of marks for boys including the 30 failed were 38 and 10. The corresponding figures for girls including the 10 failed were 35 and 9.

(i) Find the mean and standard deviation of marks obtained by the 30 boys who failed in the test.

(ii) The moderation committee argued that percentage of passes among girls is higher because the girls are very studious and if the intention is to pass those who are really intelligent, a higher pass marks should be used for girls. Wiothout quetioning the propriety of this argument, suggest what the pass mark should be which would allow only 70% of the girls to pass.

(iii) The prize committee decided to award prizes to the best 40 candidates (irrespective of sex) judged on the basis of marks obtained in the test. Estimate the number of girls who would receive prizes.

Ans. (i) $\bar{x} = 22{\cdot}83$, $\sigma_2 = 8{\cdot}27$ (ii) 39 (iii) 15

18. Find the mean and variance of first n-natural numbers.

(Agra Univ. B.Sc., 1993)

Ans. $\bar{x} = \dfrac{n+1}{2}$, $\sigma_2 = \dfrac{n^2-1}{12}$

19. In a frequency distribution, the n intervals are 0 to 1, 1 to 2, ..., $(n-1)$ to n with equal frequencies. Find the mean deviation and variance.

20. If the mean and standard deviation of a variable x are m and σ respectively, obtain the mean and standard deviation of $(ax + b)/c$, where a, b and c are constants.

Ans. $\bar{u} = \dfrac{1}{c}(a\bar{x} + b)$, $\sigma_u = \left|\dfrac{a}{c}\right| \sigma$

21. In a series of measurements we obtain m_1 values of magnitude x_1, m_2 values of magnitude x_2, and so on. If $\bar{x}$ is the mean value of all the measurements, prove that the standard deviation is

$$\sqrt{\frac{\Sigma m_r (k - x_r)^2}{\Sigma m_r} - \delta^2}$$

where $\bar{x} = k + \delta$ and k is any constant. **Delhi Univ. B.Sc. (Stat. Hons.), 1992**

22. (a) Show that in a discrete series if deviations are small compared with mean M so that $(x/M)^2$ and higher powers of (x/M) are neglected, prove that

(i) $MH = G^2$ *(II)* $M - 2G + H = 0$,

where G is geometric mean and H is harmonic mean.

(b) The mean and standard deviation of a variable x are m and σ respectively. If the deviations are small compared with the value of the mean, show that

(i) Mean $(\sqrt{x}) = \sqrt{m}\left(1 - \dfrac{\sigma^2}{8m^2}\right)$

(ii) Mean $\left(\dfrac{1}{\sqrt{x}}\right) = \dfrac{1}{\sqrt{m}}\left(1 + \dfrac{3\sigma^2}{8m^2}\right)$ approximately.

(M.S. Baroda U. B.Sc. 1993)

(*c*) If the deviation $X_i = x_i - M$ is very small in comparison with mean M and $(X_i/M)^2$ and higher powers of (X_i/M) are neglected, prove that

$$V = \sqrt{\frac{2(M - G)}{M}}$$

where G is the geometric mean of the values $x_1, x_2, \ldots, x_n$ and V is the coefficient of dispersion (σ/M). **(Lucknow Univ. B.Sc., 1993)**

23. From a sample of observations the arithmetic mean and variance are calculated. It is then found that one of the values, x_1, is in error and should be replaced by x_1'. Show that the adjustment to the variance to correct this error is

$$\frac{1}{n}(x'_1 - x_1)\left(x'_1 + x_1 - \frac{x_1' - x_1 + 2T}{n}\right)$$

where T is the total of the original results.

(Meerut Univ. B.Sc., 1992; Delhi Univ. B.Sc. (Stat. Hons.), 1989, 1985]

Hint. $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$

$$= \frac{1}{n}\left(x_1^2 + x_2^2 + \ldots + x_n^2\right) - \frac{T^2}{n^2}$$

where $T = x_1 + x_2 + \ldots + x_n$.

Let σ_1^2 be the corrected variance. Then

$$\sigma_1^2 = \frac{1}{n}\left\{x_1'^2 + x_2^2 + \ldots + x_n^2\right\} - \left\{\frac{T - x_1 + x_1'}{n}\right\}^2$$

Adjustment to the variance to correct the error is :

$$\sigma_1^2 - \sigma^2 = \frac{1}{n}\left\{x_1'^2 - x_1^2\right\} - \frac{1}{n^2}\left\{\left(T - x_1 + x_1''\right)^2 - T^2\right\}$$

$$= \frac{1}{n}\left\{x_1' + x_1\right\}\left\{x_1' - x_1\right\} - \frac{1}{n^2}\left\{\left(x_1' - x_1\right) \times (2T - x_1 + x_1')\right\}$$

24. Show that, if the variable takes the values $0, 1, 2, \ldots, n$ with frequencies proportional to the binomial coefficients ${}^nC_0, {}^nC_1, {}^nC_2, \ldots, {}^nC_n$ respectively then the mean of the distribution is $(n/2)$, the mean square deviation about $x = 0$ is $n(n + 1)/4$ and the variance is $n/4$

[Delhi Univ. B.Sc. (Stat. Hons.), 1991]

Hint. $N = \Sigma f = {}^nC_0 + {}^nC_1 + {}^nC_2 + \ldots + {}^nC_n = (1 + 1)^n = 2^n$

$$\Sigma fx = 0.{}^nC_0 + 1.{}^nC_1 + 2.{}^nC_2 + 3.{}^nC_3 + \ldots + n.{}^nC_n$$

$$= n\left\{1 + (n - 1) + \frac{(n - 1)(n - 2)}{2!} + \ldots + 1\right\}$$

$$= n(1 + 1)^{n-1} = n \cdot 2^{(n-1)}$$

Hence mean $(\bar{x}) = \frac{1}{N}\Sigma fx = \frac{n \cdot 2^{(n-1)}}{2^n} = \frac{n}{2}$

The mean square deviation s^2, (say), about the point $x = 0$ is given by

$$s^2 = \frac{1}{N}\sum fx^2 = \frac{1}{2^n}\left[1^2.{}^nC_1 + 2^2.{}^nC_2 + 3^2.{}^nC_2 + \ldots + n^2.{}^nC_n\right]$$

$$= \frac{n}{2^n}\left[1 + 2(n-1) + \frac{3}{2}(n-1)(n-2) + \ldots + n\right]$$

$$= \frac{n}{2^n}\left(\left\{1 + (n-1) + \frac{(n-1)(n-2)}{2!} + \ldots + 1\right\}\right.$$

$$\left. + \{(n-1) + (n-1)(n-2) + \ldots + (n-1)\}\right)$$

$$= \frac{n}{2^n}\left[\left({}^{n-1}C_0 + {}^{n-1}C_1 + {}^{n-1}C_2 + \ldots + {}^{n-1}C_{n-1}\right)\right.$$

$$\left. + \{(n-1)({}^{n-2}C_0 + {}^{n-2}C_1 + \ldots + {}^{n-2}C_{n-2})\}\right]$$

$$= \frac{n}{2^n}\left[(1+1)^{n-1} + (n-1)(1+1)^{n-2}\right] = \frac{n(n+1)}{4}$$

$$\therefore \qquad \sigma^2 = \frac{n(n+1)}{4} - \frac{n^2}{4} = \frac{n}{4}.$$

25. (*a*) Let r be the range and s be the standard deviation of a set of observations $x_1, x_2, \ldots, x_n$; then prove by general reasoning or otherwise that $s \le r$.

Hint. Since $x_i - \bar{x} \le r$, $i = 1, 2, \ldots, n$, we have

$$s^2 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^2 \le \frac{1}{N}\sum_{i=1}^{n} f_i (r^2)$$

$$\Rightarrow \qquad s^2 \le r^2 \frac{1}{N}\sum_{i=1}^{n} f_i = r^2 \quad \Rightarrow \quad s \le r$$

(*b*) Let r be the range and

$$S = \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{1}{2}}$$

be the standard deviation of a set of observations $x_1, x_2, \ldots, x_n$, then prove that

$$S \le r\left(\frac{n}{n-1}\right)^{\frac{1}{2}}$$ **[Punjab Univ. B.Sc (Stat. Hons.), 1993]**

3·9. Moments. The rth moment of a variable x about any point $x = A$, usually denoted by μ_r' is given by

$$\mu_r' = \frac{1}{N}\sum_i f_i (x_i - A)^r, \quad \sum_i f_i = N \qquad \ldots(3\cdot14)$$

$$= \frac{1}{N}\sum_i f_i d_i^r, \qquad \ldots(3\cdot14a)$$

where $d_i = x_i - A$.

The rth moment of a variable about the mean $\bar{x}$, usually denoted by μ_r is given by

$$\mu_r = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^r = \frac{1}{N}\sum_i f_i z_i' \qquad \ldots(3\cdot15)$$

where $z_i = x_i - \bar{x}$.

In particular

$$\mu_0 = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^0 = \frac{1}{N}\sum_i f_i = 1$$

and $\mu_1 = \frac{1}{N}\sum_i f_i (x_i - \bar{x}) = 0$, being the algebraic sum of deviations from the mean. Also

$$\mu_2 = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^2 = \sigma^2 \qquad ...(3·16)$$

These results, viz., $\mu_0 = 1$, $\mu_1 = 0$, and $\mu_2 = \sigma^2$, are of fundamental importance and should be committed to memory.

We know that if $d_i = x_i - A$, then

$$\bar{x} = A + \frac{1}{N}\sum_i f_i d_i = A + \mu_1' \qquad ...(3·17)$$

3·9·1. Relation between moments about mean in terms of moments about any point and *vice versa*.

We have

$$\mu_r = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^r = \frac{1}{N}\sum_i f_i (x_i - A + A - \bar{x})^r$$

$$= \frac{1}{N}\sum_i f_i (d_i + A - \bar{x})^r, \text{ where } d_i = x_i - A$$

Using (3·17), we get

$$\mu_r = \frac{1}{N}\sum_i f_i (d_i - \mu_1')^r$$

$$= \frac{1}{N}\sum f_i [d_i^r - {}^rC_1 d_i^{r-1}\mu_1' + {}^rC_2 d_i^{r-2}\mu_1'^2 - {}^rC_3 d_i^{r-3}\mu_1'^3 + \ldots + (-1)^r \mu_1'^r]$$

$$= \mu_r' - {}^rC_1\mu_{r-1}'\mu_1' + {}^rC_2\mu_{r-2}'\mu_1'^2 - \ldots + (-1)^r\mu_1'^r \qquad ...(3·18)$$

[On using (3·14a)]

In particular, on putting r = 2, 3 and 4 in (3·18), we get

$$\mu_2 = \mu_2' - \mu_1'^2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 \qquad ...(3.19)$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

Conversely,

$$\mu_r' = \frac{1}{N}\sum_i f_i (x_i - A)^r = \frac{1}{N}\sum_i f_i (x_i - \bar{x} + \bar{x} - A)^r$$

$$= \frac{1}{N}\sum_i f_i (z_i + \mu_1')^r$$

where $x_i - \bar{x} = z_i$ and $\bar{x} = A + \mu_1'$

Thus $\mu_r' = \frac{1}{N} \sum_i f_i \left(z_i^r + {}^rC_1 z_i^{r-1} \mu_1' + {}^rC_2 z_i^{r-2} \mu_1'^2 + \ldots + \mu_1'^r\right)$

$= \mu_r + {}^rC_1 \mu_{r-1} \mu_1' + {}^rC_2 \mu_{r-2} \mu_1'^2 + \ldots + \mu_1'^r$. [From (3·15)]

In particular, putting $r = 2$, 3 and 4 and noting that $\mu_1 = 0$, we get

$$\mu_2' = \mu_2 + \mu_1'^2$$

$$\mu_3' = \mu_3 + 3\mu_2 \mu_1' + \mu_1'^3 \quad \ldots (3·20)$$

$$\mu_4' = \mu_4 + 4\mu_3 \mu_1' + 6\mu_2 \mu_1'^2 + \mu_1'^4$$

These formulae enable us to find the moments about any point, once the mean and the moments about mean are known.

3.9.2 Effect of Change of Origin and Scale on Moments.

Let $u = \frac{x - A}{h}$, so that $x = A + hu$, $\bar{x} = A + h\bar{u}$ and $x - \bar{x} = h(u - \bar{u})$

Thus, rth moment of x about any point $x = A$ is given by

$$\mu_r' = \frac{1}{N} \sum_i f_i (x_i - A)^r = \frac{1}{N} \sum_i f_i (hu_i)^r = h^r \cdot \frac{1}{N} \cdot \sum_i f_i u_i^r$$

And the rth moment of x about mean is

$$\mu_r = \frac{1}{N} \sum_i f_i (x_i - \bar{x})^r = \frac{1}{N} \sum_i f_i [h(u_i - \bar{u})]^r$$

$$= h^r \frac{1}{N} \sum_i f_i (u_i - \bar{u})^r$$

Thus the rth moment of the variable x about mean is h^r times the rth moment of the variable u about its mean.

3·9·3. Sheppard's Corrections for Moments. In case of grouped frequency distribution, while calculating moments we assume that the frequencies are concentrated at the middle point of the class intervals. If the distribution is symmetrical or slightly symmetrical and the class intervals are not greater than one-twentieth of the range, this assumption is very nearly true. But since the assumption is not in general true, some error, called the 'grouping error', creeps into the calculation of the moments. W.F. Sheppard proved that if

(*i*) the frequency distribution is continuous, and

(*ii*) the frequency tapers off to zero in both directions,

the effect due to grouping at the mid-point of the intervals can be corrected by the following formulae, known as Sheppard's corrections :

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12} \quad \ldots (3·21)$$

$$\mu_3 \text{ (corrected)} = \mu_3$$

$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{1}{2} h^2 \mu_2 + \frac{7}{240} h^4$$

where h is the width of the class interval.

3·9·4. Charlier's Checks. The following identities

$$\Sigma f(x+1) = \Sigma fx + N;\ \Sigma f(x+1)^2 = \Sigma fx^2 + 2\Sigma fx + N$$

$$\Sigma f(x+1)^3 = \Sigma fx^3 + 3\Sigma fx^2 + 3\Sigma fx + N$$

$$\Sigma f(x+1)^4 = \Sigma fx^4 + 4\Sigma fx^3 + 6\Sigma fx^2 + 4\Sigma fx + N,$$

are often used in checking the accuracy in the calculation of first four moments and are known as Charlier's Checks.

3·10. Pearson's β and γ Coefficients. Karl Pearson defined the following four coefficients, based upon the first four moments about mean :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3},\ \gamma_1 = +\sqrt{\beta_1} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2},\ \gamma_2 = \beta_2 - 3 \qquad \ldots (3\cdot22)$$

It may be pointed out that these coefficients are pure numbers independent of units of measurement. The practical utility of these coefficients is discussed in § 3·13 and § 3·14.

Remark. Sometimes, another coefficient based on moments, *viz*, Alpha (α) coefficient is used. Alpha coefficients are defined as :

$$\alpha_1 = \frac{\mu_1}{\sigma} = 0,\ \alpha_2 = \frac{\mu_2}{\sigma^2} = 1,\ \alpha_3 = \frac{\mu^3}{\sigma^3} = \sqrt{\beta_1} = \gamma_1;\ \alpha_4 = \frac{\mu_4}{\sigma^4} = \beta_2$$

3·11. Factorial Moments. Factorial moment of order r about the origin of the frequency distribution $x_i\,|\,f_i$, $(i = 1, 2, \ldots n)$, is defined as

$$\mu_{(r)}' = \frac{1}{N}\sum_{i=1}^{n} f_i\, x_i^{(r)} \qquad \ldots (3\cdot23)$$

where $x^{(r)} = x\,(x-1)\,(x-2)\ldots(x-r+1)$ and $N = \sum_{i=1}^{n} f_i$

Thus the factorial moment of order r about any point $x = a$ is given by

$$\mu_{(r)}' = \frac{1}{N}\sum_i f_i\ (x_i - a)^{(r)} \qquad \ldots (3\cdot24)$$

where $(x-a)^{(r)} = (x-a)\,(x-a-1)\ldots(x-a-r+1)$

In particular from (3·23), we have

$$\mu_{(1)}' = \frac{1}{N}\sum_i f_i x_i = \mu_1' \text{ (about origin)} = \text{Mean } (\bar{x})$$

$$\mu_{(2)}' = \frac{1}{N}\sum_i f_i x_i^{(2)} = \frac{1}{N}\sum_i f_i x_i\,(x_i - 1)$$

$$= \frac{1}{N}\sum_i f_i x_i^2 - \frac{1}{N}\sum_i f_i x_i = \mu_2' - \mu_1'$$

$$\mu_{(3)}' = \frac{1}{N}\sum_i f_i x_i^{(3)} = \frac{1}{N}\sum_i f_i x_i (x_i - 1)(x_i - 2)$$

$$= \frac{1}{N}\sum_i f_i x_i^3 - 3\,\frac{1}{N}\sum_i f_i x_i^2 + 2\,\frac{1}{N}\cdot\sum_i f_i x_i$$

$$= \mu_3' - 3\mu_2' + 2\mu_1'$$

$$\mu_{(4)}' = \frac{1}{N}\sum_i f_i x_i^{(4)} = \frac{1}{N}\sum_i f_i x_i (x_i - 1)(x_i - 2)(x_i - 3)$$

$$= \frac{1}{N}\sum_i f_i x_i (x_i^3 - 6x_i^2 + 11x_i - 6)$$

$$= \frac{1}{N}\sum_i f_i x_i^4 - 6.\ \frac{1}{N}\sum f_i x_i^3 + 11.\ \frac{1}{N}\sum f_i x_i^2 - 6.\ \frac{1}{N}\sum f_i x_i$$

$$= \mu_4' - 6\mu_3' + 11\mu_2' - 6\mu_1'$$

Conversely, we will get

$$\mu_1' = \mu_{(1)}'$$

$$\mu_2' = \mu_{(2)}' + \mu_{(1)}'$$

$$\mu_3' = \mu_{(3)}' + 3\mu_{(2)}' + \mu_{(1)}' \qquad \ldots (3\cdot25)$$

$$\mu_4' = \mu_{(4)}' + 6\mu_{(3)}' + 7\mu_{(2)}' + \mu_{(1)}'$$

3·12. Absolute Moments. For the frequency distribution x_i / f_i $i = 1, 2, \ldots n$, the rth absolute moment of the variable about the origin is given by

$$\frac{1}{N}\sum_{i=1}^{n} f_i \left| x_i^r \right|, N = \sum f_i \qquad \ldots (3\cdot26)$$

where $\left| x_i^r \right|$ represents the absolute or modulus value of x_i^r.

The rth absolute moment of the variable about the mean $\bar{x}$ is given by

$$\frac{1}{N}\sum_{i=1}^{n} f_i \left| x_i - \bar{x} \right|^r \qquad \ldots (3\cdot26a)$$

Example 3·8. *The first four moments of a distribution about the value 4 of the variable are −1·5, 17, −30 and 108. Find the moments about mean, β_1 and β_2.*

Find also the moments about (i) the origin, and (ii) the point $x = 2$.

Solution. In the usual notations, we are given $A = 4$ and

$$\mu_1' = -1\cdot5,\ \mu_2' = 17,\ \mu_3' = -30 \text{ and } \mu_4' = 108.$$

Moments about mean : $\mu_1 = 0$

$$\mu_2 = \mu_2' - \mu_1'^2 = 17 - (-1\cdot5)^2 = 17 - 2\cdot25 = 14\cdot75$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= -30 - 3 \times (17) \times (-1\cdot5) + 2\,(-1\cdot5)^3$$

$$= -30 + 76\cdot5 - 6\cdot75 = 39\cdot75$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$
$$= 108 - 4(-30)(-1{\cdot}5) + 6(17)(-1{\cdot}5)^2 - 3(-1{\cdot}5)^4$$
$$= 108 - 180 + 229{\cdot}5 - 15{\cdot}1875 = 142{\cdot}3125$$

Hence $$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(39{\cdot}75)^2}{(14{\cdot}75)^3} = 0{\cdot}4924$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{(142{\cdot}3125)}{(14{\cdot}75)^2} = 0{\cdot}6541$$

Also $$\bar{x} = A + \mu_1' = 4 + (-1{\cdot}5) = 2{\cdot}5$$

Moments about origin. We have

$\bar{x} = 2{\cdot}5$, $\mu_2 = 14{\cdot}75$, $\mu_3 = 39{\cdot}75$ and $\mu_1 = 142{\cdot}31$ (approx).

We know $\bar{x} = A + \mu_1'$, where μ_1' is the first moment about the point $x = A$. Taking $A = 0$, we get the first moment about origin as μ_1' = mean $= 2{\cdot}5$.

Using (3·20), we get

$$\mu_2' = \mu_2 + \mu_1'^2 = 14{\cdot}75 + (2{\cdot}5)^2 = 14{\cdot}75 + 6{\cdot}25 = 21$$
$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + \mu_1'^3 = 39{\cdot}75 + 3(14{\cdot}75)(2{\cdot}5) + (2{\cdot}5)^3$$
$$= 39{\cdot}75 + 110{\cdot}625 + 15{\cdot}625 = 166$$
$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu_1'^2 + \mu_1'^4$$
$$= 142{\cdot}3125 + 4(39{\cdot}75)(2{\cdot}5) + 6(14{\cdot}75)(2{\cdot}5)^2 + (2{\cdot}5)^4$$
$$= 142{\cdot}3125 + 397{\cdot}5 + 553{\cdot}125 + 39{\cdot}0625$$
$$= 1132.$$

Moments about the point $x = 2$. We have $\bar{x} = A + \mu_1'$. Taking $A = 2$, the first moment about the point $x = 2$ is

$$\mu_1' = \bar{x} - 2 = 2{\cdot}5 - 2 = 0{\cdot}5$$

Hence

$$\mu_2' = \mu_2 + \mu_1'^2 = 14{\cdot}75 + 0{\cdot}25 = 15$$
$$\mu_3' = \mu_3 + 3\mu_2\mu_1' + \mu_1'^3 = 39{\cdot}75 + 3(14{\cdot}75)(0{\cdot}5) + (0{\cdot}5)^3$$
$$= 39{\cdot}75 + 22{\cdot}125 + 0{\cdot}125 = 62$$
$$\mu_4' = \mu_4 + 4\mu_3\mu_1' + 6\mu_2\mu_1'^2 + \mu_1'^4$$
$$= 142{\cdot}3125 + 4(39{\cdot}75)(0{\cdot}5) + 6(14{\cdot}75)(0{\cdot}5)^2 + (0{\cdot}5)^4$$
$$= 142{\cdot}3125 + 79{\cdot}5 + 22{\cdot}125 + 0{\cdot}0625$$
$$= 244$$

Example 3·9. *Calculate the first four moments of the following distribution about the mean and hence find* β_1 and β_2.

x:	0	1	2	3	4	5	6	7	8
f:	1	8	28	56	70	56	28	8	1

Solution. CALCULATION OF MOMENTS

x	f	$d = x - 4$	fd	fd^2	fd^3	fd^4
0	1	–4	–4	16	–64	256
1	8	–3	–24	72	–216	648
2	28	–2	–56	112	–224	448
3	56	–1	–56	56	–56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256
Total	256	0	0	512	0	2,816

Moments about the points $x = 4$ are

$$\mu_1' = \frac{1}{N}\Sigma fd = 0,\ \mu_2' = \frac{1}{N}\Sigma fd^2 = \frac{512}{256} = 2,$$

$$\mu_3' = \frac{1}{N}\Sigma fd^3 = 0 \text{ and } \mu_4' = \frac{1}{N}\Sigma fd^4 = \frac{2816}{256} = 11$$

Moments about mean are :

$$\mu_1 = 0,\ \mu_2 = \mu_2' - \mu_1'^2 = 2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 0$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 11$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0,\ \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{11}{4} = 2{\cdot}75$$

Example 3·10 *For a distribution the mean is 10, variance is 16, γ_2 is + 1 and β_2 is 4. Obtain the first four moments about the orgin, i.e., zero. Comment upon the nature of distribution.*

Solution. We are given

Mean = 10, $\mu_2 = 16$, $\gamma_1 = +1, \beta_1 = 4$

First four moments about origin (μ_1', μ_2', μ_3', μ_4')

μ_1' = First moment about origin = Mean = 10

$$\mu_2 = \mu_2' - \mu_1'^2 \Rightarrow \mu_2' = \mu_2 + \mu_1'^2 \Rightarrow \mu_2' = 16 + 10^2 = 116$$

we have $\gamma_1 = +1 \Rightarrow \dfrac{\mu_3}{\mu_2^{3/2}} = 1$

$$\Rightarrow \mu_3 = \mu_2^{3/2} = (16)^{3/2} = 4^3 = 64$$

$$\therefore \mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$\Rightarrow \mu_3' = \mu_3 + 3\mu_2'\mu_1' - 2\mu_1'^3$$

$$= 64 + 3 \times 116 \times 10 - 2 \times 1000 = 3544 - 2000 = 1544$$

Now $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = 4 \Rightarrow \mu_4 = 4 \times 16^2 = 1024$

and $\mu_4 = \mu'_4 - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - \mu_1'^4$

$\Rightarrow \quad \mu_4' = 1024 + 4 \times 1544 \times 10 - 6 \times 116 \times 100 + 3 \times 10000$
$= 92784 - 69600 = 23184.$

Comments of Nature of distribution : [c.f. § 3·13 and § 3·14] Since $\gamma_1 = +1$, the distribution is moderately positively skewed, i.e, if we draw the curve for the given distribution, it will have longer tail towards the right. Further since $\beta_2 = 4 > 3$, the distribution is leptokurtic, i.e., it will be more peaked than the normal curve.

Example 3·11. *If for a random variable x, the absolute moment of order k exists for ordinary k = 1, 2, ..., n–1, then the following inequalities*

(i) $\beta_\kappa^{2\kappa} \le \beta_{\kappa-1} \cdot \beta_{\kappa+1}^{\kappa}$, *(ii)* $\beta_\kappa^{1/\kappa} \le \beta_{\kappa+1}^{1/(\kappa+1)}$

holds for k=1, 2, ..., n–1, where β_κ *is the kth absolute moment about the origin.*

[Delhi Univ. B.Sc. (Stat.Hons.) 1989]

Solution. If $x_i \mid f_i$, $i = 1, 2, ..., n$ is the given frequency distribution, then

$$\beta_\kappa = \frac{1}{N} \Sigma f_i \left| x_i^k \right| \qquad ...(1)$$

Let u and v be arbitrary real numbers, then the expression

$$\sum_{i=1}^{n} f_i \left[u \mid x_i^{(k-1)/2} \mid + v \mid x_i^{(k+1)/2} \mid \right]^2, \text{ is non-negative.}$$

$$\Rightarrow \sum_{i=1}^{n} f_i \left[u \mid x_i \mid^{(k-1)/2} + v \mid x_i \mid^{(k+1)/2} \right]^2 \ge 0$$

$$\Rightarrow \mu^2 \Sigma f_i |x_i|^{k-1} + v^2 \Sigma f_i |x_i|^{k+1} + 2uv \Sigma f_i |x_i|^k \ge 0$$

Dividing throughout by N and using relation (1), we get

$$u^2 \beta_{\kappa-1} + v^2 \beta_{\kappa+1} + 2uv\beta_\kappa \ge 0, \quad i.e., \quad u^2\beta_{\kappa-1} + 2uv\beta_\kappa + v^2 \beta_{\kappa+1} \ge 0 \qquad ...(2)$$

We know that the condition for the expression $a x^2 + 2hxy + b y^2$ to be non - negative for all values of x and y is that

$$\begin{vmatrix} a & h \\ h & b \end{vmatrix} \ge 0$$

Using this result, we get from (2)

$$\begin{vmatrix} \beta_{k-1} & \beta_k \\ \beta_k & \beta_{k+1} \end{vmatrix} \ge 0$$

$$\Rightarrow \qquad \beta_{\kappa-1} \cdot \beta_{\kappa+1} - \beta_\kappa^2 \ge 0 \qquad ...(3)$$

Raising both sides of (3) to power k, we get

$$\beta_\kappa^{2\kappa} \ge \beta_{\kappa-1}^{\kappa} \cdot \beta_{\kappa+1}^{k} \qquad ...(4)$$

Putting k=1, 2, ..., k-1, k successively in (4), we get

$$\begin{aligned} \beta_1^2 &\le \beta_0 \beta_2 \\ \beta_2^4 &\le \beta_1^2 \beta_3^2 \\ \beta_3^6 &\le \beta_2^3 \beta_4^3 \\ &\cdots \\ &\cdots \\ \beta_{\kappa-1}^{2(\kappa-1)} &\le \beta_{\kappa-2}^{\kappa-1} \cdot \beta_\kappa^{\kappa-1} \end{aligned}$$

$$\beta_k^{2k} \leq \beta_{k-1}^k \beta_{k+1}^k$$

Multiplying these inequalities and noting that $\beta_0 = 1$, we get

$$\beta_k^{k+1} \leq \beta_{k+1}^k \text{ for } k = 1, 2, \ldots, n-1.$$

Raising both sides of the inequality to the power $\frac{1}{k(k+1)}$, we get

$$\beta_k^{1/k} \leq \beta_{k+1}^{1/(k+1)} \quad \ldots (5)$$

Remark. Result (5) shows that $\beta_k^{1/k}$ is an increasing function of k.

EXERCISE 3 (b)

1. (*a*) Define the raw and central moments of a frequency distribution. Obtain the relation between the central moments of order r in terms of the raw moments. What are Sheppard's corrections to the central moments ?

(*b*) Define moments. Establish the relationship between the moments about mean, *i.e.*, Central Moments in terms of moments about any arbitrary point and *vice versa*.

The first three moments of a distribution about the value 2 of the variable are 1, 16 and – 40. Show that the mean is 3, the variance is 15 and $\mu_3 = -86$. Also show that the first three moments about $x = 0$ are 3, 24 and 76.

(*c*) For a distribution the mean is 10, variance is 16, γ_1 is + 1 and β_2 is 4. Find the first four moments about the origin.

Ans. $\mu_1' = 10$, $\mu_2' = 116$, $\mu_3' = 1544$ and $\mu_4' = 23184$.

(*d*) (*i*) Define 'moment'. What is its use ? Express first four central moments in terms of moments about the origin. What is the effect of change of origin and scale on μ_3 ?

(*ii*) The first three moments of a distribution about the point $X = 7$ are 3, 11 and 15 respectively. Obtain mean, variance and β_1.

2. The first four moments of distribution about the value 5 of the variable are 2, 20, 40 and 50. Obtain as far as possible, the various characteristics of the distribution on the basis of the information given.

Ans. Mean = 7, $\mu_2 = 16$, $\mu_3 = -64$, $\mu_4 = 162$, $\beta_1 = 1$ and $\beta_2 = 0{\cdot}63$.

3. (*a*) If the first four moments of a distribution about the value 5 are equal to – 4, 22, – 117 and 560, determine the corresponding moments (*i*) about the mean, (*ii*) about zero.

(*b*) What is Sheppard's correction? What will be the corrections for the first four moments ?

The first four moments of a distribution about $x = 4$ are 1, 4, 10, 45. Show that the mean is 5 and the variance is 3 and μ_3 and μ_4 are 0 and 26 respectively,

(*c*) In certain distribution, the first four moments about the point 4 are –1·5, 17, – 13 and 308. Calculate β_1 and β_2.

(*d*) The first four moments of a frequency distribution about the point 5 are –0.55, 4·46, – 0·43 and 68·52. Find β_1 and β_2.

Ans. $\mu_2 = 4{\cdot}1575$, $\mu_3 = 6{\cdot}5962$, $\mu_4 = 75{\cdot}3944$, $\beta_1 = 0{\cdot}6055$, $\beta_2 = 4{\cdot}3619$.

4. (*a*) For the following data, calculate (*i*) Mean, (*ii*) Median, (*iii*) Semi-inter-quartile range, (*iv*) Coefficient of variation, and (*v*) β_1 and β_2 coefficients.

Wages in Rupees :	170—180	180—190	190—200	200—210	210—220	220—230	230—240	240—250
No. of Persons :	52	68	85	92	100	95	70	28

Ans. Mean = 209 (approx.); Median = 209·8; Q.D. = 15·8; σ = 19·7; *C.V.* = 9·4; $\beta_1 = 0{\cdot}003$; $\beta_2 = 26{\cdot}105$.

(*b*) Find the second, third and fourth central moments of the frequency distribution given below. Hence find (*i*) a measure of skewness (γ_1) and (*ii*) measure of kurtosis (γ_2).

Class Limits	*Frequency*
100·0 – 114·9	5
115·0 – 119·9	15
120·0 – 124·9	20
125·0 – 129·9	35
130·0 – 134·9	10
135·0 – 139·9	10
140·0 – 144·9	5

Also apply Sheppard's corrections for moments.

Ans. $\mu_2 = 2{\cdot}16$, $\mu_3 = 0{\cdot}804$, $\mu_4 = 12{\cdot}5232$

$\gamma_1 = \sqrt{\beta_1} = 0{\cdot}25298$; $\gamma_2 = \beta_2 - 3 = -0{\cdot}317$.

(*c*) The standard deviation of a symmetrical distribution is 5. What must be the value of the fourth moment about the mean in order that the distribution be (*i*) leptokurtic, (*ii*) mesokurtic, and (*iii*) platykurtic ?

Hint : $\mu_1 = \mu_3 = 0$ (because distribution is symmetrical).

$$\sigma = 5 \Rightarrow \sigma^2 = \mu_2 = 25$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{625}$$

(*i*) Distribution is leptokurtic if $\beta_2 > 3$, *i.e.*, if $\frac{\mu_4}{625} > 3 \Rightarrow \mu_4 > 1875$

(*ii*) Distribution is mesokurtic if $\beta_2 = 3 \Rightarrow$ if $\mu_4 = 1875$

(*iii*) Distribution is platykurtic if $\beta_2 < 3 \Rightarrow$ if $\mu_4 < 1875$

5. Show that for discrete distribution $\beta_2 > 1$.

[Allahabad Univ. M.A., 1993; Delhi Univ. B.Sc. (Stat. Hons), 1992]

Hint. We have to show that $\mu_4/\mu_2^2 > 1$, *i.e.*, $\mu_4 > \mu_2^2$. *If* x_i /f_i, $i = 1, 2, \ldots,$ n, be the given discrete distribution, then we have to prove that

$$\frac{1}{N}\sum_i f_i\,(x_i - \bar{x})^4 > \left(\frac{1}{N}\sum_i f_i\,(x_i - \bar{x})^2\right)^2$$

Putting $(x_i - \bar{x})^2 = z_i$, we have to show that

$$\frac{1}{N}\sum_i f_i z_i^2 > \left(\frac{1}{N}\sum_i f_i z_i\right)^2$$

$$i.e., \quad \frac{1}{N}\sum_i f_i z_i^2 - \left(\frac{1}{N}\sum_i f_i z_i\right)^2 > 0$$

$$i.e., \quad \sigma_z^2 > 0,$$

which is always true, since variance is always positive.

Hence $\beta_2 > 1$.

6. (a) The scores in Economics of 250 candidates appearing at an examination have,

Mean $= \bar{x} = 39{\cdot}72$

Variance $= \sigma^2 = 97{\cdot}80$

Third Central moment $= \mu_3 = -114{\cdot}18$

Fourth central moment $= \mu_4 = 28{,}396{\cdot}14$

It was later found on scrutiny that the score 61 of a candidate has been wrongly recorded as 51. Make necessary corrections in the given values of the mean and the central moment. **(Gujarat Univ. M.A., 1993)**

(b) For a distribution of 250 heights, calculations showed that the mean, standard deviation, β_1 and β_2 were 54 inches, 3 inches 0 and 3 inches respectively. It was, however, discovered on checking that the two items 64 and 52 in the original data were wrongly written in place of correct values 62 and 52 inches respectively. Calculate the correct frequency constants.

Ans. Correct Mean = 54, S.D. = 2·97, $\mu_3 = -2{\cdot}18$, $\mu_4 = 218{\cdot}42$, $\beta_1 = 0{\cdot}0070$ and $\beta_2 = 2{\cdot}81$

7. In calculating the moments of a frequency distribution based on 100 observations, the following results are obtained :

Mean = 9, Variance = 19, $\beta_1 = 0{\cdot}7$ (μ_3 + ive), $\beta_2 = 4$

But later on its was found that one observation 12 was read as 21. Obtain the correct value of the first four central moments.

Ans. Corrected mean = 8·91, $\mu_2 = 17{\cdot}64$, $\mu_3 = 57{\cdot}05$, $\mu_4 = 1257{\cdot}15$, $\beta_1 = 0{\cdot}59$ and $\beta_2 = 4{\cdot}04$.

8. (a) Show that if a range of six times the standard deviation covers at least 18 class intervals, Sheppard's correction will make a difference of less then 0·5 percent in the corrected value of the standard deviation.

Hint. If h is the magnitude of the class interval, then we want :

$$6\sigma > 18h \Rightarrow \sigma > 3h \Rightarrow h^2 < \frac{1}{9}\sigma^2 \Rightarrow -h^2 > -\frac{1}{9}\sigma^2$$

$$\therefore \quad \mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12} \geq \sigma^2 - \frac{1}{9 \times 12}\sigma^2 = \sigma^2\left(1 - \frac{1}{108}\right)$$

$$\Rightarrow \text{ s.d. (corrected)} \geq \sigma\left(1 - \frac{1}{108}\right)^{1/2} \simeq \sigma\left(1 - \frac{1}{2} \times \frac{1}{108}\right)$$

$$\therefore \text{ Required adjustement} = \sigma - \sigma \text{ (corrected)} < \frac{\sigma}{216} < \frac{\sigma}{200} = \frac{1}{2}\% \text{ of s.d.}$$

(*b*) Show that, if the class intervals of a grouped distribution is less than one-third of the calculated standard deviation, Sheppard's adjustment makes a difference of less than $\frac{1}{2}$ % in the estimate of the standard deviation

9. (*a*) If ∂_r is the *r*th absolute moment about zero, use the mean value of

$$[u \mid x \mid^{(r-1)/2} + v \mid x \mid^{(r+1)/2}]^2$$

to show that

$$(\delta_r)^{2r} \leq (\partial_{r-1})^r (\partial_{r+1})^r$$

From this derive the following inequalities :

(*i*) $(\partial_r)^{r+1} \leq (\delta_{r+1})^r$, (*ii*) $(\partial_r)^{1/r} \leq (\delta_{r+1})^{1/(r+1)}$

(*b*) For a random variable X moments of all order exist. Denoting by μ_j and ∂_j, the *j*th central moment and *j*th absolute moment respectively, show that

(*i*) $(\mu_{2j+1})^2 \leq \mu_{2j}\ \mu_{2j+2}$,

(*ii*) $(\partial_j)^{1/j} \leq (\partial_{j+1})^{1/(j+1)}$ **(Karnataka Univ. B.Sc., 1993)**

10. If β_1 and β_2 are the Pearsons's coefficients of skewness and Kurtosis respectively, show that $\beta_2 > \beta_1 + 1$. **(Bangalore Univ. B.Sc., 1993)**

3·13. Skewness. Literally, skewness means '*lack of symmetry*'. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. A distribution is said to be skewed if

(*i*) Mean, median and mode fall at different points,

i.e., Mean ≠ Median ≠ Mode,

(*ii*) Quartiles are not equidistant from median, and

(*iii*) The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

Measures of Skewness. Various measures of skewness are

(1) $S_k = M - M_d$ (2) $S_k = M - M_0$.

where M is the mean, M_d, the median and M_0, the mode of the distribution.

(3) $S_k = (Q_3 - M_d) - (M_d - Q_1)$.

These are the absolute measures of skewness. As in dispersion, for comparing two series we do not calculate these absolute measures but we calculate the relative measures called the *co-efficients of skewness* which are pure numbers independent of units of measurement. The following are the *coefficients of Skewness.*

1. *Prof. Karl Pearson's Coefficient of Skewness.*

$$S_k = \frac{(M - M_0)}{\sigma} \qquad \ldots (3\cdot27)$$

where σ is the standard deviation of the distribution.

If mode is ill-defined, then using the relation, $M_0 = 3M_d - 2M$, for a moderately asymmetrical distribution, we get

$$S_k = \frac{3(M - M_d)}{\sigma} \qquad \ldots (3\cdot27a)$$

The limits for Karl Pearson's coefficient of skewness are ± 3. In practice, these limits are rarely attained.

Skewness is positive if $M > M_0$ or $M > M_d$ and negative if $M < M_0$ or $M < M_d$.

II. *Prof. Bowley's Coefficient of Skewness.* Based on quartiles,

$$S_K = \frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)} = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1} \quad \ldots (3\cdot28)$$

Remarks 1. Bowley's coefficient of skewness is also known as *Quartile coefficient of skewness* and is especially useful in situations where quartiles and median are used, *viz.*,

(*i*) When the mode is ill-defined and extreme observations are present in the data.

(*ii*) When the distribution has open end classes or unequal class intervals.

In these situations Pearson's coefficient of skewness cannot be used.

2. From (3·28), we observe that

$$S_k = 0 \text{ , } if\ Q_3 - M_d = M_d - Q_1$$

This implies that for a symmetrical distribution ($S_k = 0$), median is equidistant from the upper and lower quartiles. Moreover skewness is positive if :

$$Q_3 - M_d > M_d - Q_1 \quad \Rightarrow \quad Q_3 + Q_1 > 2M_d$$

and skewness is negative if

$$Q_3 - M_d < M_d - Q_1 \quad \Rightarrow \quad Q_3 + Q_1 < 2M_d$$

3. *Limits for Bowley's Coefficient of Skewness.* We know that for two real positive numbers a and b (*i.e.*, $a > 0$ and $b > 0$), the moduls value of the difference $(a - b)$ is always less than or equal to the modules value of the sum $(a + b)$, *i.e.*,

$$|a - b| \leq |a + b| \quad \Rightarrow \quad \left|\frac{a - b}{a + b}\right| \leq 1 \quad \ldots (*)$$

We also know that $(Q_3 - M_d)$ and $(M_d - Q_1)$ are both non-negative. Thus, taking $a = Q_3 - M_d$ and $b = M_d - Q_1$, in (*), we get

$$\left|\frac{(Q_3 - M_d) - (M_d - Q_1)}{(Q_3 - M_d) + (M_d - Q_1)}\right| \leq 1$$

$$\Rightarrow \quad |S_k \text{ (Bowley)}| \leq 1$$

$$\Rightarrow \quad -1 \leq Sk \text{ (Bowley)} \leq 1.$$

Thus, Bowley's coefficient of skewness ranges from – 1 to 1.

Further, we note from (3·28) that :

$S_k = +1$, if $M_d - Q_1 = 0$, *i.e.*, if $M_d = Q_1$

$S_k = -1$, if $Q_3 - M_d = 0$, *i.e.*, if $Q_3 = M_d$.

4. It should be clearly understood that the values of the coefficients of skewness obtained by Bowley's formula and Pearson's formula are not comparable, although in each case, $S_k = 0$, implies the absence of skewness, *i.e.*, the distribution is symmetrical. It may even happen that one of them gives positive skewness while the other gives negative skewness.

5. In Bowley's coefficient of skewness the disturbing factor of variation is eliminated by dividing the absolute measure of skewness, viz., $(Q_3 - Md) - (Md - Q_1)$ by the measure of dispersion $(Q_3 - Q_1)$, i.e., quartile range.

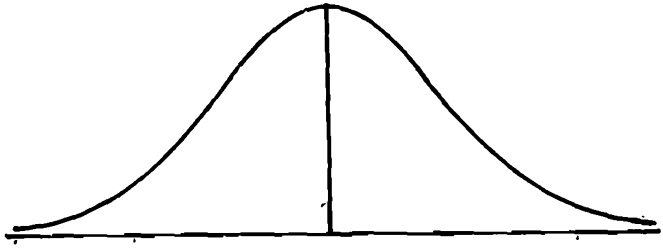
6. The only and perhaps quite serious limitations of this coefficient is that it is based only on the central 50% of the data and ignores the remaining 50% of the data towards the extremes.

III. Based upon moments, co-efficient of skewness is

$$S_k = \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2\,(5\beta_2 - 6\beta_1 - 9)} \qquad ...(3·29)$$

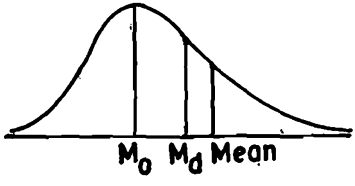
where symbols have their usual meaning. Thus $S_k = 0$ if either $\beta_1 = 0$ or $\beta_2 = -3$. But since $\beta_2 = \mu_4/\mu_2^2$, cannot be negative, $S_k = 0$ if and only if $\beta_1 = 0$. Thus for a symmetrical distribution $\beta_1 = 0$. In this respect β_1 is taken to be a measure of skewness. The co-efficient, in (3·29) is to be regarded as without sign.

We observe in (3·27) and (3·28) that skewness can be positive as well as negative. The skewness is positive if the larger tail of the distribution lies towards

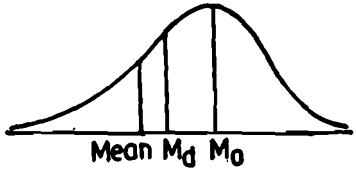


$\bar{x}$ (*Mean*) = M_0 = M_d
(Symmetrical Distribution)

the higher values of the variate (the right), i.e., if the curve drawn with the help of the given data is tretched more to the right than to the left and is negative



(Positively Skewed Distribution)



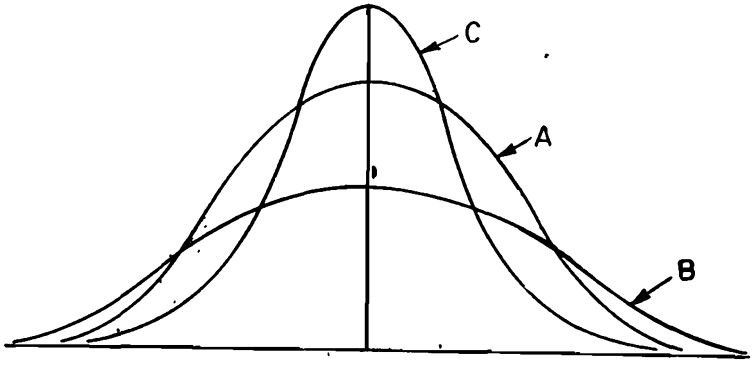
(Negatively Skewed Distribution)

in the contracy case.

3·14. Kurtosis. If we know the measures of central tendency, dispersion and skewness, we still cannot form a complete idea about the distribution as will be clear from the following figure in which all the three curves *A*, *B* and *C* are symmetrical about the mean '*m*' and have the same range.

In addition to these measures we should know one more measure which Prof. Karl Pearson calls as the '*Convexity of curve*' or *Kurtosis*. Kurtosis enables us to have an idea about the flatness or peakedness of the curve. It is measured by the co-efficient β_2 or its derivation r_2 given by

$$\beta_2 = \mu_4/\mu_2^2, \quad \gamma_2 = \beta_2 - 3$$



Curve of the type '*A*' which is neither flat nor peaked is called the *normal curve or mesokurtic curve* and for such a curve $\beta_2 = 3$, i.e., $\gamma_2 = 0$. Curve of the type '*B*' which is flatter than the normal curve is known as *platykurtic* and for such a curve $\beta_2 < 3$, *i.e.*, $\gamma_2 < 0$. Curve of the type '*C*' which is more peaked than the normal curve is called *leptokurtic* and for such a curve $\beta_2 > 3$, i.e., $\gamma_2 > 0$.

EXERCISE 3 (c)

1. What do you understand by skewness ? How is it measured ? Distinguish clearly, by giving figures, between positive and negative skewness.

2. Explain the methods of measuring skewness and kurtosis of a frequency distribution.

3. Show that for any frequency distribution :

(i) Kurtosis is *greater* than unity.

(ii) Co-efficient of skewness is *less* than 1 numerically.

4. Why do we calculate in general, only the first four moments about mean of a distribution and not the higher moments ?

5. (a) Obtain Karl Pearsons's measure of skewness for the following data:

Values	*Frequency*	*Values*	*Frequency*
5 – 10	6	25 – 30	15
10 – 15	8	30 – 35	11
15 – 20	17	35 – 40	2
20 – 25	21		

(b) Assume that a firm has selected a random sample of 100 from its production line and has obtain the data shown in the table below :

Class interval	*Frequency*	*Class interval*	*Frequency*
130 – 134	3	150 – 154	19
135 – 139	12	155 – 159	12
140 – 144	21	160 – 164	5
145 – 149	28		

Compute the following :

(a) The arithmetic mean, (b) the standard deviation,

(c) Karl Pearson's coefficient of skewness.

Ans. (a) 147·2, (b) 7·2083 (c) 0·0711

6. (a) For the frequency distribution given below, calculate the coefficient of skewness based on quartiles.

Annual Sales (*Rs.* '000)	*No. of Firms*	*Annual Sales* (*Rs.* '000)	*No. of firms*
Less than 20	30	Less than 70	644
Less than 30	225	Less than 80	650
Less than 40	465	Less than 90	665
Less than 50	580	Less than 100	680
Less than 60	634		

(b) (i) Karl Pearsons's coefficient of skewness of a distribution is 0·32, its s.d. is 6·5 and mean is 29·6. Find the mode of the distribution.

(ii) If the mode of the above distribution is 24·8, what will be the s.d. ?

7. (a) In a frequency distribution, the co-efficient of skewness based upon the quartiles lis 0·6. If the sum oif the upper and lower quartiles is 100 and median is 38, find the value of the upper and lower quartiles.

Hint. We are given

$$S_k = \frac{Q_3 + Q_1 - 2\,Md}{Q_3 - Q_1} = 0{\cdot}6 \qquad \ldots(*)$$

Also $Q_3 + Q_1 = 100$ and Median = 38

Subsituting (*), we get

$$\frac{100 - 2 \times 38}{Q_3 - Q_1} = 0{\cdot}6$$

$$\Rightarrow \qquad Q_3 - Q_1 = 40$$

Simplifying we get $Q_1 = 30,\ Q_3 = 70$

(b) A frequency distribution gives the following results :

(i) C.V. = 5 (ii) Karl Pearsons's co-efficient of skewness = 0·5

(iii) $\sigma = 2$.

Find the mean and mode of the distribution.

(c) find the C.V. of a frequency distribution given that its mean is 120, mode is 123 and Karl Pearons's co-efficient of skewness is – 0·3.

Ans. C.V. = 8·33

(d) The first three moments of distribution about the value 2 are 1, 16 and 40 respectively. Examine the skewness of the distribution.

8. The first three moments about the origin 51 Kg. calculated from the data on the weights of 25 college students are

$$\mu_1' = + 0{\cdot}4 \text{ kg.}, \sqrt{\mu_2'} = 1{\cdot}2 \text{ kg. and } (\mu_3')^{1/2} = - 0{\cdot}25\text{kg.}$$

Determine the mean, the standard deviation and coefficient oif skewness.

9. The first three moments about the origin are given by

$$\mu_1' = \frac{n+1}{2}, \ \mu_2' = \frac{(n+1)(2n+1)}{6} \text{ and } \mu_3' = \frac{n\,9n+1)^2}{4}$$

Examine the skewness of the data.

10. Find out the kurtosis of the data given below :

Class interval	0 – 10	10 – 20	20 – 30	30 – 40
Frequency	1	3	4	2.

11. Data were obtained for distribution of passengers, entering Bombay local trains over time at intervals of 15 minutes for morning and evening rush hours separately, and the following results were obtained.

	Morning hours	*Evening hours*
Arithmetic mean (Peak Hours)	8 hrs. 38 min.	5 hrs. 40 min.
Standard deviation	38·5 min.	34·9 min.
Coefficient of skewness (in 15 min. unit)	- 0·32	+ 0·17
Kurtosis measure	2·0	2·2

Interpret the result and discuss giving reasons, whether you approve of the measure of 'peak hour'.

12. (a) The standard deviation of a symmertrical distribution is 5., What must be the value of the fourth moment about the mean in order that the distribution be (i) Leptokurtic, (ii) mesokurtic, and (iii) platykurtic.

Hint. $\mu_1 = \mu_3 = 0$ (Because distribution is symmetrical), $\sigma = 5 \Rightarrow \sigma^2 = \mu_2 = 25$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{625}$$

(i) Distt. is leptokurtic if $\beta_2 > 3$ i.e., if $\frac{\mu_4}{625} > 3 \Rightarrow \mu_4 > 1875$

(ii) Distt. is mesokurtic if $\beta_2 = 3 \Rightarrow$ if $\mu_4 < 1875$

(*iii*) Distt. is platykurtic if $\beta_2 < 3 \Rightarrow$ if $\mu_4 < 1875$.

(*b*) Find the second, third and fourth central moments of the frequency distribution given below., Hence, find (*i*) a measure of skewness, and (*ii*) a measure of kurtosis (γ_2).

Class limits	*Frequency*
110·0 — 114·9	5
115·0— 119·9	15
120·0— 124·9	20
125·0— 129·9	35
130·0— 134·9	10
135·0— 139·9	10
140·0— 144·9	5

Ans. $\mu_2 = 2{\cdot}16$, $\mu_3 = 0{\cdot}804$, $\mu_4 = 12{\cdot}5232$.

$$\gamma_1 = \sqrt{\beta_1} = 0{\cdot}25298 \; ; \; \gamma_2 = \beta_2 - 3 = -\; 0{\cdot}317.$$

13. (*a*) Define Pearsonian coefficients β_1 and β_2 and discuss their utility in statistics. **[Delhi Univ. B.Sc. (Hons.), 1993]**

(*b*) What do you mean by skewness and kurtosis of a distribution ? Show that the Pearson's Beta coefficients satisfy the inequality $\beta_2 - \beta_1 - 1 \geq 0$. Also deduce that $\beta_2 \geq 1$. **(Delhi Univ. B.Sc. (Stat. Hons.), 1991)**

(*c*) Define the Pearson's coefficients γ_1 and γ_2 and discuss their utility in Statistics.

OBJECTIVE TYPE QUESTIONS

1. Match the correct parts to make a valid statement.

(*a*) Range (*i*) $(Q_3 - Q_1)/2$

(*b*) Quartile Deviation (*ii*) $\sqrt{\frac{1}{N} \Sigma f_i (x_i - \bar{x})^2}$

(*c*) Mean Deviation (*iii*) $\frac{S.D.}{Mean} \times 100$

(*d*) Standard Deviation (*iv*) $\frac{1}{N} \Sigma f_i \mid (x_1 - \bar{x}) \mid$

(*e*) Coefficient of Variation (*v*) $X_{max} - X_{min}$

II. Which value of '*a*' gives the minimum ?

(*i*) Mean square deviation from '*a*'

(*ii*) Mean deviation from '*a*'

III. Mean of 100 observations is 50 and S.D. is 10. What will be the new mean and S.D., if

(*i*) 5 is added to each observation,

(*ii*) each observation is multiplied by 3,

(*iii*) 5 is subtracted from each observation and then it is divided by 4?

IV. Fill in the blanks :

(*i*) (*a*) Absolute sum of deviation is minimum from.....

(b) Least value of root mean square deviation is .....

(ii) The sum of squares of deviations is least when measured from .....

(iii) The sum of 10 items is 12 and the sum of their squares is 16·9.

(iv) In any distribution, the standard deviation is always ..... the mean deviation.

(v) The relationship between root mean square deviation and standard deviation σ is .....

(vi) If 25% of the items are less than 10 and 25% are more than 40, the coefficient of quartile deviation is .....

(vii) The median and standard deviation of a distribution are 20 and 4 respectively. If each item is increased by 2, the median will be ..... and the new standard deviation will be .....

(viii) In a symmetric distribution, the mean and the mode are .....

(xi) In symmetric distribution, the upper and the lower quartiles are equi-distant from .....

(x) If the mean, mode and standard deviation of a frequency distribution are 41, 45 and 8 respectively, then its pearson's coefficient of skewness is .....

(xi) For a symmetrical distribution β_1 = ...

(xii) If $\beta_2 > 3$ the distribution is said to be .....

(xiii) For a symmetric distribution μ_2 = ...

μ_{2n+1} = ...

(xiv) If the mean and the mode of a given distribution are equal then its coefficient of skewness is .....

(xv) If the kurtosis of a distribution is 3, it is called ..... distribution.

(xvi) In a perfectly symmetrical distribution 50% of items are above 60 and 75% items are below 75. Therefore, the coefficient of quartile deviation is ..... and coefficient of skewness is .....

(xvii) Relation between β_1 and β_2 is given by .....

V. For the following questions given correct answers :

(i) Sum of absolute deviations about median is

(a) Least, (b) greatest, (c) zero, (d) equal.

(ii) The sum of squares of deviations is least when measured from

(a) Median, (b), (c) Mean, (d) Mode, (e) none of them.

(iii) In any discrete series (when all the values are not same) the relationship between M.D. about mean and S.D. is

(a) M.D. = S.D., (b) M.D. $\geq$ S.D., (c) M.D. < S.D.,

(d) M.D. $\leq$ S.D.

(e) None of these.

(iv) If each of a set of observations of a variable is multiplied by a constant (non-zero) value, the variance of the resultant variable.

(a) is unaltered, (b) increases (c) decreases, (d) is unknown.

(v) The appropriate measure whenever the extreme items are to be disregarded and when the distribution contains indefinite classes at the end is

(a) Median, (b) Mode, (c) Quartile deviation,
(d) Standard Deviation

(vi) A.M., G.M. and H.M. in any series are equal when

(a) the distribution is symmetric, (b) all the values are same,
(c) the distribution is positively skewed,
(d) the distribution is unimodal.

(vii) The limits for quartile coefficient of skewness are

(a) ± 3, (b) 0 and 3, (c) ± 1, (d) ± ∞

(viii) The statement that the variance is equal to the second central moment'

(a) always true, (b) sometimes true, (c) never true,
(d) ambiguous.

(ix) The standard deviation of a distribution is 5. The value of the fourth central moment (μ_4), in order that the distribution be mesokurtic should be

(a) equal to 3, (b) greater than 1,875, (c) equal to 1,875,
(d) less than 1,875.

(x) In a frequency curve of scores the mode was found to be higher than the mean. This shows that the distribution is

(a) Symmetric, (b) negatively skewed, (c) positvely skewed,
(d) normal.

(xi) For any frequency distribution, the kurtosis is

(a) greater than 1, (b) less than 1, (c) equal to 1.

(xii) The measure of kurtosis is

(a) $\beta_2 = 0$, (b) $\beta_2 = 3$, (c) $\beta_2 = 4$.

(xiii) For the distribution

(a) $\mu_4 = 0$, (b) Median = 0,
(c) The distribution of x is symmetrical.

X:	− 4	− 3	− 2	− 1	0	1	2	3	4	Total
f:	2	4	5	7	10	7	5	4	2	46

(xiv) For a symmetric distribution

(a) $\mu_2 = 0$, (b) $\mu_2 > 0$; (c) $\mu_3 > 0$

VI. State which of the following statements are Ture and which False. In each of false statements given the correct statement.

(i) Mean, standard deviation and varaince have the same unit.

(ii) Standard deviation of every distribution is unique and always exists.

(iii) Median is the value of the variance which divides the total frequency it two equal parts.

(iv) Mean – Mode = 3 (mean – median) is often approximately satisified.

(v) Mean deviation = $\frac{4}{5}$ (standard deviation) is always satisfied.

(vi) $\beta_2 \geq 1$ is always satisfied

(vii) $\beta_1 = 0$ is a conclusive test for a distribution to be symmetrical.

CHAPTER FOUR

Theory of Probability

4·1. Introduction. If an experiment is repeated under essentially homogeneous and similar conditions we generally come across two types of situations:

(i) The result or what is usually known as the *'outcome'* is unique or certain.

(ii) The result is not unique but may be one of the several possible outcomes.

The phenomena covered by *(i)* are known as *'deterministic'* or *'predictable'* phenomena. By a deterministic phenomenon we mean one in which the result can be predicted with certainty. For example :

(a) For a perfect gas,

$$V \propto \frac{1}{P} \quad i.e., \; PV = constant,$$

provided the temperature remains the same.

(b) The velocity 'v' of a particle after time 't' is given by

$$v = u + at$$

where u is the initial velocity and a is the acceleration. This equation uniquely determines v if the right-hand quantities are known.

(c) Ohm's Law, *viz.*, $C = \frac{E}{R}$

where C is the flow of current, E the potential difference between the two ends of the conductor and R the resistance, uniquely determines the value C as soon as E and R are given.

A deterministic model is defined as a model which stipulates that the conditions under which an experiment is performed determine the outcome of the experiment. For a number of situations the deterministic model suffices. However, there are phenomena [as covered by *(ii)* above] which do not lend themselves to deterministic approach and are known as *'unpredictable'* or *'probabilistic'* phenomena. For example :

(i) In tossing of a coin one is not sure if a head or tail will be obtained.

(ii) If a light tube has lasted for t hours, nothing can be said about its further life. It may fail to function any moment.

In such cases we talk of chance or probability which is taken to be a quantitative measure of certainty.

4·2. Short History. Galileo (1564-1642), an Italian mathematician, was the first to attempt at a quantitative measure of probability while dealing with some problems related to the theory of dice in gambling. But the first foundation of the mathematical theory if probability was laid in the mid-seventeenth century by two French mathematicians, B. Pascal (1623-1662) and P. Fermat (1601-1665), while

solving a number of problems posed by French gambler and noble man Chevalier-De-Mere to Pascal. The famous *'problem of points'* posed by De-Mere to Pascal is : "Two persons play a game of chance. The person who first gains a certain number of points wins the stake. They stop playing before the game is completed. How is the stake to be decided on the basis of the number of points each has won?" The two mathematicians after a lengthy correspondence between themselves ultimately solved this problem and this correspondence laid the first foundation of the science of probability. Next stalwart in this field was J. Bernoulli (1654-1705) whose 'Treatise on Probability' was published posthumously by his nephew N. Bernoulli in 1713. De-Moivre (1667-1754) also did considerable work in this field and published his famous 'Doctrine of Chances' in 1718. Other main contributors are : T. Bayes (Inverse probability), P.S. Laplace (1749-1827) who after extensive research over a number of years finally published 'Theoric analytique des probabilities' in 1812. In addition to these, other outstanding contributors are Levy, Mises and R.A. Fisher.

Russian mathematicians also have made very valuable contributions to the modern theory of probability. Chief contributors, to mention only a few of them are : Chebyshev (1821-94) who founded the Russian School of Statisticians; A. Markoff (1856-1922); Liapounoff (Central Limit Theorem); A. Khintchine (Law of Large Numbers) and A. Kolmogorov, who axiomised the calculus of probability.

4·3. Definitions of Various Terms. In this section we will define and explain the various terms which are used in the definition of probability.

Trial and Event. Consider an experiment which, though repeated under essentially identical conditions, does not give unique results but may result in any one of the several possible outcomes. The experiment is known as a *trial* and the outcomes are known as *events* or *cases*. For example :

(i) Throwing of a die is a trial and getting 1(or 2 or 3, ...or 6) is an event.

(ii) Tossing of a coin is a trial and getting head (H) or tail (T) is an event.

(iii) Drawing two cards from a pack of well-shuffled cards is a trial and getting a king and a queen are events.

Exhaustive Events. The total number of possible outcomes in any trial is known as exhaustive events or exhaustive cases. For example :

(i) In tossing of a coin there are two exhaustive cases, *viz.*, head and tail (the possibility of the coin standing on an edge being ignored).

(ii) In throwing of a die, there are six exhaustive cases since any one of the 6 faces 1, 2, ...,6 may come uppermost.

(iii) In drawing two cards from a pack of cards the exhaustive number of cases is ${}^{52}C_2$, since 2 cards can be drawn out of 52 cards in ${}^{52}C_2$ ways.

(iv) In throwing of two dice, the exhaustive number of cases is $6^2 = 36$, since any of the 6 numbers 1 to 6 on the first die can be associated with any of the six numbers on the other die.

In general in throwing of n dice the exhaustive number of cases is 6^n.

Favourable Events or Cases. The number of cases favourable to an event in a trial is the number of outcomes which entail the happening of the event. For example,

(i) In drawing a card from a pack of cards the number of cases favourable to drawing of an ace is 4, for drawing a spade is 13 and for drawing a red card is 26.

(ii) In throwing of two dice, the number of cases favourable to getting the sum 5 is : (1,4) (4,1) (2,3) (3,2), *i.e.*, 4.

Mutually exclusive events. Events are said to be *mutually exclusive* or *incompatible* if the happening of any one of them precludes the happening of all the others, *i.e.*, if no two or more of them can happen simultaneously in the same trial. For example :

(i) In throwing a die all the 6 faces numbered 1 to 6 are mutually exclusive since if any one of these faces comes, the possibility of others, in the same trial, is ruled out.

(ii) Similarly in tossing a coin the events head and tail are mutually exclusive.

Equally likely events. Outcomes of a trial are set to be equally likely if taking into consideration all the relevant evidences, there is no reason to expect one in preference to the others. For example

(i) In tossing an unbiased or uniform coin, head or tail are equally likely events.

(ii) In throwing an unbiased die, all the six faces are equally likely to come.

Independent events. Several events are said to be independent if the happening (or non-happening) of an event is not affected by the supplementary knowledge concerning the occurrence of any number of the remaining events. For example

(i) In tossing an unbiased coin the event of getting a head in the first toss is independent of getting a head in the second, third and subsequent throws.

(ii) If we draw a card from a pack of well-shuffled cards and replace it before drawing the second card, the result of the second draw is independent of the first draw. But, however, if the first card drawn is not replaced then the second draw is dependent on the first draw.

4·3·1. Mathematical or Classical or 'a priori' Probabality

Definition. If a trial results in n exhaustive, mutually exclusive and equally likely cases and m of them are favourable to the happening of an event E, then the probability 'p' of happening of E is given by

$$p = P(E) = \frac{\text{Favourable number of cases}}{\text{Exhaustive number of cases}} = \frac{m}{n} \qquad \ldots(4\cdot1)$$

Sometimes we express (4·1) by saying that 'the odds in favour of E are $m : (n - m)$ or the odds against E are $(n - m) : n$.'

Since the number of cases favourable to the 'non-happening' of the event E are $(n - m)$, the probability 'q' that E will not happen is given by

$$q = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - p \quad \Rightarrow \quad p + q = 1 \qquad \ldots(4\cdot1a)$$

Obviously p as well as q are non-negative and cannot exceed unity, *i.e*, $0 \leq p \leq 1$, $0 \leq q \leq 1$.

Remarks. 1. Probability 'p' of the happening of an event is also known as the probability of success and the probability 'q' of the non- happening of the event as the probability of failure.

2. If $P(E) = 1$, E is called a *certain event* and if $P(E) = 0$, E is called an *impossible* event.

3. Limitations of Classical Definition. This definition of Classical Probability breaks down in the following cases :

(i) If the various outcomes of the trial are not equally likely or equally probable. For example, the probability that a candidate will pass in a certain test is not 50% since the two possible outcomes, *viz.*, sucess and failure (excluding the possibility of a compartment) are not equally likely.

(ii) If the exhaustive number of cases in a trial is infinite.

4·3·2. Statistical or Empirical Probability

Definition (Von Mises). If a trial is repeated a number of times under essentially homogeneous and identical conditions, then the limiting value of the ratio of the number of times the event happens to the number of trials, as the number of trials become indefinitely large, is called the probability of happening of the event. (It is assumed that the limit is finite and unique).

Symbolically, if in n trials an event E happens m times, then the probability 'p' of the happening of E is given by

$$p = P(E) = \lim_{n \to \infty} \frac{m}{n} \qquad \ldots(4\cdot2)$$

Example 4·1. *What is the chance that a leap year selected at random will contain 53 Sundays?*

Solution. In a leap year (which consists of 366 days) there are 52 complete weeks and 2 days over. The following are the possible combinations for these two 'over' days:

(i) Sunday and Monday, *(ii)* Monday and Tuesday, *(iii)* Tuesday and Wednesday, *(iv)* Wednesday and Thursday, *(v)* Thursday and Friday, *(vi)* Friday and Saturday, and *(vii)* Saturday and Sunday.

In order that a leap year selected at random should contain 53 Sundays, one of the two 'over' days must be Sunday. Since out of the above 7 possibilities, 2 *viz.*, *(i)* and *(vii)*, are favourable to this event,

$$\therefore \qquad \text{Required probability} = \frac{2}{7}$$

Example 4·2. *A bag contains 3 red, 6 white and 7 blue balls. What is the probability that two balls drawn are white and blue?*

Solution. Total number of balls = 3 + 6 + 7 = 16.

Now, out of 16 balls, 2 can be drawn in ${}^{16}C_2$ ways.

$\therefore$ Exhaustive number of cases = ${}^{16}C_2 = \frac{16 \times 15}{2} = 120.$

Out of 6 white balls 1 ball can be drawn in ${}^{6}C_1$ ways and out of 7 blue balls 1 ball can be drawn in ${}^{7}C_1$ ways. Since each of the former cases can be associated with each of the latter cases, total number of favourable cases is : ${}^{6}C_1 \times {}^{7}C_1 = 6 \times 7 = 42.$

$\therefore$ Required probability $= \frac{42}{120} = \frac{7}{20}.$

Example 4·3. (a) *Two cards are drawn at random from a well-shuffled pack of 52 cards. Show that the chance of drawing two aces is 1/221.*

(b) From a pack of 52 cards, three are drawn at random. Find the chance that they are a king, a queen and a knave.

(c) Four cards are drawn from a pack of cards. Find the probability that (i) all are diamond, (ii) there is one card of each suit, and (iii) there are two spades and two hearts.

Solution. *(a)* From a pack of 52 cards 2 cards can be drawn in ${}^{52}C_2$ ways, all being equally likely.

$\therefore$ Exhaustive number of cases = ${}^{52}C_2$

In a pack there are 4 aces and therefore 2 aces can be drawn in ${}^{4}C_2$ ways.

$\therefore$ Required probability $= \frac{{}^{4}C_2}{{}^{52}C_2} = \frac{4 \times 3}{2} \times \frac{2}{52 \times 51} = \frac{1}{221}$

(b) Exhaustive number of cases = ${}^{52}C_3$

A pack of cards contains 4 kings, 4 queens and 4 knaves. A king, a queen and a knave can each be drawn in ${}^{4}C_1$ ways and since each way of drawing a king can be associated with each of the ways of drawing a queen and a knave, the total number of favourable cases = ${}^{4}C_1 \times {}^{4}C_1 \times {}^{4}C_1$

$\therefore$ Required probability $= \frac{{}^{4}C_1 \times {}^{4}C_1 \times {}^{4}C_1}{{}^{52}C_3} = \frac{4 \times 4 \times 4 \times 6}{52 \times 51 \times 50} = \frac{16}{5525}$

(c) Exhaustive number of cases = ${}^{52}C_4$.

(i) Required probability $= \frac{{}^{13}C_4}{{}^{52}C_4}$

(ii) Required probability $= \frac{{}^{13}C_1 \times {}^{13}C_1 \times {}^{13}C_1 \times {}^{13}C_1}{{}^{52}C_4}$

(iii) Required probability $= \frac{{}^{13}C_2 \times {}^{13}C_2}{{}^{52}C_4}$

Example 4·4. *What is the probability of getting 9 cards of the same suit in one hand at a game of bridge?*

Solution. One hand in a game of bridge consists of 13 cards.

$\therefore$ Exhaustive number of cases $= {}^{52}C_{13}$

Number of ways in which, in one hand, a particular player gets 9 cards of one suit are ${}^{13}C_9$ and the number of ways in which the remaining 4 cards are of some other suit are ${}^{39}C_4$. Since there are 4 suits in a pack of cards, total number of favourable cases $= 4 \times {}^{13}C_9 \times {}^{39}C_4$.

$$\therefore \quad \text{Required probability} = \frac{4 \times {}^{13}C_9 \times {}^{39}C_4}{{}^{52}C_{13}}$$

Example 4·5. *(a) Among the digits 1, 2, 3, 4, 5, at first one is chosen and then a second selection is made among the remaining four digits. Assuming that all twenty possible outcomes have equal probabilities, find the probability that an odd digit will be selected (i) the first time, (ii) the second time, and (iii) both times.*

(b) From 25 tickets, marked with the first 25 numerals, one is drawn at random. Find the chance that

(i) it is a multiple of 5 or 7,

(ii) it is a multiple of 3 or 7.

Solution. *(a)* Total number of cases $= 5 \times 4 = 20$

(i) Now there are 12 cases in which the first digit drawn is odd, *viz.*, (1, 2), (1, 3), (1, 4), (1, 5), (3, 1), (3, 2), (3, 4), (3, 5), (5, 1), (5, 2), (5, 3) and (5, 4).

$\therefore$ The probability that the first digit drawn is odd

$$= \frac{12}{20} = \frac{3}{5}$$

(ii) Also there are 12 cases in which the second digit drawn is odd, *viz.*, (2, 1), (3, 1), (4, 1), (5, 1), (1, 3), (2, 3), (4, 3), (5, 3), (1, 5), (2, 5), (3, 5) and (4, 5).

$\therefore$ The probability that the second digit drawn is odd

$$= \frac{12}{20} = \frac{3}{5}$$

(iii) There are six cases in which both the digits drawn are odd, *viz.*, (1, 3), (1, 5), (3, 1), (3, 5), (5, 1) and (5, 3).

$\therefore$ The probability that both the digits drawn are odd

$$= \frac{6}{20} = \frac{3}{10}$$

(b) (i) Numbers (out of the first 25 numerals) which are multiples of 5 are 5, 10, 15, 20 and 25, *i.e.*, 5 in all and the numbers which are multiples of 7 are 7, 14 and 21, *i.e.*, 3 in all. Hence required number of favourable cases are 5+3=8.

$$\therefore \quad \text{Required probability} = \frac{8}{25}$$

(ii) Numbers (among the first 25 numerals) which are multiples of 3 are 3, 6, 9, 12, 15, 18, 21, 24, *i.e.*, 8 in all, and the numbers which are multiples of 7 are 7,

14, 21, *i.e.*, 3 in all. Since the number 21 is common in both the cases, the required number of distinct favourable cases is $8 + 3 - 1 = 10$.

$$\therefore \quad \text{Required probability} = \frac{10}{25} = \frac{2}{5}$$

Example 4·6. *A committee of 4 people is to be appointed from 3 officers of the production department, 4 officers of the purchase department, two officers of the sales department and 1 chartered accountant. Find the probability of forming the committee in the following manner:*

(i) There must be one from each category.

(ii) It should have at least one from the purchase department.

(iii) The chartered accountant must be in the committee.

Solution. There are 3+4+2+1=10 persons in all and a committee of 4 people can be formed out of them in ${}^{10}C_4$ ways. Hence exhaustive number of cases is

$${}^{10}C_4 = \frac{10 \times 9 \times 8 \times 7}{4\ !} = 210$$

(i) Favourable number of cases for the committee to consist of 4 members, one from each category is :

$${}^4C_1 \times {}^3C_1 \times {}^2C_1 \times 1 = 4 \times 3 \times 2 = 24$$

$$\therefore \quad \text{Required probability} = \frac{24}{210} = \frac{8}{70}$$

(ii) P [Committee has at least one purchase officer]

$= 1 - P$ [Committee has no purchase officer]

In order that the committee has no purchase officer, all the 4 members are to be selected from amongst officers of production department, sales department and chartered accountant, *i.e.*, out of 3+2+1=6 members and this can be done in ${}^6C_4 = \frac{6 \times 5}{1 \times 2} = 15$ ways. Hence

$$P \text{ [Committee has no purchase officer]} = \frac{15}{210} = \frac{1}{14}$$

$$\therefore \quad P \text{ [Committee has at least one purchase officer]} = 1 - \frac{1}{14} = \frac{13}{14}$$

(iii) Favourable number of cases that the committee consists of a chartered accountant as a member and three others are :

$$1 \times {}^9C_3 = \frac{9 \times 8 \times 7}{1 \times 2 \times 3} = 84 \text{ ways,}$$

since a chartered accountant can be selected out of one chartered accountant in only 1 way and the remaining 3 members can be selected out of the remaining $10 - 1 = 9$ persons in ${}^9C_3$ ways. Hence the required probability $= \frac{84}{210} = \frac{2}{5}$.

Example 4·7. *(a) If the letters of the word 'REGULATIONS' be arranged at random, what is the chance that there will be exactly 4 letters between R and E?*

(b) What is the probability that four S's come consecutively in the word 'MISSISSIPPI'?

Solution. *(a)* The word *'REGULATIONS'* consists of 11 letters. The two letters R and E can occupy ${}^{11}P_2$, *i.e.*, $11 \times 10 = 110$ positions.

The number of ways in which there will be exactly 4 letters between R and E are enumerated below:

(i) R is in the 1st place and E is in the 6th place.

(ii) R is in the 2nd place and E is in the 7th place.

... ...

... ...

... ...

(vi) R is in the 6th place and E is in the 11th place.

Since R and E can interchange their positions, the required number of favourable cases is $2 \times 6 = 12$

$\therefore$ The required probability $= \frac{12}{110} = \frac{6}{55}$.

(b) Total number of permutations of the 11 letters of the word *'MISSISSIPPI'*, in which 4 are of one kind *(viz., S)*, 4 of other kind *(viz., I)*, 2 of third kind *(viz., P)* and 1 of fourth kind *(viz., M)* are

$$\frac{11\,!}{4\,!\;4\,!\;2\,!\;1\,!}$$

Following are the 8 possible combinations of 4 S's coming consecutively:

(i) $S\ S\ S\ S$

(ii) $—\ S\ S\ S\ S$

(iii) $—\ —\ S\ S\ S\ S$

⋮ ⋮

(viii) $—\ —\ —\ —\ —\ —\ —\ S\ S\ S\ S$

Since in each of the above cases, the total number of arrangements of the remaining 7 letters, *viz.*, *MIIIPPI* of which 4 are of one kind, 2 of other kind and one of third kind are $\frac{7\,!}{4\,!\;2\,!\;1\,!}$, the required number of favourable cases

$$= \frac{8 \times 7\,!}{4\,!\;2\,!\;1\,!}$$

$\therefore$ Required probability $= \frac{8 \times 7\,!}{4\,!\;2\,!\;1\,!} + \frac{11\,!}{4\,!\;4\,!\;2\,!\;1\,!}$

$$= \frac{8 \times 7\,! \times 4\,!}{11\,!} = \frac{4}{165}$$

Example 4·8. *Each coefficient in the equation $ax^2 + bx + c = 0$ is determined by throwing an ordinary die. Find the probability that the equation will have real roots.* **[Madras Univ. B. Sc. (Stat. Main), 1992]**

Solution. The roots of the equation $ax^2 + bx + c = 0$ ...(*)
will be real if its discriminant is non-negative, *i.e.*, if

$$b^2 - 4ac \geq 0 \quad \Rightarrow \quad b^2 \geq 4ac$$

Since each co-efficient in equation (*) is determined by throwing an ordinary die, each of the co-efficients a, b and c can take the values from 1 to 6.

$\therefore$ Total number of possible outcomes (all being equally likely)

$$= 6 \times 6 \times 6 = 216$$

The number of favourable cases can be enumerated as follows:

ac	a	c	$4ac$	b (*so that* $b^2 \geq 4ac$)	*No. of cases*
1	1	1	4	2, 3, 4, 5	$1 \times 5 = 5$
2 (*i*)	1	2	8	3, 4, 5, 6	$2 \times 4 = 8$
(*ii*)	2	1			
3 (*i*)	1	3	12	4, 5, 6	$2 \times 3 = 6$
(*ii*)	3	1			
4 (*i*)	1	4			
(*ii*)	4	1	16	4, 5, 6	$3 \times 3 = 9$
(*iii*)	2	2			
5 (*i*)	1	5	20	5, 6	$2 \times 2 = 4$
(*ii*)	5	1			
6 (*i*)	1	6			
(*ii*)	6	1	24	5, 6	$4 \times 2 = 8$
(*iii*)	3	2			
(*iv*)	2	3			
7 ($ac = 7$ *is not possible*)					
8 (*i*)	2	4	32	6	$2 \times 1 = 2$
(*ii*)	4	2			
9	3	3	36	6	1
					Total = 43

Since $b^2 \geq 4ac$ and since the maximum value of b^2 is 36, $ac = 10, 11, 12, \ldots$ etc. is not possible.

Hence total number of favourable cases = 43.

$\therefore$ Required probability $= \dfrac{43}{216}$

Example 4·9. *The sum of two non-negative quantities is equal to 2n. Find the chance that their product is not less than $\frac{3}{4}$ times their greatest product.*

Solution. Let $x > 0$ and $y > 0$ be the given quantities so that $x + y = 2n$.

We know that the product of two positive quantities whose sum is constant is greatest when the quantities are equal. Thus the product of x and y is maximum when $x = y = n$.

$\therefore$ Maximum product $= n \cdot n = n^2$

Now $P\left[xy < \frac{3}{4}n^2\right] = P\left[xy \geq \frac{3}{4}n^2\right] = P\left[x(2n-x) \geq \frac{3}{4}n^2\right]$

$= P\left[(4x^2 - 8nx + 3n^2) \leq 0\right]$

$= P\left[(2x - 3n)(2x - n) \leq 0\right]$

$= P\left[x \text{ lies between } \frac{n}{2} \text{ and } \frac{3n}{2}\right]$

$\therefore$ Favourable range $= \frac{3n}{2} - \frac{n}{2} = n$

Total range $= 2n$

$\therefore$ Required probability $= \frac{n}{2n} = \frac{1}{2}$

Example 4·10. *Out of (2n+1) tickets consecutively numbered three are drawn at random. Find the chance that the numbers on them are in A.P.*

[Calicut Univ. B.Sc., 1991; Delhi Univ. B.Sc.(Stat. Hons.), 1992]

Solution. Since out of $(2n + 1)$ tickets, 3 tickets can be drawn in ${}^{2n+1}C_3$ ways,

$$\text{Exhaustive number of cases} = {}^{2n+1}C_3 = \frac{(2n+1)\,2n\,(2n-1)}{3\,!}$$

$$= \frac{n(4n^2-1)}{3}$$

To find the favourable number of cases we are to enumerate all the cases in which the numbers on the drawn tickets are in A.P with common difference, (say $d = 1, 2, 3, \ldots, n-1, n$).

If $d = 1$, the possible cases are as follows:

$$\left.\begin{array}{c} 1,\ 2,\ 3 \\ 2,\ 3,\ 4 \\ \cdot\ \cdot\ \cdot \\ \cdot\ \cdot\ \cdot \\ \cdot\ \cdot\ \cdot \\ 2n-1,\ n,\ 2n+1 \end{array}\right\}, \textit{ i.e., } (2n-1) \text{ cases in all}$$

If $d = 2$, the possible cases are as follows :

$$\left.\begin{array}{ccc} 1, & 3, & 5 \\ 2, & 4, & 6 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 2n-3, & 2n-1, & 2n+1 \end{array}\right\}, \textit{ i.e., } (2n-3) \text{ cases in all}$$

and so on.

If $d = n-1$, the possible cases are as follows:

$$\left.\begin{array}{lll} 1, & n, & 2n-1 \\ 2, & n+1, & 2n \\ 3, & n+2, & 2n+1 \end{array}\right\}, \text{ i.e., 3 cases in all}$$

If $d = n$, there is only one case, *viz.*, $(1, n+1, 2n+1)$.

Hence total number of favourable cases

$$= (2n-1) + (2n-3)) + \ldots + 5 + 3 + 1$$
$$= 1 + 3 + 5 + \ldots + (2n-1),$$

which is a series in A.P. with $d = 2$ and n terms.

$\therefore$ Number of favourable cases $= \frac{n}{2}[1 + (2n-1)] = n^2$

$\therefore$ Required probability $= \frac{n^2}{n(4n^2-1)/3} = \frac{3n}{(4n^2-1)}$

EXCERCISE 4 (a)

1. (*a*) Give the classical and statistical definitions of probability. What are the objections raised in these definitions?

[Delhi Univ. B.Sc. (Stat. Hons.), 1988, 1985]

(*b*) When are a number of cases said to be equally likely? Give an example each of the following :

(*i*) the equally likely cases,

(*ii*) four cases which are not equally likely, and

(*iii*) five cases in which one case is more likely than the other four.

(*c*) What is meant by mutually exclusive events? Give an example of

(*i*) three mutually exclusive events,

(*ii*) three events which are not mutually exclusive.

[Meerut Univ. B.Sc. (Stat.), 1987]

(*d*) Can

(*i*) events be mutually exclusive and exhaustive?

(*ii*) events be exhaustive and indepenent?

(*iii*) events be mutually exclusive and independent?

(*iv*) events be mutually exhaustive, exclusive and independent?

2. (*a*) Prove that the probability of obtaining a total of 9 in a single throw with two dice is one by nine.

(*b*) Prove that in a single throw with a pair of dice the probability of getting the sum of 7 is equal to 1/6 and the probability of getting the sum of 10 is equal to 1/12.

(*c*) Show that in a single throw with two dice, the chance of throwing more than seven is equal to that of throwing less than seven.

Ans. 5/12 **[Delhi Univ. B.Sc., 1987, 1985]**

(*d*) In a single throw with two dice, what is the number whose probability is minimum?

(e) Two persons A and B throw three dice (six faced). If A throws 14, find B's chance of throwing a higher number. **[Meerut Univ. B.Sc.(Stat.), 1987]**

3. *(a)* A bag contains 7 white, 6 red and 5 black balls. Two balls are drawn at random. Find the probability that they will both be white.

Ans. 21/153

(b) A bag contains 10 white, 6 red, 4 black and 7 blue balls. 5 balls are drawn at random. What is the probability that 2 of them are red and one black?

Ans. ${}^6C_2 \times {}^4C_1 / {}^{27}C_5$

4. *(a)* From a set of raffle tickets numbered 1 to 100, three are drawn at random. What is the probability that all the tickets are odd-numbered?

Ans. ${}^{50}C_3 / {}^{100}C_3$

(b) A number is chosen from each of the two sets :

$$(1, 2, 3, 4, 5, 6, 7, 8, 9); \quad (4, 5, 6, 7, 8, 9)$$

If p_1 is the probability that the sum of the two numbers be 10 and p_2 the probability that their sum be 8, find $p_1 + p_2$.

(c) Two different digits are chosen at random from the set 1,2,3,...,8. Show that the probability that the sum of the digits will be equal to 5 is the same as the probability that their sum will exceed 13, each being 1/14. Also show that the chance of both digits exceeding 5 is 3/28. **[Nagpur Univ. B.Sc., 1992]**

5. What is the chance that *(i)* a leap year selected at random will contain 53 Sundays? *(ii)* a non-leap year selected at random would contain 53 Sundays.

Ans. (i) 2/7, (ii) 1/7

6. *(a)* What is the probability of having a knave and a queen when two cards are drawn from a pack of 52 cards?

Ans. 8/663

(b) Seven cards are drawn at random from a pack of 52 cards. What is the probability that 4 will be red and 3 black?

Ans. ${}^{26}C_4 \times {}^{26}C_3 / {}^{52}C_7$

(c) A card is drawn from an ordinary pack and a gambler bets that it is a spade or an ace. What are the odds against his winning the bet?

Ans. 9:4

(d) Two cards are drawn from a pack of 52 cards. What is the chance that

(i) they belong to the same suit?

(ii) they belong to different suits and different denominations.

[Bombay Univ. B.Sc., 1986]

7. *(a)* If the letters of the word RANDOM be arranged at random, what is the chance that there are exactly two letters between A and O.

(b) Find the probability that in a random arrangement of the leters of the word 'UNIVERSITY', the two I's do not come together.

(c) In random arrangements of the letters of the word 'ENGINEERING', what is the probability that vowels always occur together?

[Kurushetra Univ. B.E., 1991]

(d) Letters are drawn one at a time from a box containing the letters A, H, M, O, S, T. What is the probability that the letters in the order drawn spell the word 'Thomas'?

8. A letter is taken out at random out of 'ASSISTANT' and a letter out of 'STATISTIC'. What is the chance that they are the same letters?

9. *(a)* Twelve persons amongst whom are x and y sit down at random at a round table. What is the probability that there are two persons between x and y?

(b) A and B stand in a line at random with 10 other people. What is the probability that there will be 3 persons between A and B?

10. *(a)* If from a lot of 30 tickets marked 1, 2, 3, ..., 30 four tickets are drawn, what is the chance that those marked 1 and 2 are among them?

Ans. 2/145

(b) A bag contains 50 tickets numbered 1, 2, 3, ..., 50 of which five are drawn at random and arranged in ascending order of the magnitude ($x_1 < x_2 < x_3 < x_4 < x_5$). What is the probability that $x_3 = 30$?

Hint. *(a)* Exhaustive number of cases $= {}^{30}C_4$

If, of the four tickets drawn, two tickets bear the numbers 1 and 2, the remaining 2 must have come out of 28 tickets numbered from 3 to 30 and this can be done in ${}^{28}C_2$ ways.

$\therefore$ Favourable number of cases $= {}^{28}C_2$

(b) Exhaustive number of cases $= {}^{50}C_5$

If $x_3 = 30$, then the two tickets with numbers x_1 and x_2 must have come out of 29 tickets numbered from 1 to 29 and this can be done in ${}^{29}C_2$ ways, and the other two tickets with numbers x_4 and x_5 must have been drawn out of 20 tickets numbered from 31 to 50 and this can be done in ${}^{20}C_2$ ways.

$\therefore$ No. of favourable cases $= {}^{29}C_2 \times {}^{20}C_2$.

11. Four persons are chosen ar random from a group containing 3 men, 2 women and 4 children. Show that the chance that exactly two of them will be children is 10/21. **[Delhi Univ. B.A.1988]**

Ans. $\dfrac{{}^4C_2 \times {}^5C_2}{{}^9C_4} = \dfrac{10}{21}$

12. From a group of 3 Indians, 4 Pakistanis and 5 Americans a sub-committee of four people is selected by lots. Find the probability that the sub-committee will consist of

(i) 2 Indians and 2 Pakistanis

(ii) 1 Indian, 1 Pakistani and 2 Americans

(iii) 4 Americans [Madras Univ. B.Sc.(Main Stat.), 1987]

Ans. (*i*) $\frac{{}^3C_2 \times {}^4C_2}{{}^{12}C_4}$, (*ii*) $\frac{{}^3C_1 \times {}^4C_1 \times {}^5C_2}{{}^{12}C_4}$, (*iii*) $\frac{{}^5C_4}{{}^{12}C_4}$

13. In a box there are 4 granite stones, 5 sand stones and 6 bricks of identical size and shape. Out of them 3 are chosen at random. Find the chance that:

(i) They all belong to different varieties.

(ii) They all belong to the same variety.

(iii) They are all granite stones. **(Madras Univ. B.Sc., Oct. 1992)**

14. If n people are seated at a round table, what is the chance that two named individuals will be next to each other?

Ans. 2/(n-1)

15. Four tickets marked 00, 01, 10 and 11 respectively are placed in a bag. A ticket is drawn at random five times, being replaced each time. Find the probability that the sum of the numbers on tickets thus drawn is 23.

[Delhi Univ. B.Sc.(Subs.), 1988]

16. From a group of 25 persons, what is the probability that all 25 will have different birthdays? Assume a 365 day year and that all days are equally likely.

[Delhi Univ. B.Sc.(Maths Hons.), 1987]

Hint. $(365 \times 364 \times \ldots \times 341) \div (365)^{25}$

4·4. Mathematical Tools : Preliminary Notions of Sets. The set theory was developed by the German mathematician, G. Cantor (1845–1918).

4·4·1. Sets and Elements of Sets. A set is a well defined collection or aggregate of all possible objects having given properties and specified according to a well defined rule. The objects comprising a set are called elements, members or points of the set. Sets are often denoted by capital letters, *viz.*, A, B, C, etc. If x is an element of the set A, we write symbolically $x \in A$ (x belongs to A). If x is not a member of the set A, we write $x \notin A$ (x does not belong to A). Sets are often described by describing the properties possessed by their members. Thus the set A of all non-negative rational numbers with square less than 2 will be written as $A = \{x : x \text{ rational}, x \geq 0, x^2 < 2\}$.

If every element of the set A belongs to the set B, *i.e.*, if $x \in A \Rightarrow x \in B$, then we say that A is a subset of B and write symbolically $A \subseteq B$ (A is contained in B) or $B \supseteq A$ (B contains A). Two sets A and B are said to be *equal* or *identical* if $A \subseteq B$ and $B \subseteq A$ and we write $A = B$ or $B = A$.

A *null* or an *empty* set is one which does not contain any element at all and is denoted by ϕ.

Remarks. 1. Every set is a subset of itself.

2. An empty set is subset of every set.

3. A set containing only one element is conceptually distinct from the element itself, but will be represented by the same symbol for the sake of convenience.

4. As will be the case in all our applications of set theory, especially to probability theory, we shall have a fixed set S (say) given in advance, and we shall

be concerned only with subsets of this given set. The underlying set S may vary from one application to another, and it will be referred to as *universal set* of each particular discourse.

4·4·2. Operation on Sets

The union of two given sets A and B, denoted by $A \cup B$, is defined as a set consisting of all those points which belong to either A or B or both. Thus symbolically,

$$A \cup B = \{ x : x \in A \text{ or } x \in B \}.$$

Similarly

$$\bigcup_{i=1}^{n} A_i = \{ x : x \in A_i \text{ for at least one } i = 1, 2, \ldots, n \}$$

The *intersection* of two sets A and B, denoted by $A \cap B$, is defined as a set consisting of all those elements which belong to both A and B. Thus

$$A \cap B = \{ x : x \in A \text{ and } x \in B \}.$$

Similarly

$$\bigcap_{i=1}^{n} A_i = \{ x : x \in A_i \text{ for all } i = 1, 2, \ldots, n \}$$

For example, if $A = \{1, 2, 5, 8, 10\}$ and $B = \{2, 4, 8, 12\}$, then

$A \cup B = \{1, 2, 4, 5, 8, 10, 12\}$ and $A \cap B = \{2, 8\}$.

If A and B have no common point, *i.e.*, $A \cap B = \phi$, then the sets A and B are said to be *disjoint, mutually exclusive* or *non-overlapping*.

The *relative difference* of a set A from another set B, denoted by $A-B$ is defined as a set consisting of those elements of A which do not belong to B. Symbolically,

$$A - B = \{ x : x \in A \text{ and } x \notin B \}.$$

The *complement* or *negative* of any set A, denoted by $\bar{A}$ is a set containing all elements of the universal set S, (say), that are not elements of A, *i.e.*, $\bar{A} = S - A$.

4·4·3. Algebra of Sets

Now we state certain important properties concerning operations on sets. If A, B and C are the subsets of a universal set S, then the following laws hold:

Commutative Law : $A \cup B = B \cup A, A \cap B = B \cap A$

Associative Law : $(A \cup B) \cup C = A \cup (B \cup C)$

$(A \cap B) \cap C = A \cap (B \cap C)$

Distributive Law : $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Complementary Law : $A \cup \bar{A} = S, A \cap \bar{A} = \phi$

$A \cup S = S, (\because A \subset S), A \cap S = A$

$A \cup \phi = A, A \cap \phi = \phi$

Difference Law : $A - B = A \cap \bar{B}$

$A - B = A - (A \cap B) = (A \cup B) - B$

$A - (B - C) = (A - B) \cup (A - C)$.

$$(A \cup B) - C = (A - C) \cup (B - C)$$
$$A - (B \cup C) = (A - B) \cap (A - C)$$
$$(A \cap B) \cup (A - B) = A\,,\ (A \cap B) \cap (A - B) = \phi$$

De-Morgan's Law — Dualization Law.

$$\overline{(A \cup B)} = \overline{A} \cap \overline{B} \text{ and } \overline{(A \cap B)} = \overline{A} \cup \overline{B}$$

More generally

$$\overline{\left(\bigcup_{i=1}^{n} A_i\right)} = \bigcap_{i=1}^{n} \overline{A_i} \text{ and } \overline{\left(\bigcap_{i=1}^{n} A_i\right)} = \bigcup_{i=1}^{n} \overline{A_i}$$

Involution Law : $\overline{(\overline{A})} = A$

Idempotency Law : $A \cup A = A\,,\ A \cap A = A$

4·4·4. Limit of Sequence of Sets

Let $\{A_n\}$ be a sequence of sets in S. The *limit supremum* or *limit superior* of the sequence, usually written as $\lim \sup A_n$, is the set of all those elements which belong to A_n for infinitely many n. Thus

$$\lim_{n\to\infty} \sup A_n = \{ x : x \in A_n \text{ for infinitely many } n \} \qquad \ldots(4\cdot3)$$

The set of all those elements which belong to A_n for all but a finite number of n is called *limit infinimum or limit inferior* of the sequence and is denoted by $\lim \inf A_n$. Thus

$$\lim_{n\to\infty} \inf A_n = \{ x : x \in A_n \text{ for all but a finite number of } n \} \qquad \ldots(4\cdot3\,a)$$

The sequence $\{A_n\}$ is said to have a limit if and only if $\lim \sup A_n = \lim \inf A_n$ and this common value gives the limit of the sequence.

Theorem 4·1. *lim sup* $A_n = \bigcap_{m=1}^{\infty} \left(\bigcup_{n=m}^{\infty} A_n \right)$

and *lim inf* $A_n = \bigcup_{m=1}^{\infty} \left(\bigcap_{n=m}^{\infty} A_n \right)$

Def. $\{A_n\}$ is a monotone (infinite) sequence of sets if either

(i) $A_n \subset A_{n+1}\ \forall\ n$ or *(ii)* $A_n \supset A_{n+1}\ \forall\ n$.

In the former case the sequence $\{A_n\}$ is said to be *non-decreasing sequence* and is usually expressed as $A_n \uparrow$ and in the latter case it is said to be *non-increasing sequence* and is expressed as $A_n \downarrow$.

For a monotone sequence (non-increasing or non-decreasing), the limit always exists and we have,

$$\lim_{n\to\infty} A_n = \begin{cases} \bigcup_{n=1}^{\infty} A_n \text{ in case } (i)\,,\ i.e.,\ A_n \uparrow \\ \bigcap_{n=1}^{\infty} A_n \text{ in case } (ii)\,,\ i.e.,\ A_n \downarrow \end{cases}$$

4·4·5. Classes of Sets. A group of sets will be termed as a *class* (of sets). Below we shall define some useful types of classes.

A ring R is a *non-empty* class of sets which is closed under the formation of 'finite unions' and 'difference',

i.e., if $A \in \mathbf{R}$, $B \in \mathbf{R}$, then $A \cup B \in \mathbf{R}$ and $A - B \in \mathbf{R}$.

Obviously ϕ is a member of every ring.

A *field* F (or an *algebra*) is a non-empty class of sets which is closed under the formation of finite unions and under complementation. Thus

(i) $A \in \mathbf{F}, B \in \mathbf{F} \Rightarrow A \cup B \in \mathbf{F}$ and

(ii) $A \in \mathbf{F} \Rightarrow \bar{A} \in \mathbf{F}$.

A σ-*ring* C is a non-empty class of sets which is closed under the formation of 'countable unions' and 'difference'. Thus

(i) $A_i \in \mathbf{C},\ i = 1, 2, \ldots \quad \Rightarrow \quad \bigcup_{i=1}^{\infty} A_i \in \mathbf{C}$

(ii) $A \in \mathbf{C}, B \in \mathbf{C} \quad \Rightarrow \quad A - B \in \mathbf{C}$.

More precisesly σ-ring is a ring which is closed under the formation of countable unions.

A σ field (or σ-algebra) B is a non-empty class of sets that is closed under the formation of 'countable unions' and complementations,

i.e.,

(i) $A_i \in \mathbf{B}, i = 1, 2, \ldots \quad \Rightarrow \quad \bigcup_{i=1}^{\infty} A_i \in \mathbf{B}$.

(ii) $A \in \mathbf{B} \quad \Rightarrow \quad \bar{A} \in \mathbf{B}$.

σ-field may also be defined as a field which is closed under the formation of countable unions.

4·5. Axiomatic Approach to Probability. The axiomatic approach to probability, which closely relates the theory of probability with the modern metric theory of functions and also set theory, was proposed by A.N. Kolmogorov, a Russian mathematician, in 1933. The axiomatic definition of probability includes 'both' the classical and the statistical definitions as particular cases and overcomes the deficiencies of each of them. On this basis, it is possible to construct a logically perfect structure of the modern theory of probability and at the same time to satisfy the enchanced requirements of modern natural science. The axiomatic development of mathematical theory of probability relies entirely upon the logic of deduction.

The diverse theorems of probability, as were available prior to 1933, were finally brought together into a unified axiomised system in 1933. It is important to remark that probability theory, in common with all axiomatic mathematical systems, is concerned solely with relations among undefined things.

The axioms thus provide a set of rules which define relationships between abstract entities. These rules can be used to deduce theorems, and the theorems can

be brought together to deduce more complex theorems. These theorems have no empirical meaning although they can be given an interpretation in terms of empirical phenomenon. The important point, however, is that the mathematical development of probability theory is, in no way, conditional upon the interpretation given to the theory.

More precisely, under axiomatic approach, the probability can be deduced from mathematical concepts. To start with some concepts are laid down. Then some statements are made in respect of the properties possessed by these concepts. These properties, often termed as "*axioms*" of the theory, are used to frame some theorems. These theorems are framed without any reference to the real world and are deductions from the axioms of the theory.

4·5·1. Random Experiment, Sample Space. The theory of probability provides *mathematical models* for "real-world phenomenon" involving games of chance such as the tossing of coins and dice. We feel intuitively that statements such as

(i) "The probability of getting a "head" in one toss of an unbiased coin is 1/2"

(ii) "The probability of getting a "four" in a single toss of an unbiased die is 1/6",

should hold. We also feel that the probability of obtaining *either* a "5" or a "6" in a single throw of a die, should be the sum of the probabilities of a "5" and a "6", *viz.*, 1/6+1/6=1/3. That is, probabilities should have some kind of *additive* property. Finally, we feel that in a large number of repetitions of, for example, a coin tossing experiment, the proportion of heads should be approximately 1/2. That is, the probability should have a *frequency interpretation.*

To deal with these properties sensibly, we need a *mathematical description* or *model* for the probabilistic situation we have. Any such probabilistic situation is referred to as a *random experiment*, denoted by **E**. **E** may be a coin or die throwing experiment, drawing of cards from a well-shuffled pack of cards, an agricultural experiment to determine the effects of fertilizers on yield of a commodity, and so on.

Each performance in a random experiment is called a *trial.* That is, all the trials conducted under the same set of conditions form a random experiment. The result of a trial in a random experiment is called an outcome, an elementary event or a sample point. The totality of all possible outcomes (*i.e.*, sample points) of a random experiment constitutes the *sample space.*

Suppose $e_1, e_2, ..., e_n$ are the possible outcomes of a random experiment E such that no two or more of them can occur simultaneously and exactly one of the outcomes $e_1, e_2, ..., e_n$ must occur. More specifically, with an experiment E, we associated a set $S = (e_1, e_2, ..., e_n)$ of possible outcomes with the following properties:

(i) each element of S denotes a possible outcome of the experiement,

(ii) any trial results in an outcome that corresponds to one and only one element of the set S.

The set S associated with an experiment E, real or conceptual, satisfying the above two properties is called the *sample space* of the experiment.

Remarks. 1. The sample space serves as universal set for all questions concerned with the experiment.

2. A sample space S is said to be finite (infinite) sample sapce if the number of elements in S is finite (infinite). For example, the sample space associated with the experiment of throwing the coin until a head appears, is infinite, with possible *sample points*

$$\{\omega_1, \omega_2, \omega_3, \omega_4, \ldots.\}$$

where $\omega_1 = H$, $\omega_2 = TH$, $\omega_3 = TTH$, $\omega_4 = TTTH$, and so on, H denoting a head and T a tail.

3. A sample space is called discrete if it contains only finitely or infinitely many points which can be arranged into a simple sequence $\omega_1, \omega_2, \ldots$, while a sample space containing non-denumerable number of points is called a continuous sample space. In this book, we shall restrict ourselves to discrete sample spaces only.

4·5·2. Event. Every non-empty subset A of S, which is a disjoint union of single element subsets of the sample space S of a random experiment E is called an event. The notion of an event may also be defined as follows:

"Of all the possible outcomes in the sample space of an experiment,some outcomes satisfy a specified description,which we call an event."

Remarks. 1. As the empty set ϕ is a subset of S, ϕ is also an event, known as *impossible* event.

2. An event A, in particular, can be a single element subset of S, in which case it is known as *elementary* event.

4·5·3. Some Illustrations — Examples. We discuss below some examples to illustrate the concepts of sample space and event.

1. Consider tossing of a coin singly. The possible outcomes for this experiment are (writing H for a "head" and T for a "tail") : H and T. Thus the sample space S consists of two points $\{\omega_1, \omega_2\}$, corresponding to each possible outcome or elementary event listed.

i.e., $S = \{\omega_1, \omega_2\} = \{H,T\}$ and $n(S) = 2$,

where $n(S)$ is the total number of sample points in S.

If we consider two tosses of a coin, the possible outcomes are HH, HT, TH, TT. Thus, in this case the sample space S consists of four points $\{\omega_1, \omega_2, \omega_3, \omega_4\}$, corresponding to each possible outcome listed and $n(S) = 4$. Combinations of these outcomes form what we call events. For example, the event of getting at lea t one head is the set of the outcomes $\{HH,HT,TH\} = \{\omega_1, \omega_2, \omega_3\}$. Thus, mathematically, the events are subsets of S.

2. Let us consider a single toss of a die. Since there are six possible outcomes, our sample space S is now a space of six points $\{\omega_1, \omega_2, \ldots, \omega_6\}$ where ω_i corresponds to the appearance of number i. Thus $S=\{\omega_1, \omega_2, \ldots, \omega_6\}=\{1,2,\ldots,6\}$ and $n(S)=6$. The event that the outcome is even is represented by the set of points $\{\omega_2, \omega_4, \omega_6\}$.

3. A coin and a die are tossed together. For this experiment, our sample space consists of twelve points $\{\omega_1, \omega_2, \ldots, \omega_{12}\}$ where ω_i $(i = 1, 2, \ldots, 6)$ represents a head on coin together with appearance of ith number on the die and ω_i $(i = 7, 8, \ldots, 12)$ represents a tail on coin together with the appearance of ith number on die. Thus

$$S = \{\omega_1, \omega_2, \ldots, \omega_{12}\} = \{(H, T) \times (1, 2, \ldots, 6)\} \text{ and } n(S) = 12$$

Remark. If the coin and die are unbiased, we can see intuitively that in each of the above examples, the outcomes (sample points) are equally likely to occur.

4. Consider an experiment in which two balls are drawn one by one from an urn containing 2 white and 4 blue balls such that when the second ball is drawn, the first is *not* replaced.

Let us number the six balls as 1, 2, 3, 4, 5 and 6, numbers 1 and 2 representing a white ball and numbers 3, 4, 5, and 6 representing a blue ball. Suppose in a draw we pick up balls numbered 2 and 6. Then (2,6) is called an outcome of the experiment. It should be noted that the outcome (2,6) is different from the outcome (6,2) because in the former case ball No. 2 is drawn first and ball No.6 is drawn next while in the latter case, 6th ball is drawn first and the second ball is drawn next.

The sample space consists of thirty points as listed below:

$\omega_1 = (1,2)$	$\omega_2 = (1,3)$	$\omega_3 = (1,4)$	$\omega_4 = (1,5)$	$\omega_5 = (1,6)$
$\omega_6 = (2,1)$	$\omega_7 = (2,3)$	$\omega_8 = (2,4)$	$\omega_9 = (2,5)$	$\omega_{10} = (2,6)$
$\omega_{11} = (3,1)$	$\omega_{12} = (3,2)$	$\omega_{13} = (3,4)$	$\omega_{14} = (3,5)$	$\omega_{15} = (3,6)$
$\omega_{16} = (4,1)$	$\omega_{17} = (4,2)$	$\omega_{18} = (4,3)$	$\omega_{19} = (4,5)$	$\omega_{20} = (4,6)$,
$\omega_{21} = (5,1)$	$\omega_{22} = (5,2)$	$\omega_{23} = (5,3)$	$\omega_{24} = (5,4)$	$\omega_{25} = (5,6)$
$\omega_{26} = (6,1)$	$\omega_{27} = (6,2)$	$\omega_{28} = (6,3)$	$\omega_{29} = (6,4)$	$\omega_{30} = (6,5)$

Thus

$$S = \{\omega_1, \omega_2, \omega_3, \ldots, \omega_{30}\} \text{ and } n(S) = 30$$

$$\Rightarrow \quad S = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\} - \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$$

The event

(i) the first ball drawn is white

(ii) the second ball drawn is white

(iii) both the balls drawn are white

(iv) both the balls drawn are black

are represented respectively by the following sets of points:

$\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8, \omega_9, \omega_{10}\}$,

$\{\omega_1, \omega_6, \omega_{11}, \omega_{12}, \omega_{16}, \omega_{17}, \omega_{21}, \omega_{22}, \omega_{26}, \omega_{27}\}$,

$\{\omega_1, \omega_6\}$, and

$\{\omega_{13}, \omega_{14}, \omega_{15}, \omega_{18}, \omega_{19}, \omega_{20}, \omega_{23}, \omega_{24}, \omega_{25}, \omega_{28}, \omega_{29}, \omega_{30}\}$.

5. Consider an experiment in which two dice are tossed. The sample space S for this experiment is given by

$$S = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$$

and $n(S) = 6 \times 6 = 36$.

Let E_1 be the event that 'the sum of the spots on the dice is greater than 12', E_2 be the event that 'the sum of spots on the dice is divisible by 3', and E_3 be the event that 'the sum is greater than or equal to two and is less than or equal to 12'. Then these events are represented by the following subsets of S :

$$E_1 = \{\phi\}, \; E_3 = S \text{ and}$$

$$E_2 = \{(1, 2), (1, 5), (2, 1), (2, 4), (3, 3), (3, 6), (4, 2), (4, 5), (5, 1), (5, 4), (6, 3), (6, 6)\}$$

Thus $n(E_1)=0$, $n(E_2)=12$, and $n(E_3)=36$

Here E is an 'impossible event' and E_3 a 'certain event'.

6. Let E denote the experiment of tossing a coin three times in succession or tossing three coins at a time. Then the sample space S is given by

$$\begin{aligned} S &= \{H, T\} \times \{H, T\} \times \{H, T\} \\ &= \{H, T\} \times \{HH, HT, TH, TT\} \\ &= \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} \\ &= \{\omega_1, \omega_2, \omega_3, \ldots, \omega_8\}, \text{ say.} \end{aligned}$$

If E_1 is the event that 'the number of heads exceeds the number of tails', E_2, the event of 'getting two heads' and E_3, the event of getting 'head in the first trial' then these are represented by the following sets of points :

$$E_1 = \{\omega_1, \omega_2, \omega_3, \omega_5\},$$

$$E_2 = \{\omega_2, \omega_3, \omega_5\}$$

and $$E_3 = \{\omega_1, \omega_2, \omega_3, \omega_4\}.$$

7. In the foregoing examples the sample sapce is finite. To construct an experiment in which the sample sapce is countably infinite, we toss a coin repeatedly until head or tail appears twice in succession. The sample space of all the possible outcomes may be represented as :

$$S = \{HH, TT, THH, HTT, HTHH, THTT, THTHH, HTHTT, \ldots\}.$$

4·5·4. Algebra of Events. For events A, B, C

(i) $A \cup B = \{\omega \in S : \omega \in A \text{ or } \omega \in B\}$

(ii) $A \cap B = \{\omega \in S : \omega \in A \text{ and } \omega \in B\}$

(iii) $\overline{A}$ (*A complement*) $= \{\omega \in S : \omega \notin A\}$

(iv) $A - B = \{\omega \in S : \omega \in A \text{ but } \omega \notin B\}$

(v) Similar generalisations for $\bigcup_{i=1}^{n} A_i$, $\bigcap_{i=1}^{n} A_i$, $\bigcup_{i} A_i$ etc.

(vi) $A \subset B \Rightarrow$ for every $\omega \in A$, $\omega \in B$.

(vii) $B \supset A \Rightarrow A \subset B$.

(viii) $A = B$ if and only if A and B have the same elements, *i.e.*, if $A \subset B$ and $B \subset A$.

(ix) A and B disjoint (mutually exclusive) $\Rightarrow A \cap B = \phi$ (*null set*).

(x) $A \cup B$ can be denoted by $A + B$ if A and B are disjoint.

(xi) $A \Delta B$ denotes those ω belonging to exactly one of A and B, *i.e.*,

$$A \Delta B = A \overline{B} \cup \overline{A} B$$

Remark. Since the events are subsets of S, all the laws of set theory *viz.*, commutative laws, associative laws, distributive laws, De-Morgan's law, etc., hold for algebra of events.

Table – *Glossary of Probability Terms*

Statement	*Meaning in terms of set theory*
1. At least one of the events A or B occurs.	$\omega \in A \cup B$
2. Both the events A and B occur.	$\omega \in A \cap B$
3. Neither A nor B occurs	$\omega \in \overline{A} \cap \overline{B}$
4. Event A occurs and B does not occur	$\omega \in A \cap \overline{B}$
5. Exactly one of the events A or B occurs.	$\omega \in A \Delta B$
6. Not more than one of the events A or B occurs.	$\omega \in (A \cap \overline{B}) \cup (\overline{A} \cap B) \cup (\overline{A} \cap \overline{B})$
7. If event A occurs, so does B	$A \subset B$
8. Events A and B are mutually exclusive.	$A \cap B = \phi$
9. Complementary event of A.	$\overline{A}$
10. Sample space	universal set S

Example 4·11. *A, B and C are three arbitrary events. Find expressions for the events noted below, in the context of A, B and C.*

(i) only A occurs,

(ii) Both A and B, but not C, occur,

(iii) All three events occur,

(iv) At least one occurs,

(v) At least two occur,

(vi) One and no more occurs,

(vii) Two and no more occur,

(viii) None occurs.

Solution.

(i) $A \cap \overline{B} \cap \overline{C}$, *(ii)* $A \cap B \cap \overline{C}$, *(iii)* $A \cap B \cap C$,

(iv) $A \cup B \cup C$,

(v) $(A \cap B \cap \overline{C}) \cup (A \cap \overline{B} \cap C) \cup (\overline{A} \cap B \cap C) \cup (A \cap B \cap C)$

(vi) $(A \cap \overline{B} \cap \overline{C}) \cup (\overline{A} \cap B \cap \overline{C}) \cup (\overline{A} \cap \overline{B} \cap C)$

(vii) $(A \cap B \cap \overline{C}) \cup (\overline{A} \cap B \cap C) \cup (A \cap \overline{B} \cap C)$

(viii) $\overline{A} \cap \overline{B} \cap \overline{C}$ *or* $\overline{A \cup B \cup C}$

EXERCISE 4(b)

1. *(i)* If A, B and C are any three events, write down the theoretical expressions for the following events:

(a) Only A occurs, *(b)* A and B occur but C does not,

(c) A, B, and C all the three occur, *(d)* at least one occurs

(e) At least two occur, *(f)* one does not occur,

(g) Two do not occurs, and *(h)* None occurs.

(ii) A, B and C are three events. Express the following events in appropriate symbols:

(a) Simultaneous occurrence of A, B and C.

(b) C currence of at least one of them.

(c) A, B and C are mutually exclusive events.

(d) Every point of A is contained in B.

(e) The event B but not A occurs. **[Gauhati Univ. B.Sc., Oct.1990]**

2. A sample space S contains four points x_1, x_2, x_3 and x_4 and the values of a set function $P(A)$ are known for the following sets :

$A_1 = (x_1, x_2)$ and $P(A_1) = \frac{4}{10}$; $A_2 = (x_3, x_4)$ and $P(A_2) = \frac{6}{10}$;

$A_3 = (x_1, x_2, x_3)$ and $P(A_3) = \frac{4}{10}$; $A_4 = (x_2, x_3, x_4)$ and $P(A_4) = \frac{7}{10}$

Show that :

(i) the total number of sets (including the "null" set of number points) of points of x is 16.

(ii) Although the set containing no sample point has zero probability, the converse is not always true, *i.e.*, a set may have zero probability and yet it may be the set of a number of points.

3. Describe explicitly the sample spaces for each of the following experiments:

(i) The tossing of four coins.

(ii) The throwing of three dice.

(iii) The tossing of ten coins with the aim of observing the numbers of tails coming up.

(iv) Two cards are selected from a standard deck of cards.

(v) Four successive draws *(a)* with replacement, and *(b)* without replacement, from a bag containing fifty coloured balls out of which ten are white, twenty blue and twenty red.

(vi) A survey of families with two children is conduced and the sex of the children (the older child first) is recorded.

(vii) A survey of families with three children is made and the sex of the children (in order of age, oldest child first) are recorded.

(viii) Three distinguishable objects are distributed in three numbered cells.

(ix) A poker hand (five cards) is dealt from an ordinary deck of cards.

(x) Selecting r screws from the lot produced by a machine, a screw can be defective or non-defective.

4. In an experiment a coin is thrown five times. Write down the sample space. How many points are there in the sample space?

5. Describe sample space appropriate in each of the following cases :

(i) n-tosses of a coin with head or tails as outcome in each toss.

(ii) Successive tosses of a coin until a head turns up.

(iii) A survey of families with two children is conducted and the sex of the children (the older child first) is recorded.

(iv) Two successive draws, *(a)* with replacement *(b)* without replacement, from a bag containing 4 coloured toys out of which one is white, one black and 2 red toys. **[M.S.Baroda Univ. B.Sc., 1991]**

6. *(a)* An experiment consists of tossing an unbiased coin until the same result appears twice on succession for the first time. To every possible outcome requiring n tosses attribute probability $1/2^n$. Describe the sample space.

(b) A coin is tossed until there are either two consecutive heads or two consecutive tails or the number of tosses becomes five. Describe the sample space along with the probability associated with each sample point, if every sequence of n tosses has probabilty 2^{-n}. **[Civil Services (main), 1983]**

7. Urn 1 contains two white, one red and 3 black balls. Urn 2 contains one white, 3 red and 2 black balls. An experiment consists of first selecting an urn and then drawing a ball from this urn. Define a suitable sample space for this experiment.

8. Suppose an experiment has n outcomes $A_1, A_2, ..., A_n$ and that it is repeated r times. Let $x_1, x_2, ..., x_n$ record the number of occurrences of $A_1, A_2, ..., A_n$. Describe the sample space. Show that the number of sample points is

$$\binom{n+r-1}{r-1}$$

9. A manufacturer buys parts from four different vendors numbered 1, 2, 3 and 4. Referring to orders placed on two successive days, (1,4) denotes the event that on the first day, the order was given to vendor 1 and on the second day it was given to vendor 4. Letting A represent the event that vendor 1 gets at least one of these two orders, B the event that the same vendor gets both orders and C the event that vendors 1 and 3 do not get either order. List the elements of :

(a) entire sample space, *(b)* A, *(c)* B, *(d)* C, *(e)* $\overline{A}$, *(f)* $\overline{B}$, *(g)* $B \cup C$, *(h)* $A \cap B$, *(i)* $A \cap C$, *(j)* $\overline{A \cup B}$, and *(k)* $A - B$

[Hint. *(a)* The elements of entire sample space are

(1,1); (1, 2); (1, 3); (1, 4); (2, 1); (2, 2); (2, 3); (2, 4);
(3, 1); (3, 2); (3, 3); (3, 4); (4, 1); (4, 2); (4, 3); (4, 4).

(b) The elements of A are
(1, 1); (1, 2); (1, 3); (1, 4); (2, 1); (3, 1); (4, 1).

(c) The elements of B are (1, 1); (2, 2); (3, 3) and (4, 4).

(d) The elements of C are (2, 2); (2, 4); (4, 2); (4, 4).

(e) The element of $\bar{A}$ are :
(2, 2); (2, 3); (2, 4); (3, 2); (3, 3); (3, 4); (4, 2); (4, 3); (4, 4).

(f) The elements of $\bar{B}$ are : (1, 2); (1, 3); (1, 4); (2, 1); (2, 3); (2, 4); (3, 1); (3, 2); (3, 4); (4, 1); (4, 2); (4, 3).

(g) The elements of B $\cup$ C are (1, 1); (2, 2); (3, 3); (4, 4); (2, 4); (4, 2).

(h) The elements of A $\cap$ B are (1, 1).

(i) $A \cap C = \phi$

(j) Since $\overline{A \cup B} = \bar{A} \cap \bar{B}$. The elements of $\overline{A \cup B}$ are (2, 3); (2, 4); (3, 2); (3, 4); (4, 2); (4, 3).

(k) The elements of $A - B$ are (1, 2); (1, 3); (1, 4); (2, 1); (3, 1); (4, 1).

4·6. Probability — Mathematical Notion. We are now set to give the mathematical notion of the occurrence of a random phenomenon and the mathematical notion of probability. Suppose in a large number of trials the sample space S contains N sample points. The event A is defined by a description which is satisfied by N_A of the occurrences. The frequency interpretation of the probability $P(A)$ of the event A, tells us that $P(A)=N_A/N$.

A purely mathematical definition of probability cannot give us the actual value of $P(A)$ and this must be considered as a function defined on all events. With this in view, a mathematical definition of probability is enunciated as follows:

"Given a sample description space, probability is a function which assigns a non-negative real number to every event A, denoted by P(A) and is called the probability of the event A."

4·6·1. Probability Function. $P(A)$ is the probability function defined on a σ-field B of events if the following properties or axioms hold :

1. For each $A \in$ B, $P(A)$ is defined, is real and $P(A) \geq 0$
2. $P(S) = 1$
3. If $\{A_n\}$ is any finite or infinite sequence of disjoint events in B, then

$$P\left(\bigcup_{i=1}^{n} A_n\right) = \sum_{i=1}^{n} P(A_i) \qquad \ldots(4\cdot4)$$

The above three axioms are termed as the axiom of positiveness, certainty and union (additivity), respectively.

Remarks. 1. The set function P defined on σ-field B, taking its values in the real line and satisfying the above three axioms is called the probability measure.

2. The same definition of probability applies to *uncountable sample space* except that special restrictions must be placed on S and its subsets. It is important to realise that for a complete description of a probability measure, three things must

be specified, *viz.*, the sample space S, the σ-field (σ-algebra) B formed from certain subset of S and set function P. The triplet (S, B, P) is often called the *probability space*. In most elementary applications, S is finite and the σ-algebra B is taken to be the collection of all subsets of S.

3. It is interesting to see that there are some formal statements of the properties of events derived from the frequency approach. Since $P(A)=N_A/N$, it is easy to see that $P(A) \geq 0$, as in Axiom 1. Next since $N_S = N$, $P(S)=1$, as in Axiom 2. In case of two mutually exclusive (or disjoint) events A and B defined by sample points N_A and N_B, the sample points belonging to $A \cup B$ are $N_A + N_B$. Therefore,

$$P(A \cup B) = \frac{N_A + N_B}{N} = \frac{N_A}{N} + \frac{N_B}{N} = P(A) + P(B), \text{ as in axiom 3.}$$

Extended Axiom of Addition. If an event A can materialise in the occurrence of any one of the pairwise disjoint events $A_1, A_2, \ldots$ so that

$$A = \bigcup_{i=1}^{\infty} A_i ; \; A_i \cap A_j = \phi \;\; (i \neq j)$$

then

$$P(A) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) \qquad \ldots(1)$$

Axiom of Continuity. If $B_1, B_2, \ldots, B_n, \ldots$ be a countable sequences of events such that

$$(i)\; B_i \supset B_{i+1}, \; (i = 1, 2, 3, \ldots)$$

and

$$(ii) \bigcap_{n=1}^{\infty} B_n = \phi$$

i.e., if each succeeding event implies the preceeding event and if their simultaneous occurrence is an impossible event then

$$\lim_{n \to \infty} P(B_n) = 0 \qquad \ldots(2)$$

We shall now prove that these two axioms, *viz.*, the extended axiom of addition and axiom of continuity are equivalent, *i.e.*, each implies the other, *i.e.*, (1) ⇔ (2).

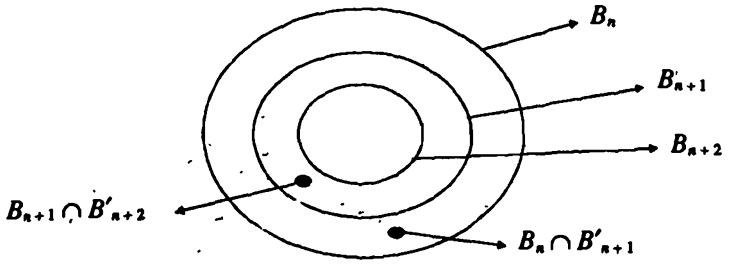
Theorem 4·1. *Axiom of continuity follows from the extended axiom of addition and vice versa.*

Proof. *(a)* (1) ⇒(2). Let $\{B_n\}$ be a countable sequence of events such that

$$B_1 \supset B_2 \supset B_3 \supset \ldots, \supset B_n \supset B_{n+1} \supset \ldots$$

and let for any $n \geq 1$,

$$\bigcap_{k \geq n} B_k = \phi \qquad (*)$$



Then it is obvious from the diagram that

$$B_n = B_n B'_{n+1} \cup B_{n+1} B'_{n+2} \cup \ldots \cup \left(\bigcap_{k \geq n} B_k \right)$$

$$\Rightarrow \quad B_n = \left(\bigcup_{k=n}^{\infty} B_k B'_{k+1} \right) \cup \left(\bigcap_{k \geq n} B_k \right),$$

where the events $B_k B'_{k+1}$; $(k=n, n+1, \ldots)$ are pairwise disjoint and each is disjoint with $\bigcap_{k \geq n} B_k$.

Thus B_n has been expressed as the countable union of pairwise disjoint events and hence by the extended axiom of addition, we get

$$P(B_n) = \sum_{k=n}^{\infty} P(B_k B'_{k+1}) + P\left(\bigcap_{k \geq n} B_k \right)$$

$$= \sum_{k=n}^{\infty} P(B_k B'_{k+1}), \qquad (**)$$

since, from (*)

$$P\left(\bigcap_{k \geq n} B_k \right) = P(\phi) = 0$$

Further, from (**), since

$$\sum_{k=1}^{\infty} P(B_k B'_{k+1}) = P(B_1) \leq 1,$$

the right hand sum in (**), being the remainder after n terms of a convergent series tends to zero as $n \to \infty$.

Hence

$$\lim_{n \to \infty} P(B_n) = \lim_{n \to \infty} \sum_{k=n}^{\infty} P(B_k B'_{k+1}) = 0$$

Thus the extended axiom of addition implies the axiom of continuity.

(b) Conversely (2) $\Rightarrow$ (1), *i.e., the extended axiom of addition follows from the axiom of continuity.*

Let $\{A_n\}$ be a countable sequence of pairwise disjoint events and let

$$A = \bigcup_{i=1}^{\infty} A_i$$

$$= \left(\bigcup_{i=1}^{n} A_i\right) \cup \left(\bigcup_{i=n+1}^{\infty} A_i\right) \quad ...(3)$$

Let us define a countable sequence $\{B_n\}$ of events by

$$B_n = \bigcup_{i=n}^{\infty} A_i \quad ...(4)$$

Obviously B_n is a decreasing sequence of events, *i.e.*,

$$B_1 \supset B_2 \supset \ldots \supset B_n \supset B_{n+1} \supset \ldots. \quad ...(5)$$

Also we have

$$A = \left(\bigcup_{i=1}^{n} A_i\right) \cup B_{n+1} \quad ...(6)$$

Since A_i's are pairwise disjoint, we get

$$A_i \cap B_{n+1} = \phi, \ (i = 1, 2, \ldots, n) \quad ...(6a)$$

From (4) we see that if the event B_n has occurred it implies the occurrence of any one of the events $A_{n+1}, A_{n+2}, \ldots$ Without loss of generality let us assume that this event is A_i $(i = n+1, n+2, \ldots)$. Further since A_i's are pairwise disjoint, the occurrence of A_i implies that events $A_{i+1}, A_{i+2}, \ldots$ do not occur leading to the conclusion that $B_{i+1}, B_{i+2}, \ldots$ will not occur.

$$\Rightarrow \quad \bigcap_{i=n}^{\infty} B_i = \phi \quad ...(7)$$

From (5) and (7), we observe that both the conditions of axiom of continuity are satisfied and hence we get

$$\lim_{n \to \infty} P(B_n) = 0 \quad ...(8)$$

From (6), we get

$$P(A) = P\left[\left(\bigcup_{i=1}^{n} A_i\right) \cup B_{n+1}\right]$$

$$= \sum_{i=1}^{n} P(A_i) + P(B_{n+1})$$

(By axiom of Additivity)

$$\Rightarrow \quad P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n \to \infty} \sum_{i=1}^{n} P(A_i) + \lim_{n \to \infty} (B_{n+1})$$

$$= \sum_{i=1}^{\infty} P(A_i), \qquad \text{[From (8)]}$$

which is the extended axiom of addition.

THEOREMS ON PROBABILITIES OF EVENTS

Theorem 4·2. *Probability of the impossible event is zero, i.e.,* $P(\phi) = 0$.

Proof. Impossible event contains no sample point and hence the certain event S and the impossible event ϕ are mutually exclusive.

Hence $S \cup \phi = S$

$\therefore$ $P(S \cup \phi) = P(S)$

$\Rightarrow$ $P(S) + P(\phi) = P(S)$ [By Axiom 3]

$\Rightarrow$ $P(\phi) = 0$

Remark. It may be noted $P(A)=0$, does not imply that A is necessarily an empty set. In practice, probability '0' is assigned to the events which are so rare that they happen only once in a lifetime. For example, if a person who does not know typing is asked to type the manuscript of a book, the probability of the event that he will type it correctly without any mistake is 0.

As another illustration, let us consider the random tossing of a coin. The event that the coin will stand erect on its edge, is assigned the probability 0.

The study of continuous random variable provides another illustration to the fact that $P(A)=0$, does not imply $A=\phi$, because in case of continous random variable X, the proability at a point is always zero, *i.e.*, $P(X=c)=0$ [See Chapter 5].

Theorem 4·3. Probability of the complementary event $\bar{A}$ of A is given by

$$P(\bar{A}) = 1 - P(A)$$

Proof. A and $\bar{A}$ are disjoint events.

Moreover, $A \cup \bar{A} = S$

From axioms 2 and 3 of probability, we have

$$P(A \cup \bar{A}) = P(A) + P(\bar{A}) = P(S) = 1$$

$\Rightarrow$ $P(\bar{A}) = 1 - P(A)$

Cor. 1. We have $P(A) = 1 - P(\bar{A})$

$\Rightarrow$ $P(A) \leq 1$ $(\because P(\bar{A}) \geq 0)$

Cor. 2. $P(\phi) = 0$, since $\phi = \bar{S}$

and $P(\phi) = P(\bar{S}) = 1 - P(S) = 1 - 1 = 0.$

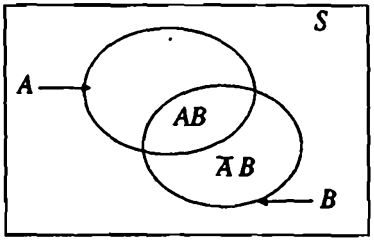
Theorem 4·4. *For any two events A and B,*

$$P(\bar{A} \cap B) = P(B) - P(A \cap B)$$ **[Mysore Univ. B.Sc., 1992]**

Proof.

$\bar{A} \cap B$ and $A \cap B$ are disjoint events and

$(A \cap B) \cup (\bar{A} \cap B) = B$



Hence by axiom 3, we get

$$P(B) = P(A \cap B) + P(\bar{A} \cap B)$$

$$\Rightarrow P(\bar{A} \cap B) = P(B) - P(A \cap B)$$

Remark. Similarly, we shall get $P(A \cap \bar{B}) = P(A) - P(A \cap B)$

Theorem 4·5. *Probability of the union of any two events A and B is given by*

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

Proof. $A \cup B$ can be written as the union of the two mutually disjoint events, A and $B \cap \bar{A}$.

$$\therefore \quad P(A \cup B) = P[A \cup (B \cap \bar{A})] = P(A) + P(B \cap \bar{A})$$

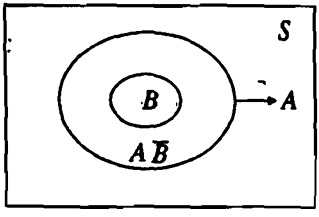
$$= P(A) + P(B) - P(A \cap B) \qquad (c.f.\ \text{Theorem 4·4})$$

Theorem 4·6. *If $B \subset A$, then*

(i) $P(A \cap \bar{B}) = P(A) - P(B)$,

(ii) $P(B) \leq P(A)$

Proof. *(i)* When $B \subset A$, B and $A \cap \bar{B}$ are mutually exclusive events and their union is A



Therefore

$$P(A) = P[B \cup (A \cap \bar{B})]$$

$$= P(B) + P(A \cap \bar{B}) \qquad \text{[By axiom 3]}$$

$$\Rightarrow \quad P(A \cap \bar{B}) = P(A) - P(B)$$

(ii) Using axiom 1,

$$P(A \cap \bar{B}) \geq 0 \quad \Rightarrow \quad P(A) - P(B) \geq 0$$

Hence $P(B) \leq P(A)$

Cor. Since $(A \cap B) \subset A$ and $(A \cap B) \subset B$,

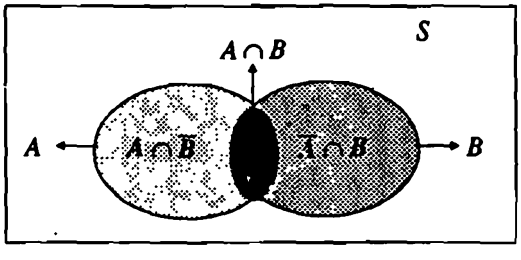
$P(A \cap B) \leq P(A)$ and $P(A \cap B) \leq P(B)$

4·6·2. Law of Addition of Probabilities

Statement. *If A and B are any two events [subsets of sample space S] and are not disjoint, then*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad \ldots(4·5)$$

Proof.



We have

$$A \cup B = A \cup (\bar{A} \cap B)$$

Since A and $(\bar{A} \cap B)$ are disjoint,

$$P(A \cup B) = P(A) + P(\bar{A} \cap B)$$
$$= P(A) + [P(\bar{A} \cap B) + P(A \cap B)] - P(A \cap B)$$
$$= P(A) + P[(\bar{A} \cap B) \cup (A \cap B)] - P(A \cap B)$$

$[\because (\bar{A} \cap B)$ and $(A \cap B)$ *are disjoint*$]$

$$\Rightarrow \quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Remark. An alternative proof is provided by Theorems 4·4 *and* 4·5.

4·6·3. Extention of General Law of Addition of Probabilities. *For n events* $A_1, A_2, \ldots, A_n$ *we have*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum\sum_{1 \le i < j \le n} P(A_i \cap A_j) + \sum\sum\sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k)$$
$$- \ldots + (-1)^{n-1} P(A_1 \cap A_2 \cap \ldots \cap A_n) \qquad \ldots(4\cdot6)$$

Proof. For two events A_1 and A_2, we have

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \qquad \ldots(*)$$

Hence (4·6) is true for $n = 2$.

Let us now suppose that (4·6) is true for $n = r$, (say). Then

$$P\left(\bigcup_{i=1}^{r} A_i\right) = \sum_{i=1}^{r} P(A_i) - \sum\sum_{1 \le i < j \le r} P(A_i \cap A_j) + \ldots + (-1)^{r-1} P(A_1 \cap A_2 \cap \ldots \cap A_r) \qquad \ldots(**)$$

Now

$$P\left(\bigcup_{i=1}^{r+1} A_i\right) = P\left[\left(\bigcup_{i=1}^{r} A_i\right) \cup A_{r+1}\right]$$

$$= P\left(\bigcup_{i=1}^{r} A_i\right) + P(A_{r+1}) - P\left[\left(\bigcup_{i=1}^{r} A_i\right) \cap A_{r+1}\right] \qquad \ldots[\text{Using } (*)]$$

$$= P\left(\bigcup_{i=1}^{r} A_i\right) + P(A_{r+1}) - P\left[\bigcup_{i=1}^{r} (A_i \cap A_{r+1})\right] \qquad \text{(Distributive Law)}$$

$$= \sum_{i=1}^{r} P(A_i) - \sum\sum_{1 \le i < j \le r} P(A_i \cap A_j) + \ldots$$
$$\ldots + (-1)^{r-1} P(A_1 \cap A_2 \cap \ldots \cap A_r) + P(A_{r+1})$$
$$- P\left[\bigcup_{i=1}^{r} (A_i \cap A_{r+1})\right] \qquad \ldots[\text{From } (**)]$$

$$= \sum_{i=1}^{r+1} P(A_i) - \sum\sum_{1 \le i < j \le r} P(A_i \cap A_j) + \ldots$$
$$+ (-1)^{r-1} P(A_1 \cap A_2 \cap \ldots \cap A_r)$$

$$- [\sum_{i=1}^{r} P(A_i \cap A_{r+1}) - \sum\sum_{1 \le i < j \le r} P(A_i \cap A_j \cap A_{r+1})$$

$$+ \ldots + (-1)^{r-1} P(A_1 \cap A_2 \cap \ldots \cap A_r \cap A_{r+1})] \quad \ldots[\text{From } (**)]$$

$$\Rightarrow \quad P(\bigcup_{i=1}^{r+1} A_i) = \sum_{i=1}^{r+1} P(A_i) - [\sum\sum_{1 \le i < j \le r} P(A_i \cap A_j) + \sum_{i=1}^{r} P(A_i \cap A_{r+1})]$$

$$+ \ldots + (-1)^r P[(A_1 \cap A_2 \cap \ldots \cap A_{r+1})]$$

$$= \sum_{i=1}^{r+1} P(A_i) - \sum\sum_{1 \le i < j \le (r+1)} P(A_i \cap A_j)$$

$$+ \ldots + (-1)^r P(A_1 \cap A_2 \cap \ldots \cap A_{r+1})$$

Hence if (4·6) is true for $n=r$, it is also true for $n = (r+1)$. But we have proved in (*) that (4·6) is true for $n=2$. Hence by the principle of mathematical induction, it follows that (4·6) is true for all positive integral values of n.

Remarks. 1. If we write

$$P(A_i) = p_i, P(A_i \cap A_j) = p_{ij}, P(A_i \cap A_j \cap A_k) = p_{ijk}$$

and so on and

$$S_1 = \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} P(A_i)$$

$$S_2 = \sum\sum_{1 \le i < j \le n} p_{ij} = \sum\sum_{1 \le i < j \le n} P(A_i \cap A_j)$$

$$S_3 = \sum\sum\sum_{1 \le i < j < k \le n} p_{ijk} \quad \text{and so on,}$$

then

$$P(\bigcup_{i=1}^{n} A_i) = S_1 - S_2 + S_3 - \ldots + (-1)^{n-1} S_n \qquad \ldots(4\cdot6a)$$

2. If all the events A_i, $(i = 1, 2, \ldots, n)$ are mutually disjoint then (4·6) gives

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

3. From practical point of view the theorem can be restated in a slightly different form. Let us suppose that an event A can materialise in several mutually exclusive forms, *viz.*, $A_1, A_2, \ldots, A_n$ which may be regarded as that many mutually exclusive events. If A happens then any one of the events A_i, $(i = 1, 2, \ldots, n)$ must happen and conversely if any one of the events A_i, $(i = 1, 2, \ldots, n)$ happens, then A happens. Hence the probability of happening of A is the same as the probability of happening of any one of its (unspecified) mutually exclusive forms. From this point of view, the total probability theorem can be restated as follows:

The probability of happening of an event A is the sum of the probabilities of happening of its mutually exclusive forms $A_1, A_2, \ldots, A_n$. *Symbolically,*

$$P(A) = P(A_1) + P(A_2) + \ldots + P(A_n) \qquad (4\cdot6b)$$

The probabilities $P(A_1), P(A_2), ..., P(A_n)$ of the mutually exclusive forms of A are known as the *partial probabilities*. Since $P(A)$ is their sum, it may be called the *total probability* of A. Hence the name of the theorem.

Theorem 4·7. *(Boole's inequality). For n events $A_1, A_2, ..., A_n$, we have*

$$(a) \quad P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1) \qquad ...(4·7)$$

$$(b) \quad P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i) \qquad ...(4·7a)$$

[Delhi Univ. B.Sc. (Stat Hons.), 1992, 1989]

Proof. *(a)* $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq 1$

$$\Rightarrow \quad P(A_1 \cap A_2) \geq P(A_1) + P(A_2) - 1 \qquad (*)$$

Hence (4·7) is true for $n = 2$.

Let us now suppose that (4·7) is true for $n=r$ (say), such that

$$P\left(\bigcap_{i=1}^{r} A_i\right) \geq \sum_{i=1}^{r} P(A_i) - (r-1) \qquad (**)$$

Then

$$P\left(\bigcap_{i=1}^{r+1} A_i\right) = P\left(\bigcap_{i=1}^{r} A_i \cap A_{r+1}\right)$$

$$\geq P\left(\bigcap_{i=1}^{r} A_i\right) + P(A_{r+1}) - 1 \qquad \text{[From (*)]}$$

$$\geq \sum_{i=1}^{r} P(A_i) - (r-1) + P(A_{r+1}) - 1 \qquad \text{[From (**)]}$$

$$\Rightarrow \quad P\left(\bigcap_{i=1}^{r+1} A_i\right) \geq \sum_{i=1}^{r+1} P(A_i) - r$$

$\Rightarrow$ (4·7) is true for $n = r + 1$ also.

The result now follows by the principle of mathematical induction.

(b) Applying the inequality (4·7) to the events $\overline{A}_1, \overline{A}_2, ..., \overline{A}_n$, we get

$$P(\overline{A}_1 \cap \overline{A}_2 \cap ... \cap \overline{A}_n) \geq [P(\overline{A}_1) + P(\overline{A}_2) + ... + P(\overline{A}_n)] - (n-1)$$

$$= [1 - P(A_1)] + [1 - P(A_2)] + ... + [1 - P(A_n)] - (n-1)$$

$$= 1 - P(A_1) - P(A_2) - ... - P(A_n)$$

$$\Rightarrow \quad P(A_1) + P(A_2) + ... + P(A_n) \geq 1 - P(\overline{A}_1) \cap \overline{A}_2 \cap ... \cap \overline{A}_n).$$

$$= 1 - P(\overline{A}_1 \cup A_2 \cup ... \cup \overline{A}_n).$$

$$= P(A_1 \cup A_2 \cup ... \cup A_n)$$

$$\Rightarrow \quad P(A_1 \cup A_2 \cup ... \cup A_n) \leq P(A_1) + P(A_2) + ... + P(A_n)$$

as desired.

Aliter for (b) i.e., (4·7a). We have

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

$$\leq P(A_1) + P(A_2) \qquad [\because P(A_1 \cap A_2) \geq 0] \qquad ...(***)$$

Hence (4·7*a*) is true for $n = 2$.

Let us now suppose that (4·7*a*) is true for $n=r$, (say), so that

$$P\left(\bigcup_{i=1}^{r} A_i\right) \leq \sum_{i=1}^{r} P(A_i) \qquad ...(****)$$

Now

$$P\left(\bigcup_{i=1}^{r+1} A_i\right) = P\left(\bigcup_{i=1}^{r} A_i \cup A_{r+1}\right)$$

$$\leq P\left(\bigcup_{i=1}^{r} A_i\right) + P(A_{r+1}) \qquad \text{[Using (***)]}$$

$$\leq \sum_{i=1}^{r} P(A_i) + P(A_{r+1}) \qquad \text{[Using (****)]}$$

$$\Rightarrow \quad P\left(\bigcup_{i=1}^{r+1} A_i\right) \leq \sum_{i=1}^{r+1} P(A_i)$$

Hence if (4·7*a*) is true for $n=r$, then it is also true for $n=r+1$. But we have proved in (***) that (4·7*a*) is true for $n=2$. Hence by mathematical induction we conclude that (4·7*a*) is true for all positive integral values of n.

Theorem 4·8. *For n events* $A_1, A_2, ..., A_n$,

$$P\left[\bigcup_{i=1}^{n} A_i\right] \geq \sum_{i=1}^{n} P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j)$$

[Delhi Univ. B.Sc. (Stat Hons.), 1986]

Proof. We shall prove this theorem by the method of induction.

We know that

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - [P(A_1 \cap A_2) + P(A_2 \cap A_3) + P(A_3 \cap A_1)] + P(A_1 \cap A_2 \cap A_3)$$

$$\Rightarrow P\left(\bigcup_{i=1}^{3} A_i\right) \geq \sum_{i=1}^{3} P(A_i) - \sum\sum_{1 \leq i < j \leq 3} P(A_i \cap A_j)$$

Thus the result is true for $n=3$. Let us now suppose that the result is true for $n=r$ (say), so that

$$P\left(\bigcup_{i=1}^{r} A_i\right) \geq \sum_{i=1}^{r} P(A_i) - \sum\sum_{1 \leq i < j \leq r} P(A_i \cap A_j) \qquad ...(*)$$

Now

$$P\left(\bigcup_{i=1}^{r+1} A_i\right) = P\left(\bigcup_{i=1}^{r} A_i \cup A_{r+1}\right)$$

$$= P\left(\bigcup_{i=1}^{r} A_i\right) + P(A_{r+1}) - P\left[\left(\bigcup_{i=1}^{r} A_i\right) \cap A_{r+1}\right]$$

$$= P\left(\bigcup_{i=1}^{r} A_i\right) + P(A_{r+1}) - P\left[\bigcup_{i=1}^{r} (A_i \cap A_{r+1})\right]$$

$$\geq \left[\sum_{i=1}^{r} P(A_i) - \sum\sum_{1 \leq i < j < r} P(A_i \cap A_j)\right]$$

$$+ P(A_{r+1}) - P\left[\bigcup_{i=1}^{r} (A_i \cap A_{r+1})\right] \quad \ldots(**)$$

[From (*)]

From Boole's inequality (*c.f.* Theorem 4·7 page 4·33), we get

$$P\left[\bigcup_{i=1}^{r} (A_i \cap A_{r+1})\right] \leq \sum_{i=1}^{r} P(A_i \cap A_{r+1})$$

$$\Rightarrow \quad -P\left[\bigcup_{i=1}^{r} (A_i \cap A_{r+1})\right] \geq -\sum_{i=1}^{r} P(A_i \cap A_{r+1})$$

$\therefore$ From (**), we get

$$P\left(\bigcup_{i=1}^{r+1} A_i\right) \geq \sum_{i=1}^{r+1} P(A_i) - \sum\sum_{1 \leq i < j \leq r} P(A_i \cap A_j) - \sum_{i=1}^{r} P(A_i \cap A_{r+1})$$

$$\Rightarrow \quad P\left(\bigcup_{i=1}^{r+1} A_i\right) \geq \sum_{i=1}^{r+1} P(A_i) - \sum\sum_{1 \leq i < j \leq r+1} P(A_i \cap A_j)$$

Hence, if the theorem is true for $n = r$, it is also true for $n = r+1$. But we have seen that the result is true for $n = 3$. Hence by mathematical induction, the result is true for all positive integral values of n.

4·7. Multiplication Law of Probability and Conditional Probability

Theorem 4·8. *For two events A and B*

$$\left.\begin{aligned} P(A \cap B) &= P(A) \,.\, P(B \mid A), P(A) > 0 \\ &= P(B) \,.\, P(A \mid B), \; P(B) > 0 \end{aligned}\right\} \quad \ldots(4\cdot8)$$

where $P(B \mid A)$ represents the conditional probability of occurrence of B when the event A has already happened and $P(A \mid B)$ is the conditional probability of happening of A, given that B has already happened.

Proof.

$$P(A) = \frac{n(A)}{n(S)} \; ; \; P(B) = \frac{n(B)}{n(S)} \text{ and } P(A \cap B) = \frac{n(A \cap B)}{n(S)} \quad (*)$$

For the conditional event $A \mid B$, the favourable outcomes must be one of the sample points of B, *i.e.*, for the event $A \mid B$, the sample space is B and out of the $n(B)$ sample points, $n(A \cap B)$ pertain to the occurrence of the event A. Hence

$$P(A \mid B) = \frac{n(A \cap B)}{n(B)}$$

Rewriting (*), we get

$$P(A \cap B) = \frac{n(B)}{n(S)} \quad \frac{n(A \cap B)}{n(B)} = P(B) \;.\; P(A \mid B)$$

Similarly we can prove :

$$P(A \cap B) = \frac{n(A)}{n(S)} \cdot \frac{n(A \cap B)}{n(A)} = P(A) \cdot P(B \mid A)$$

Remarks. 1. $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)}$ and $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$

Thus the conditional probabilities $P(B \mid A)$ and $P(A \mid B)$ are defined if and only if $P(A) \neq 0$ and $P(B) \neq 0$, respectively.

2. *(i)* For $P(B) > 0$, $P(A \mid B) \leq P(A)$

(ii) The conditional probability $P(A \mid B)$ is not defined if $P(B) = 0$.

(iii) $P(B \mid B) = 1$.

3. Multiplication Law of Probability for Independent Events. If A and B are independent then

$$P(A \mid B) = P(A) \quad \text{and} \quad P(B \mid A) = P(B)$$

Hence (4·8) gives :

$$P(A \cap B) = P(A)\, P(B) \qquad \ldots(4\cdot 8a)$$

provided A and B are independent.

4·7·1. Extension of Multiplication Law of Probability. *For n events $A_1, A_2, \ldots, A_n$, we have*

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1 \cap A_2) \ldots \times P(A_n \mid A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \qquad \ldots(4\cdot 8b)$$

where $P(A_i \mid A_j \cap A_k \cap \ldots \cap A_l)$ represents the conditional probability of the event A_i given that the events $A_j, A_k, \ldots, A_l$ have already happened.

Proof. We have for three events A_1, A_2, and A_3

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P[A_1 \cap (A_2 \cap A_3)] \\ &= P(A_1)\, P(A_2 \cap A_3 \mid A_1) \\ &= P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1 \cap A_2) \end{aligned}$$

Thus we find that (4·8*b*) is true for $n=2$ and $n=3$. Let us suppose that (4·8*b*) is true for $n=k$, so that

$$P(A_1 \cap A_2 \cap \ldots \cap A_k) = P(A_1)\, P(A_2 \mid A_1)\, P(A_3 \mid A_1 \cap A_2) \ldots P(A_k \mid A_1 \cap A_2 \cap \ldots \cap A_{k-1})$$

Now

$$\begin{aligned} P[(A_1 \cap A_2 \cap \ldots \cap A_k) \cap A_{k+1}] &= P(A_1 \cap A_2 \cap \ldots \cap A_k) \times P(A_{k+1} \mid A_1 \cap A_2 \cap \ldots \cap A_k) \\ &= P(A_1)\, P(A_2 \mid A_1) \ldots P(A_k \mid A_1 \cap A_2 \cap \ldots \cap A_{k-1}) \times P(A_{k+1} \mid A_1 \cap A_2 \cap \ldots \cap A_k) \end{aligned}$$

Thus (4·8*b*) is true for $n=k+1$ also. Since (4·8*b*) is true for $n=2$ and $n=3$, by the principle of mathematical induction, it follows that (4·8*b*) is true for all positive integral values of n.

Remark. If $A_1, A_2, \ldots, A_n$ are independent events then

$$P(A_2 \mid A_1) = P(A_2),\ P(A_3 \mid A_1 \cap A_2) = P(A_3) \ldots P(A_n \mid A_1 \cap A_2 \cap \ldots \cap A_{n-1}) = P(A_n)$$

Hence (4·8*b*) gives :

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)\,P(A_2) \ldots \cap P(A_n), \qquad \ldots(4\cdot8c)$$

provided $A_1, A_2, \ldots, A_n$ are independent.

Remark. Mutually Exclusive (Disjoint) Events and Independent Events. Let A and B be mutually exclusive (disjoint) events with positive probabilities ($P(A) > 0, P(B) > 0$), *i.e.*, both A and B are possible events such that

$$A \cap B = \phi \quad \Rightarrow \quad P(A \cap B) = P(\phi) = 0 \qquad \ldots(i)$$

Further, by compound probability theorem we have

$$P(A \cap B) = P(A)\,.\,P(B \mid A) = P(B)\,.P(A \mid B) \qquad \ldots(ii)$$

Since $P(A) \neq 0$; $P(B) \neq 0$, from *(i)* and *(ii)* we get

$$P(A \mid B) = 0 \neq P(A), \quad P(B \mid A) = 0 \neq P(B) \qquad \ldots(iii)$$

$\Rightarrow$ A and B are dependent events.

Hence two possible mutually disjoint events are always dependent (not independent) events.

However, if A and B are independent events with $P(A) > 0$ and $P(B) > 0$, then

$$P(A \cap B) = P(A)\ P(B) \neq 0$$

$\Rightarrow$ A and B cannot be mutually exclusive.

Hence two independent events (both of which are possible events), cannot be mutually disjoint.

4·7·2. Given n independent events A_i, (i =1,2,...,n) with respective probabilities of occurrence p_i, to find the probability of occurrence of at least one of them.

We have

$$P(A_i) = p_i \quad \Rightarrow \quad P(\overline{A}_i) = 1 - p_i\,;\ i = 1, 2, \ldots, n$$

$$[\because \overline{(A_1 \cup A_2 \cup \ldots \cup A_n)} = (\overline{A}_1 \cap \overline{A}_2 \cap \ldots \cap \overline{A}_n) \quad (\textit{De–Morgan's Law})]$$

Hence the probability of happening of at least one of the events is given by

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = 1 - P(\overline{A_1 \cup A_2 \cup \ldots \cup A_n}) \qquad \ldots(*)$$

$$= 1 - P(\overline{A}_1 \cap \overline{A}_2 \cap \ldots \cap \overline{A}_n)$$

$$= 1 - P(\overline{A}_1)\,P(\overline{A}_2) \ldots P(\overline{A}_n) \qquad \ldots(**)$$

[*c.f.* Theorem 4·14 page 4·41]

$$= 1 - [(1 - p_1)(1 - p_2) \ldots (1 - p_n)]$$

$$= \left[\sum_{i=1}^{n} p_i - \sum_{\substack{i,j=1 \\ i<j}}^{n} (p_i p_j) + \sum_{\substack{i,j,k=1 \\ i<j<k}}^{n} (p_i p_j p_k) \right.$$

$$\left. \ldots + (-1)^{n-1}\ (p_1 p_2 \ldots p_n) \right]$$

Remark. The results in (*) and (**) are very important and are used quite often in numerical problems. Result (*) stated in words gives:

P [happening of at least one of the events $A_1, A_2, \ldots, A_n$]

$= 1 - P$ (none of the events $A_1, A_2, \ldots, A_n$ happens)

or equivalently,

P {none of the given events happens}
$= 1 - P$ {at least one of them happens}.

Theorem 4·9. *For any three events A, B and C*

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

Proof. We have

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

$$\Rightarrow P[(A \cap C) \cup (B \cap C)] = P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)$$

Dividing both sides by $P(C)$, we get

$$\frac{P[(A \cap C) \cup (B \cap C)]}{P(C)} = \frac{P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)}{P(C)}, P(C) > 0$$

$$= \frac{P(A \cap C)}{P(C)} + \frac{P(B \cap C)}{P(C)} - \frac{P(A \cap B \cap C)}{P(C)}$$

$$\Rightarrow \quad \frac{P[(A \cup B) \cap C]}{P(C)} = P(A \mid C) + P(B \mid C) - P(A \cap B \mid C)$$

$$\Rightarrow \quad P[(A \cup B) \mid C] = P(A \mid C) + P(B \mid C) - P(A \cap B \mid C)$$

Theorem 4·10. *For any three events A, B and C*

$$P(A \cap \bar{B} \mid C) + P(A \cap B \mid C) = P(A \mid C)$$

Proof.

$$P(A \cap \bar{B} \mid C) + P(A \cap B \mid C) = \frac{P(A \cap \bar{B} \cap C)}{P(C)} + \frac{P(A \cap B \cap C)}{P(C)}$$

$$= \frac{P(A \cap \bar{B} \cap C) + P(A \cap B \cap C)}{P(C)}$$

$$= \frac{P(A \cap C)}{P(C)} = P(A \mid C)$$

Theorem 4·11. *For a fixed B with $P(B) > 0$, $P(A \mid B)$ is a probability function.* **[Delhi Univ. B.Sc. (Stat. Hons.), 1991; (Maths Hons.), 1992]**

Proof.

(i) $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} \geq 0$

(ii) $P(S \mid B) = \dfrac{P(S \cap B)}{P(B)} = \dfrac{P(B)}{P(B)} = 1$

(iii) If $\{A_n\}$ is any finite or infinite sequences of disjoint events, then

$$P\left[\bigcup_n A_n \mid B\right] = \frac{P[(\bigcup_n A_n) \cap B]}{P(B)} = \frac{P[(\bigcup_n A_n . B)]}{P(B)}$$

$$= \frac{\sum_n P(A_n B)}{P(B)} = \sum_n \left[\frac{P(A_n B)}{P(B)}\right] = \sum_n P(A_n \mid B)$$

Hence the theorem.

Remark. For given B satisfying $P(B) > 0$, the conditional probability $P[\cdot|B]$ also enjoys the same properties as the unconditional probability.

For example, in the usual notations, we have:

(*i*) $P[\phi \mid B] = 0$

(*ii*) $P[\overline{A} \mid B] = 1 - P[A \mid B]$

(*iii*) $P[\bigcup_{i=1}^{n} A_i \mid B] = \sum_{i=1}^{n} P[A_i \mid B]$,

where $A_1, A_2, \ldots, A_n$ are mutually disjoint events.

(*iv*) $P(A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P(A_2 \mid B) - P(A_1 A_2 \mid B)$

(*v*) If $E \subset F$, then $P(E \mid B) \leq P(F \mid B)$

and so on.

The proofs of results (*iv*) and (*v*) are given in theorems 4·9 and 4·13 respectively. Others are left as exercises to the reader.

Theorem 4·12. *For any three events, A, B and C defined on the sample space S such that $B \subset C$ and $P(A) > 0$,*

$$P(B \mid A) \leq P(C \mid A)$$

Proof.
$$P(C \mid A) = \frac{P(C \cap A)}{P(A)} \qquad \text{(By definition)}$$

$$= \frac{P[B \cap C \cap A) \cup (\overline{B} \cap C \cap A)}{P(A)}$$

$$= \frac{P[B \cap C \cap A)}{P(A)} + \frac{P(\overline{B} \cap C \cap A)}{P(A)} \qquad \text{(Using axiom 3)}$$

$$= P[(B \cap C|A) + (\overline{B} \cap C \cap A)]$$

Now $B \subset C \quad \Rightarrow \quad B \cap C = B$

$\therefore \quad P(C \mid A) = P(B \mid A) + P(\overline{B} \cap C|A)$

$\Rightarrow \quad P(C \mid A) \geq P(B \mid A)$

4·7·3. Independent Events. *An event B is said to be independent (or statistically independent) of event A, if the conditional probability of B given A i.e., $P(B \mid A)$ is equal to the unconditional probability of B, i.e., if*

$$P(B \mid A) = P(B)$$

Since

$$P(A \cap B) = P(B \mid A)\, P(A) = P(A \mid B)\, P(B)$$

and since $P(B \mid A) = P(B)$ when B is independent of A, we must have $P(A \mid B) = P(A)$ or it follows that A is also independent of B. Hence the events A and B are independent if and only if

$$P(A \cap B) = P(A)\, P(B) \qquad \ldots(4·9)$$

4·7·4. Pairwise Independent Events

Definition. *A set of events $A_1, A_2, \ldots, A_n$ are said to be pair-wise independent if*

$$P(A_i \cap A_j) = P(A_i)\, P(A_j) \quad \forall \quad i \neq j \qquad \ldots(4·10)$$

4·7·5. Conditions for Mutual Independence of n Events. Let S denote the sample space for a number of events. The events in S are said to be mutually indepdendent if the probability of the simultaneous occurrence of (any) finite number of them is equal to the product of their separate probabilities.

If $A_1, A_2, ..., A_n$ are n events, then for their mutual independence, we should have

$$(i) \quad P(A_i \cap A_j) = P(A_i)\,P(A_j), \ (i \neq j \ ; i, j = 1, 2, ..., n)$$

$$(ii) \quad P(A_i \cap A_j \cap A_k) = P(A_i)\,P(A_j)\,P(A_k), \ (i \neq j \neq k \ ; i \ ; j \ ; k = 1, 2, ..., n)$$

$$\vdots \qquad\qquad \vdots$$

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)\,P(A_2)\,...\,P(A_n)$$

It is interesting to note that the above equations give respectively ${}^nC_2, {}^nC_3, ..., {}^nC_n$ conditions to be satisfied by $A_1, A_2, ..., A_n$.

Hence the total number of conditions for the mutual independence of $A_1, A_2, ..., A_n$ is ${}^nC_2 + {}^nC_3 + ... + {}^nC_n$.

Since ${}^nC_0 + {}^nC_1 + {}^nC_2 + ... + {}^nC_n = 2^n$, we get the required number of conditions as $(2^n - 1 - n)$.

In particular for three events A_1, A_2 and A_3, $(n = 3)$, we have the following $2^3 - 1 - 3 = 4$, conditions for their mutual independence.

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)\ P(A_2) \\ P(A_2 \cap A_3) &= P(A_2)\ P(A_3) \\ P(A_1 \cap A_3) &= P(A_1)\ P(A_3) \\ P(A_1 \cap A_2 \cap A_3) &= P(A_1)\ P(A_2)\ P(A_3) \end{aligned} \qquad ...(4·11)$$

Remarks. 1. It may be observed that pairwise or mutual independence of events $A_1, A_2, ..., A_n$, is defined only when $P(A_i) \neq 0$, for $i = 1, 2, ..., n$.

2. If the events A and B are such that $P(A_i) \neq 0$, $P(B) \neq 0$ and A is independent of B, then B is independent of A.

Proof. We are given that

$$P(A \mid B) = P(A)$$

$$\Rightarrow \quad \frac{P(A \cap B)}{P(B)} = P(A)$$

$$\Rightarrow \quad P(A \cap B) = P(A)\ P(B)$$

$$\Rightarrow \quad \frac{P(B \cap A)}{P(A)} = P(B) \qquad [\because P(A) \neq 0 \text{ and } A \cap B = B \cap A]$$

$$\Rightarrow \quad P(B \mid A) = P(B),$$

which by definition of independent events, means that B is independent of A.

3. It may be noted that pairwise independence of events does not imply their mutual independence. For illustrations, see Examples 4·50 and 4·51.

Theorem 4·13. *If A and B are independent events then A and $\bar{B}$ are also independent events.*

Proof. By theorem 4·4, we have

$$P(A \cap \overline{B}) = P(A) - P(A \cap B)$$
$$= P(A) - P(A)P(B) \quad [\because A \text{ and } B \text{ are independent}]$$
$$= P(A)[1 - P(B)]$$
$$= P(A)\,P(\overline{B})$$

$\Rightarrow$ A and $\overline{B}$ are independent events.

Aliter. $P(A \cap B) = P(A)\,P(B) = P(A)\,P(B \mid A) = P(B)\,P(A \mid B)$

i.e., $P(B \mid A) = P(B) \Rightarrow B$ is independent of A.

also $P(A \mid B) = P(A) \Rightarrow A$ is independent of B.

Also $P(B \mid A) + P(\overline{B} \mid A) = 1 \Rightarrow P(B) + P(\overline{B} \mid A) = 1$

or $P(\overline{B} \mid A) = 1 - P(B) = P(\overline{B})$

$\therefore$ $\overline{B}$ is independent of A and by symmetry we say that A is independent of $\overline{B}$. Thus A and $\overline{B}$ are independent events.

Remark. Similarly, we can prove that if A and B are independent events then $\overline{A}$ and B are also independent events.

Theorem 4·14. *If A and B are independent events then $\overline{A}$ and $\overline{B}$ are also independent events.*

Proof. We are given $P(A \cap B) = P(A)\,P(B)$

Now $P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$
$$= 1 - [P(A) + P(B) - P(A \cap B)]$$
$$= 1 - [P(A) + P(B) - P(A)P(B)]$$
$$= 1 - P(A) - P(B) + P(A)P(B)$$
$$= [1 - P(B)] - P(A)[1 - P(B)]$$
$$= [1 - P(A)][1 - P(B)] = P(\overline{A})\,P(\overline{B})$$

$\therefore$ $\overline{A}$ and $\overline{B}$ are independent events.

Aliter. We know

$$P(\overline{A} \mid \overline{B}) + P(A \mid \overline{B}) = 1$$
$\Rightarrow$ $$P(\overline{A} \mid \overline{B}) + P(A) = 1 \qquad (c.f.\ \text{Theorem 4·13})$$
$\Rightarrow$ $$P(\overline{A} \mid \overline{B}) = 1 - P(A) = P(\overline{A})$$

$\therefore$ $\overline{A}$ and $\overline{B}$ are independent events.

Theorem 4·15. *If A, B, C are mutually independent events then $A \cup B$ and C are also independent.*

Proof. We are required to prove:

$P[(A \cup B) \cap C] = P(A \cup B)\,P(C)$

L.H.S. $= P[(A \cap C) \cup (B \cap C)]$ [Distributive Law]
$$= P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)$$
$$= P(A)\,P(C) + P(B)\,P(C) - P(A)\,P(B)\,P(C)$$
[$\because$ A, B and C are mutually independent]
$$= P(C)\,[P(A) + P(B) - P(A \cap B)]$$

$= P(C)\, P(A \cup B) =$ R.H.S.

Hence $(A \cup B)$ and C are independent.

Theorem 4·16. *If A, B and C are random events in a sample space and if A, B and C are pairwise independent and A is independent of $(B \cup C)$, then A, B and C are mutually independent.*

Proof. We are given

$$\left.\begin{aligned} P(A \cap B) &= P(A)P(B) \\ P(B \cap C) &= P(B)P(C) \\ P(A \cap C) &= P(A)P(C) \\ P[A \cap (B \cup C)] &= P(A)P(B \cup C) \end{aligned}\right\} \quad \ldots (*)$$

Now $P[A \cap (B \cup C)] = P[(A \cap B) \cup (A \cap C)]$

$= P(A \cap B) + P(A \cap C) - P[A \cap B) \cap (A \cap C)]$

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

and $P(A)\, P(B \cup C) = P(A)[P(B) + P(C) - P(B \cap C)]$

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

From (**) and (***), on using (*), we get

$P(A \cap B \cap C) = P(A)\, P(B \cap C) = P(A)\, P(B)\, P(C)$

Hence A, B, C are mutually independent.

Theorem 4·17. *For any two events A and B,*

$$P(A \cap B) \le P(A) \le P(A \cup B) \le P(A) + P(B)$$

[Patna Univ. B.A.(Stat. Hons.), 1992; Delhi Univ. B.Sc.(Stat. Hons.), 1989]

Proof. We have

$$A = (A \cap \overline{B}) \cup (A \cap B)$$

Using axiom 3, we have

$$P(A) = P[(A \cap \overline{B}) \cup (A \cap B)] = P(A \cap \overline{B}) + P(A \cap B)$$

Now $P[(A \cap \overline{B}) \ge 0$ (From axiom 1)

$\therefore \quad P(A) \ge P(A \cap B)$ ..(*)

Similarly $P(B) \ge P(A \cap B)$

$\Rightarrow \quad P(B) - P(A \cap B) \ge 0$

Now $P(A \cup B) = P(A) + [P(B) - P(A \cap B)]$ ...(**)

$\therefore \quad P(A \cup B) \ge P(A) \quad \Rightarrow \quad P(A) \le P(A \cup B)$ ...(***)

Also $P(A \cup B) \le P(A) + P(B)$ [From (**)]

Hence from (*), (**) and (***), we get

$$P(A \cap B) \le P(A) \le P(A \cup B) \le P(A) + P(B)$$

Aliter. Since $A \cap B \subset A$, by Theorem 4·6 *(ii)* page 4·30, we get

$$P(A \cap B) \le P(A).$$

Also $A \subset (A \cup B) \quad \Rightarrow \quad P(A) \le P(A \cup B)$

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

$$\le P(A) + P(B) \qquad [\because P(A \cap B) \ge 0]$$

Combining the above results, we get

$$P(A \cap B) \le P(A) \le P(A \cup B) \le P(A) + P(B)$$

Example 4·12. *Two dice, one green and the other red, are thrown. Let A be the event that the sum of the points on the faces shown is odd, and B be the event of at least one ace (number '1').*

(a) Describe the (i) complete sample space, (ii) events A, B, $\overline{B}$, $A \cap B$, $A \cup B$, and $A \cap \overline{B}$ and find their probabilities assuming that all the 36 sample points have equal probabilities.

(b) Find the probabilities of the events :

(i) $(\overline{A} \cup \overline{B})$ *(ii)* $(\overline{A} \cap \overline{B})$ *(iii)* $(A \cap \overline{B})$ *(iv)* $(\overline{A} \cap B)$ *(v)* $(\overline{A \cap B})$ *(vi)* $(\overline{A} \cup B)$ *(vii)* $(\overline{A \cup B})$ *(viii)* $\overline{A} \cap (A \cup B)$ *(ix)* $A \cup (\overline{A} \cap B)$ *(x)* $(A \mid B)$ and $(B \mid A)$, and *(xi)* $(\overline{A} \mid \overline{B})$ *and* $(\overline{B} \mid \overline{A})$ . .

Solution. *(a)* The sample space consists of the 36 elementary events .

$(1,1)$; $(1,2)$; $(1,3)$; $(1,4)$; $(1,5)$; $(1,6)$
$(2,1)$; $(2,2)$; $(2,3)$; $(2,4)$; $(2,5)$; $(2,6)$
$(3,1)$; $(3,2)$; $(3,3)$; $(3,4)$; $(3,5)$; $(3,6)$
$(4,1)$; $(4,2)$; $(4,3)$; $(4,4)$; $(4,5)$; $(4,6)$
$(5,1)$; $(5,2)$; $(5,3)$; $(5,4)$; $(5,5)$; $(5,6)$
$(6,1)$; $(6,2)$; $(6,3)$; $(6,4)$; $(6,5)$; $(6,6)$

where, for example, the ordered pair (4, 5) refers to the elementary event that the green die shows 4 and and the red die shows 5.

A = The event that the sum of the numbers shown by the two dice is odd.

$= \{(1,2);(2,1);(1,4);(2,3);(3,2);(4,1);(1,6);(2,5)$
$(3,4);(4,3);(5,2);(6,1);(3,6);(4,5);(5,4);(6,3)$
$(5,6);(6,5)\}$ and therefore

$$P(A) = \frac{n(A)}{n(S)} = \frac{18}{36}$$

B = The event that at least one face is 1,

$= \{(1,1)$; $(1,2)$; $(1,3)$; $(1,4)$; $(1,5)$; $(1,6)$
$(2,1)$; $(3,1)$; $(4,1)$; $(5,1)$; $(6,1)\}$ and therefore

$$P(B) = \frac{n(B)}{n(S)} = \frac{11}{36}$$

$\overline{B}$ = The event that each of the face obtained is not an ace.

$= \{(2,2);(2,3);(2,4);(2,5);(2,6);(3,2);(3,3);$
$(3,4);(3,5);(3,6);(4,2);(4,3);(4,4);(4,5);$
$(4,6);(5,2);(5,3);(5,4);(5,5);(5,6);(6,2);$
$(6,3);(6,4);(6,5);(6,6)\}$ and therefore

$$P(\overline{B}) = \frac{n(\overline{B})}{n(S)} = \frac{25}{36}$$

$A \cap B$ = The event that sum is odd and at least one face is an ace.

$= \{(1,2);(2,1);(1,4);(4,1);(1,6);(6,1)\}$

$$\therefore \quad P(A \cap B) = \frac{n(A \cap B)}{n(S)} = \frac{6}{36} = \frac{1}{6}$$

$$A \cup B = \{(1,2);(2,1);(1,4);(2,3);(3,2);(4,1);(1,6);(2,5)$$
$$(3,4);(4,3);(5,2);(6,1);(3,6);(4,5);(5,4);(6,3)$$
$$(5,6);(6,5);(1,1);(1,3);(1,5);(3,1);(5,1)\}$$

$$\therefore \quad P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{23}{36}$$

$$A \cap \bar{B} = \{(2,3);(3,2);(2,5);(3,4);(3,6);(4,3);(4,5);(5,2)$$
$$(5,4);(5,6);(6,3);(6,5)\}$$

$$P(A \cap \bar{B}) = \frac{n(A \cap \bar{B})}{n(S)} = \frac{12}{36} = \frac{1}{3}$$

(b)(i) $P(\bar{A} \cup \bar{B}) = P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - \frac{1}{6} = \frac{5}{6}$

(ii) $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - \frac{23}{36} = \frac{13}{36}$

(iii) $P(A \cap \bar{B}) = P(A) - P(A \cap B) = \frac{18}{36} - \frac{6}{36} = \frac{12}{36} = \frac{1}{3}$

(iv) $P(\bar{A} \cap B) = P(B) - P(A \cap B) = \frac{11}{36} - \frac{6}{36} = \frac{5}{36}$

(v) $P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - \frac{1}{6} = \frac{5}{6}$

(vi) $P(\bar{A} \cup B) = P(\bar{A}) + P(B) - P(\bar{A} \cap B)$
$$= \left(1 - \frac{18}{36}\right) + \frac{11}{36} - \frac{5}{36} = \frac{2}{3}$$

(vii) $P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - \frac{23}{36} = \frac{13}{36}$

(viii) $P[\bar{A} \cap (A \cup B)] = P[(A \cap \bar{A}) \cup (\bar{A} \cap B)]$
$$= P(\bar{A} \cap B) = \frac{5}{36} \qquad [\because A \cap \bar{A} = \phi]$$

(ix) $P[A \cup (\bar{A} \cap B)] = P(A) + P(\bar{A} \cap B) - P(A \cap \bar{A} \cap B)$
$$= P(A) + P(\bar{A} \cap B) = \frac{18}{36} + \frac{5}{36} = \frac{23}{36}$$

(x) $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{6/36}{11/36} = \frac{6}{11}$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{6/36}{18/36} = \frac{6}{18} = \frac{1}{3}$$

(xi) $P(\bar{A} \mid \bar{B}) = \frac{P(\bar{A} \cap \bar{B})}{P(\bar{B})} = \frac{13/36}{25/36} = \frac{13}{25}$

$$P(\bar{B} \mid \bar{A}) = \frac{P(\bar{A} \cap \bar{B})}{P(\bar{A})} = \frac{13/36}{18/36} = \frac{13}{18}$$

Example 4·13. *If two dice are thrown, what is the probability that the sum is (a) greater than 8, and (b) neither 7 nor 11?*

Solution. *(a)* If S denotes the sum on the two dice, then we want $P(S > 8)$. The required event can happen in the following mutually exclusive ways:

(i) $S = 9$ *(ii)* $S = 10$ *(iii)* $S = 11$ *(iv)* $S = 12$.

Hence by addition theorem of probability

$$P(S > 8) = P(S = 9) + P(S = 10) + P(S = 11) + P(S = 12)$$

In a throw of two dice, the sample space contains $6^2 = 36$ points.

The number of favourable cases can be enumerated as follows:

$S = 9$: (3, 6), (6, 3), (4, 5), (5, 4), *i.e.*, 4 sample points.

$\therefore \quad P(S = 9) = \frac{4}{36}$

$S = 10$: (4,6), (6,4), (5,5), *i.e.*, 3 sample points.

$\therefore \quad P(S = 10) = \frac{3}{36}$

$S = 11$: (5, 6), (6, 5), *i.e.*, 2 sample points.

$\therefore \quad P(S = 11) = \frac{2}{36}$

$S = 12$: (6, 6), *i.e.*, 1 sample point.

$\therefore \quad P(S = 12) = \frac{1}{36}$

$\therefore \quad P(S > 8) = \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{10}{36} = \frac{5}{18}$

(b) Let A denote the event of getting the sum of 7 and B denote the event of getting the sum of 11 with a pair of dice.

$S = 7$: (1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3), *i.e.*, 6 distinct sample points.

$\therefore \quad P(A) = P(S = 7) = \frac{6}{36} = \frac{1}{6}$

$S = 11$: (5, 6), (6, 5), $P(B) = P(S = 11) = \frac{2}{36} = \frac{1}{18}$

$\therefore$ Required probability $= P(\overline{A} \cap \overline{B}) = 1 - P(A \cup B)$

$= 1 - [P(A) + P(B)]$

($\because$ A and B are disjoint events)

$= 1 - \frac{1}{6} - \frac{1}{18} = \frac{7}{9}$

Example 4·14. *An urn contains 4 tickets numbered 1, 2, 3, 4 and another contains 6 tickets numbered 2, 4, 6, 7, 8, 9. If one of the two urns is chosen at random and a ticket is drawn at random from the chosen urn, find the probabilities that the ticket drawn bears the number (i) 2 or 4, (ii) 3, (iii) 1 or 9*

[Calicut Univ. B.Sc., 1992]

Solution. *(i)* Required event can happen in the following mutually exclusive ways:

(I) First urn is chosen and then a ticket is drawn.

(II) Second urn is chosen and then a ticket is drawn.

Since the probability of choosing any urn is ½, the required probability 'p' is given by

$$p = P(I) + P(II)$$

$$= \frac{1}{2} \times \frac{2}{4} + \frac{1}{2} \times \frac{2}{6} = \frac{5}{12}$$

(ii) Required probability $= \frac{1}{2} \times \frac{1}{4} + \frac{1}{2} \times 0 = \frac{1}{8}$

($\because$ in the 2nd urn there is no ticket with number 3)

(iii) Required probability $= \frac{1}{2} \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{6} = \frac{5}{24}$

Example 4·15. *A card is drawn from a well-shuffled pack of playing cards. What is the probability that it is either a spade or an ace ?*

Solution. The equiprobable sample space S of drawing a card from a well-shuffled pack of playing cards consists of 52 sample points.

If A and B denote the events of drawing a 'spade card' and 'an ace' respectively then A consists of 13 sample points and B consists of 4 sample points so that,

$$P(A) = \frac{13}{52} \quad \text{and} \quad P(B) = \frac{4}{52}$$

The compound event $A \cap B$ consists of only one sample point, *viz.*, ace of spade so that,

$$P(A \cap B) = \frac{1}{52}$$

The probability that the card drawn is either a spade or an ace is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{4}{13}$$

Example 4·16. *A box contains 6 red, 4 white and 5 black balls. A person draws 4 balls from the box at random. Find the probability that among the balls drawn there is at least one ball of each colour.* **(Nagpur Univ. B.Sc., 1992)**

Solution. The required event E that 'in a draw of 4 balls from the box at random there is at least one ball of each colour', can materialise in the following mutually disjoint ways :

(i) 1 Red, 1 White, 2 Black balls

(ii) 2 Red, 1 White, 1 Black balls

(iii) 1 Red, 2 White, 1 Black balls.

Hence by the addition theorem of probability, the required probability is given by

$$P(E) = P(i) + P(ii) + P(iii)$$
$$= \frac{{}^6C_1 \times {}^4C_1 \times {}^5C_2}{{}^{15}C_4} + \frac{{}^6C_2 \times {}^4C_1 \times {}^5C_1}{{}^{15}C_4} + \frac{{}^6C_1 \times {}^4C_2 \times {}^5C_1}{{}^{15}C_4}$$
$$= \frac{1}{{}^{15}C_4} \left[6 \times 4 \times 10 + 15 \times 4 \times 5 + 6 \times 6 \times 5 \right]$$
$$= \frac{4!}{15 \times 14 \times 13 \times 12} \left[240 + 300 + 180 \right]$$
$$= \frac{24 \times 720}{15 \times 14 \times 13 \times 12} = 0{\cdot}5275$$

Example 4·17. *Why does it pay to bet consistently on seeing 6 at least once in 4 throws of a die, but not on seeing a double six at least once in 24 throws with two dice? (de Mere's Problem).*

Solution. The probability of getting a '6' in a throw of die $=1/6$.

$\therefore$ The probability of not getting a '6' in a throw of die

$$=1-1/6=5/6.$$

By compound probability theorem, the probability that in 4 throws of a die no '6' is obtained $=(5/6)^4$

Hence the probability of obtaining '6' at least once in 4 throws of a die $=1-(5/6)^4=0{\cdot}516$

Now, if a trial consists of throwing two dice at a time, then the probability of getting a 'double' of '6' in a trial $=1/36$.

Thus the probability of not getting a 'double of 6' in a trial $=35/36$.

The probability that in 24 throws, with two dice each, no 'double of 6' is obtained $=(35/36)^{24}$

Hence the probability of getting a 'double of 6' at least once in 24 throws $=1-(35/36)^{24}=0{\cdot}491$.

Since the probability in the first case is greater than the probability in the second case, the result follows.

Example 4·18. *A problem in Statistics is given to the three students A, B and C whose chances of solving it are 1/2, 3/4, and 1/4 respectively.*

What is the probability that the problem will be solved if all of them try independently? **[Madurai Kamraj Univ. B.Sc.,1986; Delhi Univ. B.A.,1991]**

Solution. Let A, B, C denote the events that the problem is solved by the students A, B, C respectively. Then

$$P(A)=\frac{1}{2},\ P(B)=\frac{3}{4}\ \text{and}\ P(C)=\frac{1}{4}$$

The problem will be solved if at least one of them solves the problem. Thus we have to calculate the probability of occurrence of at least one of the three events A, B, C, *i.e.*, $P(A\cup B\cup C)$.

$$\begin{aligned}P(A\cup B\cup C)&=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)\\&\qquad -P(B\cap C)+P(A\cap B\cap C)\\&=P(A)+P(B)+P(C)-P(A)\,P(B)-P(A)\,P(C)\\&\qquad -P(B)\,P(C)+P(A)\,P(B)\,P(C)\end{aligned}$$

($\because A, B, C$ are independent events.)

$$\begin{aligned}&=\frac{1}{2}+\frac{3}{4}+\frac{1}{4}-\frac{1}{2}\cdot\frac{3}{4}-\frac{3}{4}\cdot\frac{1}{4}\\&\qquad -\frac{1}{2}\cdot\frac{1}{4}+\frac{1}{2}\cdot\frac{3}{4}\cdot\frac{1}{4}\\&=\frac{29}{32}\end{aligned}$$

Aliter. $P(A \cup B \cup C) = 1 - P(\overline{A \cup B \cup C})$

$$= 1 - P(\overline{A} \cap \overline{B} \cap \overline{C})$$

$$= 1 - P(\overline{A})P(\overline{B})P(\overline{C})$$

$$= 1 - \left(1 - \frac{1}{2}\right)\left(1 - \frac{3}{4}\right)\left(1 - \frac{1}{4}\right)$$

$$= \frac{29}{32}$$

Example 4·19. *If* $A \cap B = \phi$, *then show that* ...(*)

$$P(A) \leq P(\overline{B})$$

[Delhi Univ. B.Sc. (Maths Hons.) 1987]

Solution. We have

$$A = (A \cap B) \cup (A \cap \overline{B})$$

$$= \phi \cup (A \cap \overline{B}) \qquad \text{[Using *]}$$

$$= A \cap \overline{B}$$

$$\Rightarrow \quad A \subseteq \overline{B}$$

$$\Rightarrow \quad P(A) \leq P(\overline{B})$$

as desired.

Aliter. Since $A \cap B = \phi$, we have $A \subset \overline{B}$, which implies that $P(A) \leq P(\overline{B})$.

Example 4·20. *Let A and B be two events such that*

$$P(A) = \frac{3}{4} \text{ and } P(B) = \frac{5}{8}$$

show that

(a) $P(A \cup B) \geq \frac{3}{4}$

(b) $\frac{3}{8} \leq P(A \cap B) \leq \frac{5}{8}$

[Delhi Univ. B.Sc. Stat (Hons.) 1986,1988]

Solution. *(i)* We have

$$A \subset (A \cup B)$$

$$\Rightarrow \quad P(A) \leq P(A \cup B)$$

$$\Rightarrow \quad \frac{3}{4} \leq P(A \cup B)$$

$$\Rightarrow \quad P(A \cup B) \geq \frac{3}{4}$$

(ii) $A \cap B \subseteq B$

$$\Rightarrow \quad P(A \cap B) \leq P(B) = \frac{5}{8} \qquad ...(i)$$

Also $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq 1$

$$\Rightarrow \quad \frac{3}{4} + \frac{5}{8} - 1 \leq P(A \cap B)$$

$$\Rightarrow \qquad \frac{6+5-8}{8} \leq P(A \cap B)$$

$$\Rightarrow \qquad \frac{3}{8} \leq P(A \cap B) \qquad \ldots(ii)$$

From *(i)* and *(ii)* we get

$$\frac{3}{8} \leq P(A \cap B) \leq \frac{5}{8}$$

Example 4·21. *(Chebychev's Problem). What is the chance that two numbers, chosen at random, will be prime to each other ?*

Solution. If any number 'a' is divided by a prime number 'r', then the possible remainders are 0, 1, 2, ...r–1. Hence the chance that 'a' is divisible by r is $1/r$ (because the only case favourable to this is remainder being 0). Similarly, the probability that any number 'b' chosen at random is divisible by r is $1/r$. Since the numbers a and b are chosen at random, the probability that none of them is divisible by 'r' is given (by compound probability theorem) by :

$$\left(1-\frac{1}{r}\right) \times \left(1-\frac{1}{r}\right) = \left(1-\frac{1}{r}\right)^2; \quad r = 2, 3, 5, 7, \ldots$$

Hence the required probability that the two numbers chosen at random are prime to each other is given by

$$P = \prod_r \left(1-\frac{1}{r}\right)^2, \quad \text{where } r \text{ is a prime number.}$$

$$= \frac{6}{\pi^2} \qquad \text{(From trigonometry)}$$

Example 4·22. *A bag contains 10 gold and 8 silver coins. Two successive drawings of 4 coins are made such that : (i) coins are replaced before the second trial, (ii) the coins are not replaced before the second trial. Find the probability that the first drawing will give 4 gold and the second 4 silver coins.*

[Allahabad Univ. B.Sc., 1987]

Solution. Let A denote the event of drawing 4 gold coins in the first draw and B denote the event of drawing 4 silver coins in the second draw. Then we have to find the probability of $P(A \cap B)$.

(i) Draws with replacement. If the coins drawn in the first draw are replaced back in the bag before the second draw then the events A and B are independent and the required probability is given (using the multiplication rule of probability) by the expression

$$P(A \cap B) = P(A) \cdot P(B) \qquad \ldots(*)$$

Ist draw. Four coins can be drawn out of 10+8=18 coins in ${}^{18}C_4$ ways, which gives the exhaustive number of cases. In order that all these coins are of gold, they must be drawn out of the 10 gold coins and this can be done in ${}^{10}C_4$ ways. Hence

$$P(A) = {}^{10}C_4 \;/\; {}^{18}C_4$$

2nd draw. When the coins drawn in the first draw are replaced before the 2nd draw, the bag contains 18 coins. The probability of drawing 4 silver coins in the 2nd draw is given by $P(B) = {}^{8}C_4 \;/\; {}^{18}C_4$.

Substituting in (*), we have

$$P(A \cap B) = \frac{{}^{10}C_4}{{}^{18}C_4} \times \frac{{}^{8}C_4}{{}^{18}C_4}$$

(ii) Draws without replacement. If the coins drawn are not eplaced back before the second draw, then the events A and B are not independent and the required probability is given by

$$P(A \cap B) = P(A) \,.\, P(B \mid A) \qquad \ldots(**)$$

As discussed in part *(i)*, $P(A) = {}^{10}C_4 \;/\; {}^{18}C_4$.

Now, if the 4 gold coins which were drawn in the first draw are not replaced back, there are $18 - 4 = 14$ coins left in the bag and $P(B \mid A)$ is the probability of drawing 4 silver coins from the bag containing 14 coins out of which 6 are gold coins and 8 are silver coins.

Hence $P(B \mid A) = {}^{8}C_4 \;/\; {}^{14}C_4$

Substituting in (**) we get

$$P(A \cap B) = \cdot \frac{{}^{10}C_4}{{}^{18}C_4} \times \frac{{}^{8}C_4}{{}^{14}C_4}$$

Example 4·23. *A consignment of 15 record players contains 4 defectives. The record players are selected at random, one by one, and examined. Those examined are not put back. What is the probability that the 9th one examined is the last defective?*

Solution. Let A be the event of getting exactly 3 defectives in examination of 8 record players and let B the event that the 9th piece examined is a defective one.

Since it is a problem of sampling without replacement and since there are 4 defectives out of 15 record players, we have

$$P(A) = \frac{\binom{4}{3} \times \binom{11}{5}}{\binom{15}{8}}$$

$P(B \mid A)$ = Probability that the 9th examined record player is defective given that there were 3 defectives in the first 8 pieces examined.

$= 1/7,$

since there is only one defective piece left among the remaining $15 - 8 = 7$ record players.

Hence the required probability is

$$P(A \cap B) = P(A) \,.\, P(B \mid A)$$

$$= \frac{\binom{4}{3} \times \binom{11}{5}}{\binom{15}{8}} \times \frac{1}{7} = \frac{8}{195}$$

Example 4·24. *p is the probability that a man aged x years will die in a year. Find the probability that out of n men $A_1, A_2, ..., A_n$ each aged x, A_1 will die in a year and will be the first to die.* **[Delhi Univ. B.Sc., 1985]**

Solution. Let E_i, $(i = 1, 2, ..., n)$ denote the event that A_i dies in a year. Then

$$P(E_i) = p, (i = 1, 2, ..., n) \text{ and } P(\overline{E}_i) = 1 - p.$$

The probability that none of n men $A_1, A_2, ..., A_n$ dies in a year

$$= P(\overline{E}_1 \cap \overline{E}_2 \cap ... \cap \overline{E}_n) = P(\overline{E}_1) P(\overline{E}_2) ... P(\overline{E}_n)$$

(By compound probability theorem)

$$= (1 - p)^n$$

$\therefore$ The probability that at least one of $A_1, A_2, ..., A_n$, dies in a year

$$= 1 - P(\overline{E}_1 \cap \overline{E}_2 \cap ... \cap \overline{E}_n) = 1 - (1 - p)^n$$

The probability that among n men, A_1 is the first to die is $1/n$ and since this event is independent of the event that at least one man dies in a year, required probability is

$$\frac{1}{n}\left[1 - (1 - p)^n\right]$$

Example 4·25. *The odds against Manager X settling the wage dispute with the workers are 8:6 and odds in favour of manager Y settling the same dispute are 14:16.*

(i) What is the chance that neither settles the dispute, if they both try, independently of each other?

(ii) What is the probability that the dispute will be settled?

Solution. Let A be the event that the manager X will settle the dispute and B be the event that the Manager Y will settle the dispute. Then clearly

$$P(\overline{A}) = \frac{8}{8+6} = \frac{4}{7} \quad \Rightarrow \quad P(A) = 1 - P(\overline{A}) = \frac{6}{14} = \frac{3}{7}$$

$$P(B) = \frac{14}{14+16} = \frac{7}{15} \quad \Rightarrow \quad P(\overline{B}) = 1 - P(B) = \frac{16}{14+16} = \frac{8}{15}$$

The required probability that neither settles the dispute is given by :

$$P(\overline{A} \cap \overline{B}) = P(\overline{A}) \times P(\overline{B}) = \frac{4}{7} \times \frac{8}{15} = \frac{32}{105}$$

[Since A and B are independent $\Rightarrow$ $\overline{A}$ and $\overline{B}$ are also independent]

(ii) The dispute will be settled if at least one of the managers X and Y settles the dispute. Hence the required probability is given by:

$$P(A \cup B) = \text{Prob. [At least one of X and Y settles the dispute]}$$

$$= 1 - \text{Prob. [None settles the dispute]}$$

$$= 1 - P(\overline{A} \cap \overline{B}) = 1 - \frac{32}{105} = \frac{73}{105}$$

Example 4·26. *The odds that person X speaks the truth are 3:2 and the odds that person Y speaks the truth are 5:3. In what percentage of cases are they likely to contradict each other on an identical point.*

Solution. Let us define the events:

A : X speaks the truth, B : Y speaks the truth

Then $\overline{A}$ and $\overline{B}$ represent the complementary events that X and Y tell a lie respectively. We are given:

$$P(A) = \frac{3}{3+2} = \frac{3}{5} \quad \Rightarrow \quad P(\overline{A}) = 1 - \frac{3}{5} = \frac{2}{5}$$

and $$P(B) = \frac{5}{5+3} = \frac{5}{8} \quad \Rightarrow \quad P(\overline{B}) = 1 - \frac{5}{8} = \frac{3}{8}$$

The event E that X and Y contradict each other on an identical point can happen in the following mutually exclusive ways:

(i) X speaks the truth and Y tells a lie, *i.e.*, the event $A \cap \overline{B}$ happens,

(ii) X tells a lie and Y speaks the truth, *i.e.*, the event $\overline{A} \cap B$ happens.

Hence by addition theorem of probability the required probability is given by:

$$P(E) = P(i) + P(ii) = P(A \cap \overline{B}) + P(\overline{A} \cap B)$$

$$= P(A) \cdot P(\overline{B}) + P(\overline{A}) \cdot P(B),$$

[Since A and B are independent]

$$= \frac{3}{5} \times \frac{3}{8} + \frac{2}{5} \times \frac{5}{8} = \frac{19}{40} = 0{\cdot}475$$

Hence A and B are likely to contradict each other on an identical point in 47·5% of the cases.

Example 4·27. *A special dice is prepared such that the probabilities of throwing 1, 2, 3, 4, 5 and 6 points are :*

$$\frac{1-k}{6}, \frac{1+2k}{6}, \frac{1-k}{6}; \frac{1+k}{6}, \frac{1-2k}{6}, \text{ and } \frac{1+k}{6}$$

respectively. If two such dice are thrown, find the probability of getting a sum equal to 9. **[Delhi Univ. B.Sc. (Stat. Hons.), 1988]**

Solution. Let (x, y) denote the numbers obtained in a thrown of two dice, x denoting the number on the first dice and y denoting the number on the second dice. The sum $S = x+y = 9$, can be obtained in the following mutually disjoint ways:

(i) (3, 6), *(ii)* (6, 3), *(iii)* (4, 5), *(iv)* (5, 4)

Hence by addition theorem of probability:

$$P(S=9) = P(3,6) + P(6,3) + P(4,5) + P(5,4)$$

$$= P(x=3)\,P(y=6) + P(x=6)\,P(y=3) + P(x=4)\,P(y=5) + P(x=5)\,P(y=4),$$

since the number on one dice is independent of the number on the other dice.

$$\therefore \quad P(S=9) = \frac{(1-k)}{6} \cdot \frac{(1+k)}{6} + \frac{(1+k)}{6} \cdot \frac{(1-k)}{6} + \frac{(1+k)}{6} \cdot \frac{(1-2k)}{6} + \frac{(1-2k)}{6} \cdot \frac{(1+k)}{6}$$

$$= 2\left(\frac{1+k}{36}\right) \mid (1-k) + (1-2k) \mid$$

$$= \frac{1}{18}(1+k)(2-3k)$$

Example 4·28. *(a) A and B alternately cut a pack of cards and the pack is shuffled after each cut. If A starts and the game is continued until one cuts a diamond, what are the respective chances of A and B first cutting a diamond?*

(b) One shot is fired from each of the three guns. E_1, E_2, E_3 denote the events that the target is hit by the first, second and third gun respectively. If $P(E_1) = 0{\cdot}5$, $P(E_2) = 0{\cdot}6$ and $P(E_3) = 0{\cdot}8$ and, E_1, E_2, E_3 are independent events, find the probability that (a) exactly one hit is registered, (b) at least two hits are registered.

Solution. *(a)* Let E_1 and E_2, denote the events of A and B cutting a diamond respectively. Then

$$P(E_1) = P(E_2) = \frac{13}{52} = \frac{1}{4} \quad \Rightarrow \quad P(\overline{E_1}) = P(\overline{E_2}) = \frac{3}{4}$$

If A starts the game, he can first cut the diamond in the following mutually exclusive ways:

(i) E_1 happens, *(ii)* $\overline{E_1} \cap \overline{E_2} \cap E_1$ happens, *(iii)* $\overline{E_1} \cap \overline{E_2} \cap \overline{E_1} \cap \overline{E_2} \cap E_1$ happens, and so on. Hence by addition theorem of probability, the probability 'p' that A first wins is given by

$$p = P(i) + P(ii) + P(iii) + \ldots\ldots$$

$$= P(E_1) + P(\overline{E_1} \cap \overline{E_2} \cap E_1) + P(\overline{E_1} \cap \overline{E_2} \cap \overline{E_1} \cap \overline{E_2} \cap E_1) + \ldots$$

$$= P(E_1) + P(\overline{E_1})\,P(\overline{E_2})\,P(E_1) + P(\overline{E_1})\,P(\overline{E_2})\,P(\overline{E_1})\,P(\overline{E_2})\,P(E_1) + \ldots$$

(By Compound Probability Theorem)

$$= \frac{1}{4} + \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} + \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} + \ldots\ldots$$

$$= \frac{\frac{1}{4}}{1 - \frac{9}{16}} = \frac{4}{7}$$

The probability that B first cuts a diamond

$$= 1 - p = 1 - \frac{4}{7} = \frac{3}{7}$$

(b) We are given

$$P(\overline{E_1}) = 0{\cdot}5,\ P(\overline{E_2}) = 0{\cdot}4 \text{ and } P(\overline{E_3}) = 0{\cdot}2$$

(a) Exactly one hit can be registered in the following mutually exclusive ways: *(i)* $E_1 \cap \overline{E_2} \cap \overline{E_3}$ happens, *(ii)* $\overline{E_1} \cap E_2 \cap \overline{E_3}$ happens, *(iii)* $\overline{E_1} \cap \overline{E_2} \cap E_3$ happens.

Hence by addition probability theorem, the required probability 'p' is given by :

$$p = P(E_1 \cap \overline{E_2} \cap \overline{E_3}) + P(\overline{E_1} \cap E_2 \cap \overline{E_3}) + P(\overline{E_1} \cap \overline{E_2} \cap E_3)$$

$$= P(E_1)\,P(\overline{E_2})\,P(\overline{E_3}) + P(\overline{E_1})\,P(E_2)\,P(\overline{E_3}) + P(\overline{E_1})\,P(\overline{E_2})\,P(E_3)$$

(Since E_1, E_2 and E_3 are independent)

$$= 0{\cdot}5 \times 0{\cdot}4 \times 0{\cdot}2 + 0{\cdot}5 \times 0{\cdot}6 \times 0{\cdot}2 + 0{\cdot}5 \times 0{\cdot}4 \times 0{\cdot}8 = 0{\cdot}26.$$

(b) At least two hits can be registered in the following mutually exclusive ways: *(i)* $E_1 \cap E_2 \cap \overline{E_3}$ happens *(ii)* $E_1 \cap \overline{E_2} \cap E_3$ happens, *(iii)* $\overline{E_1} \cap E_2 \cap E_3$ happens. *(iv)* $E_1 \cap E_2 \cap E_3$ happens.

Required probability

$$= P(E_1 \cap E_2 \cap \overline{E_3}) + P(E_1 \cap \overline{E_2} \cap E_3) + P(\overline{E_1} \cap E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_3)$$

$$= 0{\cdot}5 \times 0{\cdot}6 \times 0{\cdot}2 + 0{\cdot}5 \times 0{\cdot}4 \times 0{\cdot}8 + 0{\cdot}5 \times 0{\cdot}6 \times 0{\cdot}8 + 0{\cdot}5 \times 0{\cdot}6 \times 0{\cdot}8$$

$$= 0{\cdot}06 + 0{\cdot}16 + 0{\cdot}24 + 0{\cdot}24 = 0{\cdot}70$$

Example 4·29. *Three groups of children contain respectively 3 girls and 1 boy, 2 girls and 2 boys, and 1 girl and 3 boys. One child is selected at random from each group. Show that the chance that the three selected consist of 1 girl and 2 boys is 13/32.* **[Madurai Univ. B.Sc.,1988; Nagpur Univ. B.Sc.,1991]**

Solution. The required event of getting 1 girl and 2 boys among the three selected children can materialise in the following three mutually disjoint cases:

Group No. →	I	II	III
(i)	Girl	Boy	Boy
(ii)	Boy	Girl	Boy
(iii)	Boy	Boy	Girl

Hence by addition theorem of probability,

$$\text{Required probability} = P(i) + P(ii) + P(iii) \qquad \ldots(*)$$

Since the probability of selecting a girl from the first group is 3/4, of selecting a boy from the second is 2/4, and of selecting a boy from the third group is 3/4, and since these three events of selecting children from three groups are independent of each other, by compound probability theorem, we have

$$P(i) = \frac{3}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{9}{32}$$

Similarly, we have

$$P(ii) = \frac{1}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{3}{32}$$

and $$P(iii) = \frac{1}{4} \times \frac{2}{4} \times \frac{1}{4} = \frac{1}{32}$$

Substituting in (*), we get

Required probability $= \frac{9}{32} + \frac{3}{32} + \frac{1}{32} = \frac{13}{32}$

EXERCISE 4 (b)

1. *(a)* Which function defines a probability space on $S = (e_1, e_2, e_3)$

(i) $P(e_1) = \frac{1}{4}$, $P(e_2) = \frac{1}{3}$, $P(e_3) = \frac{1}{2}$

(ii) $P(e_1) = \frac{2}{3}$, $P(e_2) = -\frac{1}{3}$, $P(e_3) = \frac{2}{3}$

(iii) $P(e_1) = \frac{1}{4}$, $P(e_2) = \frac{1}{3}$, $P(e_3) = \frac{1}{2}$, and

(iv) $P(e_1) = 0$, $P(e_2) = \frac{1}{3}$, $P(e_3) = \frac{2}{3}$

Ans. *(i)* No, *(ii)* No, *(iii)* No, and *(iv)* Yes

(b) Let $S = (e_1, e_2, e_3, e_4)$, and let P be a probability function on S.

(i) Find $P(e_1)$, if $P(e_2) = \frac{1}{3}, P(e_3) = \frac{1}{6}, P(e_4) = \frac{1}{9}$,

(ii) Find $P(e_1)$ and $P(e_2)$ if $P(e_3) = P(e_4) = \frac{1}{4}$ and $P(e_1) = 2P(e_2)$, and

(iii) Find $P(e_1)$ if $P[(e_2, e_3)] = \frac{2}{3}, P[(e_2, e_4)] = \frac{1}{2}$ and $P(e_2) = \frac{1}{3}$.

Ans. *(i)* $P(e_1) = \frac{7}{18}$, *(ii)* $P(e_1) = \frac{1}{3}$, $P(e_2) = \frac{1}{6}$, and *(iii)* $P(e_1) = \frac{1}{6}$

2. *(a)* With usual notations, prove that

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

Deduce a similar result for $P(A \cup B \cup C)$, where C is one more event.

(b) For any event : E_i, $P(E_i) = p_i$, $(i = 1, 2, 3)$; $P(E_i \cap E_j) = p_{ij}$, $(i, j = 1, 2, 3)$ and $P(E_1 \cap E_2 \cap E_3) = p_{123}$, find the probability that of the three events, *(i)* at least one, and *(ii)* exactly one happens.

(c) Discuss briefly the axiomatic approach to probability, illustrating by examples how it meets the deficiencies of the classical approach.

(d) If A and B are any two events, state the results giving

(i) $P(A \cup B)$ and *(ii)* $P(A \cap B)$.

A and B are mutually exclusive events and $P(A) = \frac{1}{2}, P(B) = \frac{1}{3}$. Find $P(A \cup B)$ and $P(A \cap B)$.

3. Let $S = \left\{1, \frac{1}{2}, \left(\frac{1}{2}\right)^2, ..., \left(\frac{1}{2}\right)^n\right\}$, be a classical event space and A, B be events given by

$$A = \left\{1, \frac{1}{2}\right\}, \quad B = \left\{\left(\frac{1}{2}\right)^k \mid k \text{ is an even positive integer}\right\}$$

Find $P(\bar{A} \cap \bar{B})$ **[Calcutta Univ. B.Sc. (Stat Hons.), 1986]**

4. What is a 'probability space'? State *(i)* the 'law of total probability' and *(ii)* Boole's inequality for events not necessarily mutually exclusive.

5. *(a)* Explain the following with examples:

(i) random experiment, *(ii)* an event, *(iii)* an event space. State the axioms of probability and explain their frequency interpretations.

A man forgets the last digit of a telephone number, and dials the last digit at random. What is the probability of calling no more than three wrong numbers?

(b) Define conditional probability and give its frequency interpretation. Show that conditional probabilities satisfy the axioms of probability.

6. Prove the following laws, in each case assuming the conditional probabilities being defined.

(a) $P(E \mid E) = 1$, *(b)* $P(\phi \mid F) = 0$

(c) If $E_1 \subseteq E_2$, then $E(E_1 \mid F) < P(E_2 \mid F)$

(d) $P(\bar{E} \mid F) = 1 - P(E \mid F)$

(e) $P(E_1 \cup E_2 \mid F) = P(E_1 \mid F) + P(E_2 \mid F) - P(P(E_1 \cap E_2 \mid F)$

(f) If $P(F) = 1$ then $P(E \mid F) = P(E)$

(g) $P(E - F) = P(E) - P(E \cap F)$

(h) If $P(F) > 0$, and E and F are mutully exclusive then $P(E \mid F) = 0$

(i) If $P(E \mid F) = P(E)$, then $P(E \mid \bar{F}) = P(E)$ and $P(\bar{E} \mid F) = P(\bar{E})$

7. *(a)* If $P(\bar{A}) = a$, $P(\bar{B}) = b$, then prove that $P(A \cap B) \geq 1 - a - b$.

(b) If $P(A) = \alpha$, $P(B) = \beta$, then prove that $P(A \mid B) \geq (\alpha + \beta - 1)/\beta$.

Hint. In each case use $P(A \cup B) \leq 1$

8. Prove or disprove:

(a) (i) If $P(A \mid B) \geq P(A)$, then $P(B \mid A) \geq P(B)$

(ii) If $P(A) = P(\bar{B})$, then $A = \bar{B}$.

[Delhi Univ. B.Sc. (Maths Hons.), 1988]

(b) If $P(A) = 0$, then $A = \phi$

[Delhi Univ. B.Sc. (Maths Hons.), 1990]

Ans. Wrong.

(c) For possible events A, B, C,

(i) If $P(A) > P(B)$, then $P(A \mid C) > P(B \mid C)$

(ii) If $P(A \mid C) \geq P(B \mid C)$ and $P(A \mid \bar{C}) \geq P(B \mid \bar{C})$, then $P(A) \geq P(B)$. **[Delhi Univ. B.Sc.(Maths Hons),1989]**

(d) If $P(A) = 0$, then $P(A \cap B) = 0$.

[Delhi Univ. B.Sc. (Maths Hons.), 1986]

(e) (i) If $P(A) = P(B) = p$, then $P(A \cap B) \leq p^2$

(ii) If $P(B \mid \bar{A}) = P(B \mid A)$, then A and B are independent.

[Delhi Univ. B.Sc. (Maths Hons.), 1990]

(f) If $P(A) > 0, P(B) > 0$ and $P(A \mid B) = P(B \mid A)$, then $P(A) = P(B)$.

9. *(a)* Let A and B be two events, neither of which has probability zero. Then if A and B are disjoint, A and B are independent.

[Delhi Univ. B.Sc.(Stat. Hons.), 1986]

(b) Under what conditions does the following equality hold?

$$P(A) = P(A \mid B) + P(A \mid \bar{B})$$

[Punjab Univ. B.Sc. (Maths Hons.), 1992]

Ans. $B = S$ or $\bar{B} = S$

10. *(a)* If A and B are two events and the probability $P(B) \neq 1$, prove that

$$P(A \mid \bar{B}) = \frac{[P(A) - P(A \cap B)]}{[1 - P(B)]}$$

where $\bar{B}$ denotes the event complementary to B and hence deduce that

$$P(A \cap B) \geq P(A) + P(B) - 1$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

Also show that $P(A) > \text{ or } < P(A \mid B)$ according as

$$P(A \mid \bar{B}) > \text{ or } < P(A).$$

[Sri Venkat. Univ. B.Sc. 1992 ; Karnatak Univ. B.Sc.1991]

Hint. *(i)*

$$P(A \mid \bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{[P(A) - P(A \cap B)]}{[1 - P(B)]}$$

(ii) Since $P(A \mid \bar{B}) \leq 1$, $P(A) - P(A \cap B) \leq 1 - P(B)$

$\Rightarrow$ $P(A) + P(B) - 1 \leq P(A \cap B)$

(iii) $$\frac{P(A \mid \bar{B})}{P(A)} = \frac{P(\bar{B} \mid A)}{P(\bar{B})} = \frac{1 - P(B \mid A)}{1 - P(B)}$$

Now $P(A \mid \bar{B}) > P(A)$ if $\{1 - P(B \mid A)\} > \{1 - P(B)\}$

i.e., if $P(B \mid A) < P(B)$

i.e., if $\frac{P(B \mid A)}{P(B)} < 1$

i.e., if $\frac{P(A \mid B)}{P(A)} < 1$ *i.e.*, if $P(A) > P(A \mid B)$

(b) If A and B are two mutually exclusive events show that

$$P(A \mid \bar{B}) = P(A)/[1 - P(B)]$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

(c) If A and B are two mutually exclusive events and $P(A \cup B) \neq 0$, then

$$P(A \mid A \cup B) = \frac{P(A)}{P(A) + P(B)}$$ **[Guahati Univ. B.Sc. 1991]**

(d) If A and B are two independent events show that

$$P(A \cup B) = 1 - P(\bar{A}) P(\bar{B})$$

(e) If $\overline{A}$ denotes the non-occurrence of A, then prove that

$$P(\overline{A}_1 \cup \overline{A}_2 \cup \overline{A}_3) = 1 - P(A_1)\ P(A_2 \mid A_1)\ P(A_3 \mid A_1 \cap A_2)$$

[Agra Univ. B.Sc., 1987]

11. If A, B and C are three arbitrary events and

$$S_1 = P(A) + P(B) + P(C)$$
$$S_2 = P(A \cap B) + P(B \cap C) + P(C \cap A)$$
$$S_3 = P(A \cap B \cap C).$$

Prove that the probability that exactly one of the three events occurs is given by $S_1 - 2S_2 + 3S_3$.

12. *(a)* For the events $A_1, A_2, \ldots, A_n$ assuming

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i), \text{ prove that}$$

(i) $P\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(\overline{A}_i)$ and that

(ii) $P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$

[Sardar Patel Univ. B.Sc. Nov.1992]

(b) Let A, B and C denote events. If $P(A \mid C) \geq P(B \mid C)$ and $P(A \mid \overline{C}) \geq P(B \mid \overline{C})$, then show that $P(A) \geq P(B)$.

[Calcutta Univ. B.Sc. (Maths Hons.), 1992]

13. *(a)* If A and B are independent events defined on a given probability space $(\Omega, A, P(.))$, then prove that A and $\overline{B}$ are independent, $\overline{A}$ and $\overline{B}$ are independent. **[Delhi Univ. B.Sc. (Maths Hons.), 1988]**

(b) A, B and C are three events such that A and B are independent, $P(C) = 0$. Show that A, B and C are independent.

(c) An event A is known to be independent of the events B, $B \cup C$ and $B \cap C$. Show that it is also independent of C. **[Nagpur Univ. B.Sc.1992]**

(d) Show that if an event C is independent of two mutually exclusive events A and B, then C is also independent of $A \cup B$.

(e) The outcome of an experiment is equally likely to be one of the four points in three-dimensional space with rectangular coordinates $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$ and $(1, 1, 1)$. Let E, F and G be the events : x-coordinate=1, y-coordinate=1 and z-coordinate=1, respectively. Check if the events E, F and G are independent. **(Calcutta Univ. B.Sc., 1988)**

14. Explain what is meant by "Probability Space". You fire at a target with each of the three guns; A, B and C denote respectively the event — hit the target with the first, second and third gun. Assuming that the events are independent and have probabilities $P(A) = a$, $P(B) = b$ and $P(C) = c$, express in terms of A, B and C the following events:

(i) You will not hit the target at all.

(*ii*) You will hit the target at least twice. Find also the probabilities of these events. **[Sardar Patel Univ. B.Sc., 1990)**

15. (*a*) Suppose A and B are any two events and that $P(A) = p_1$, $P(B) = p_2$ and $P(A \cap B) = p_3$. Show that the formula of each of the following probabilities in terms of p_1, p_2 and p_3 can be expressed as follows :

(*i*) $P(\overline{A} \cup \overline{B}) = 1 - p_2$ (*ii*) $P(\overline{A} \cap \overline{B}) = 1 - p_1 - p_2 + p_3$

(*iii*) $P(A \cap \overline{B}) = p_1 - p_3$ (*iv*) $P(\overline{A} \cap B) = p_2 - p_3$

(*v*) $P(\overline{A \cap B}) = 1 - p_3$ (*vi*) $P(\overline{A} \cup B) = 1 - p_1 + p_3$

(*vii*) $P(\overline{A \cup B}) = 1 - p_1 - p_2 + p_3$ (*viii*) $P[\overline{A} \cap (A \cup B)] = p_2 - p_3$

(*ix*) $P[A \cup (\overline{A} \cap B)] = p_1 + p_2 - p_3$

(*x*) $P(A \mid B) = \dfrac{p_3}{p_2}$ and $P(B \mid A) = \dfrac{p_3}{p_1}$

(*xi*) $P(\overline{A} \mid \overline{B}) = \dfrac{1 - p_1 - p_2 + p_3}{1 - p_2}$ and $P(\overline{B} \mid \overline{A}) = \dfrac{1 - p_1 - p_2 + p_3}{1 - p_1}$

[Allahabad Univ. B.Sc. (Stat.), 1991]

(*b*) If $P(A) = 1/3$, $P(B) = 3/4$ and $P(A \cup B) = 11/12$, find $P(A \mid B)$ and $P(B \mid A)$.

(*c*) Let $P(A) = p$, $P(A \mid B) = q$, $P(B \mid A) = r$. Find the relation between the numbers p, q and r such that $\overline{A}$ and $\overline{B}$ are mutually exclusive.

[Delhi Univ. B.Sc. (Maths Hons.), 1985]

Hint. $P(AB) = P(A)\,P(B \mid A) = P(B) \cdot P(A \mid B)$

$\Rightarrow \quad P(AB) = pr = P(B).\, q \quad \Rightarrow \quad P(B) = pr/q$

If $\overline{A}$ and $\overline{B}$ are mutually disjoint, then $P(\overline{A} \cap \overline{B}) = 0$.

$\Rightarrow \quad 1 - P(A \cup B) = 0 \Rightarrow 1 - [p + (pr/q) - pr] = 0$

16. (*a*) In terms of probabilities, $p_1 = P(A)$, $p_2 = P(B)$ and $p_3 = P(A \cap B)$;

Express (*i*) $P(A \cup B)$, (*ii*) $P(A \mid B)$, (*iii*) $P(\overline{A} \cap B)$ under the condition that (*i*) A and B are mutually exclusive, (*ii*) A and B are mutually independent.

(*b*) Let A and B be the possible outcomes of an experiment and suppose $P(A) = 0{\cdot}4$, $P(A \cup B) = 0{\cdot}7$ and $P(B) = p$

(*i*) For what choice of p are A and B mutually exclusive ?

(*ii*) For what choice of p are A and B independent ?

[Aligarh Univ. B.Sc., 1988 ; Guwahati Univ. B.Sc., 1991]

Ans. (*i*) 0·3, (*ii*) 0·5

(*c*) Let A_1, A_2, A_3, A_4 be four independent events for which $P(A_1) = p$, $P(A_2) = q$, $P(A_3) = r$ and $P(A_4) = s$. Find the probability that

(*i*) at least one of the events occurs, (*ii*) exactly two of the events occur, and (*iii*) at most three of the events occur. **[Civil Services (Main), 1985]**

17. (*a*) Two six-faced unbiased dice are thrown. Find the probability that the sum of the numbers shown is 7 or their product is 12.

Ans. 2/9

(b) Defects are classifed as A, B or C, and the following probabilities have been determined from available production data :

$P(A) = 0{\cdot}20, P(B) = 0{\cdot}16, P(C) = 0{\cdot}14, P(A \cap B) = 0{\cdot}08, P(A \cap C) = 0{\cdot}05, P(B \cap C) = 0{\cdot}04$, and $P(A \cap B \cap C) = 0{\cdot}02$.

What is the probability that a randomly selected item of product will exhibit at least one type of defect ? What is the probability that it exhibits both A and B defects but is free from type C defect ? **[Bombay Univ. B.Sc., 1991]**

(c) A language class has only three students A, B, C and they independently attend the class. The probabilities of attendance of A, B and C on any given day are $1/2, 2/3$ and $3/4$ respectively. Find the probability that the total number of attendances in two consecutive days is exactly three.

[Lucknow Univ. B.Sc. 1990; Calcutta Univ. B.Sc.(Maths Hons.), 1986]

18. *(a)* Cards are drawn one by one from a full deck. What is the probability that exactly 10 cards will precede the first ace. **[Delhi Univ. B.Sc.,1988]**

Ans. $\left(\frac{48}{52} \times \frac{47}{51} \times \frac{46}{50} \times \ldots \times \frac{39}{43}\right) \times \frac{4}{42} = \frac{164}{4165}$

(b) Each of two persons tosses three fair coins. What is the probability that they obtain the same number of heads.

Ans. $\left(\frac{1}{8}\right)^2 + \left(\frac{3}{8}\right)^2 + \left(\frac{3}{8}\right)^2 + \left(\frac{1}{8}\right)^2 = \frac{5}{16}$.

19. *(a)* Given that A, B and C are mutually exclusive events, explain why each of the following is not a permissible assignment of probabilities.

(i) $P(A) = 0{\cdot}24$, $P(B) = 0{\cdot}4$ and $P(A \cup C) = 0{\cdot}2$,

(ii) $P(A) = 0{\cdot}7$, $P(B) = 0{\cdot}1$ and $P(B \cap C) = 0{\cdot}3$

(iii) $P(A) = 0{\cdot}6$, $P(A \cap \overline{B}) = 0{\cdot}5$

(b) Prove that for n arbitrary independent events $A_1, A_2, \ldots, A_n$

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n) + P(\overline{A_1})\, P(\overline{A_2}) \ldots P(\overline{A_n}) = 1.$$

(c) $A_1, A_2, \ldots, A_n$ are n independent events with

$$P(A_i) = 1 - \frac{1}{\alpha^i}, \quad i = 1, 2, \ldots, n.$$

Find the value of $P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n)$. **(Nagpur Univ. B.Sc., 1987)**

Ans. $1 - \frac{1}{\alpha^{n(n+1)/2}}$

(d) Suppose the events $A_1, A_2, \ldots, A_n$ are independent and that $P(A_i) = \frac{1}{i+1}$ for $1 \le i \le n$. Find the probability that none of the n events occurs, justifying each step in your calculations.

Ans. $1/(n+1)$

20. *(a)* A denotes getting a heart card, B denotes getting a face card (King, Queem or Jack), $\overline{A}$ and $\overline{B}$ denote the complementary events. A card is drawn at

random from a full deck. Compute the following probabilities.

(*i*) $P(A)$, (*ii*) $P(A \cap \bar{B})$, (*iii*) $P(A \cup B)$, (*iv*) $P(A \cap B)$,

(*v*) $P(\bar{A} \cup B)$.

Assume natural assignment of probabilities.

Ans. (*i*) 1/4, (*ii*) 5/26, (*iii*) 11/26, (*iv*) 3/5, (*v*) 21/26.

(*b*) A town has two doctors X and Y operating independently. If the probability that doctor X is available is 0·9 and that for Y is 0·8, what is the probability that at least one doctor is available when needed? **[Gorakhpur Univ. B.Sc., 1988]**

Ans. 0·98

21. (*a*) The odds that a book will be favourably reviewed by 3 independent critics are 5 to 2, 4 to 3 and 3 to 4 respectively. What is the probability that, of the three reviews, a majority will be favourable? **[Gauhati Univ. BSc., 1987]**

Ans. 209/343.

(*b*) A, B and C are independent witnesses of an event which is known to have occurred. A speaks the truth three times out of four, B four times out of five and C five times out of six. What is the probability that the occurrence will be reported truthfully by majority of three witnesses?

Ans. 31/60.

(*c*) A man seeks advice regarding one of two possible courses of action from three advisers who arrived at their recommendations independently. He follows the recommendation of the majority. The probability that the individual advisers are wrong are 0·1, 0·05 and 0·05 respectively. What is the probability that the man takes incorrect advise ? **[Gujarat Univ. B.Sc., 1987]**

22. (*a*) The odds against a certain event are 5 to 2 and odds in favour of another (independent) event are 6 to 5. Find the chance that at least one of the events will happen. **(Madras Univ. B.Sc.,1987)**

Ans. 52/77.

(*b*) A person takes four tests in succession. The probability of his passing the first test is p, that of his passing each succeeding test is p or $p/2$ according as he passes or fails the preceding one. He qualifies provided he passes at least three tests. What is his chance of qualifying. **[Gauhati Univ. B.Sc. (Hons.) 1988]**

23. (*a*) The probability that a 50-years old man will be alive at 60 is 0·83 and the probability that a 45-years old woman will be alive at 55 is 0·87. What is the probability that a man who is 50 and his wife who is 45 will both be alive 10 years hence?

Ans. 0·7221.

(*b*) It is 8:5 against a husband who is 55 years old living till he is 75 and 4:3 against his wife who is now 48, living till she is 68. Find the probability that (*i*) the couple will be alive 20 years hence, and (*ii*) at least one of them will be alive 20 years hence.

Ans (*i*) 15/91, (*ii*) 59/91.

(c) A husband and wife appear in an interview for two vacancies in the same post. The probability of husband's selection is 1/7 and that of wife's selection is 1/5. What is the probability that only one of them will be selected ?

Ans. 2/7 **[Delhi Univ. B.Sc., 1986]**

24. *(a)* The chances of winning of two race-horses are 1/3 and 1/6 respectively. What is the probability that at least one will win when the horses are running *(a)* in different races, and *(b)* in the same race?

Ans. *(a)* 8/18 *(b)* 1/2

(b) A problem in statistics is given to three students whose chances of solving it are 1/2, 1/3 and 1/4. What is the probability that the problem will be solved?

Ans. 3/4 **[Meerut Univ. B.Sc., 1990]**

25. *(a)* Ten pairs of shoes are in a closet. Four shoes are selected at random. Find the probability that there will be at least one pair among the four shoes selected?

Ans. $1 - \dfrac{{}^{10}C_4 \times 2^4}{{}^{20}C_4}$

(b) From 100 tickets numbered 1, 2, ... , 100 four are drawn at random. What is the probability that 3 of them will bear number from 1 to 20 and the fourth will bear any number from 21 to 100 ?

Ans. $\dfrac{{}^{20}C_3 \times {}^{80}C_1}{{}^{100}C_4}$

26. A six faced die is so biased that it is twice as likely to show an even number as an odd number when thrown. It is thrown twice. What is the probability that the sum of the two numbers thrown is odd?

Ans. 4/9

27. From a group of 8 children, 5 boys and 3 girls, three children are selected at random. Calculate the probabilities that selected group contains *(i)* no girl, *(ii)* only one girl, *(iii)* one particular girl, *(iv)* at least one girl, and *(v)* more girls than boys.

Ans. *(i)* 5/28, *(ii)* 15/28, *(iii)* 5/28, *(iv)* 23/28, *(v)* 2/7.

28. If three persons, selected at random, are stopped on a street, what are the probabilities that :

(a) all were born on a Friday;

(b) two were born on a Friday and the other on a Tuesday;

(c) none was born on a Monday.

Ans. *(a)* 1/343, *(b)* 3/343, *(c)* 216/343.

29. *(a)* *A* and *B* toss a coin alternately on the understanding that the first who obtains the head wins. If *A* starts, show that their respective chances of winning are 2/3 and 1/3.

(b) *A*, *B* and *C*, in order, toss a coin. The first one who throws a head wins. If *A* starts, find their respectivre chances of winning. (Assume that the game may

continue indefinitely.)

Ans. 4/7, 2/7, 1/7.

(c) A man alternately tosses a coin and throws a die, beginning with coin. What is the probability that he will get a head before he gets a '5 or 6' on die?

Ans. 3/4.

30. *(a)* Two ordinary six-sided dice are tossed.

(i) What is the probability that both the dice show the number 5.

(ii) What is the probability that both the dice show the same number.

(iii) Given that the sum of two numbers shown is 8, find the conditional probability that the number noted on the first dice is larger than the number noted on the second dice.

(b) Six dice are thrown simultaneously. What is the probability that all will show different faces?

31. *(a)* A bag contains 10 balls, two of which are red, three blue and five black. Three balls are drawn at random from the bag, that is every ball has an equal chance of being included in the three. What is the probability that

(i) the three balls are of different colours,

(ii) two balls are of the same colour, and

(iii) the balls are all of the same colour?

Ans. *(i)* 30/120, *(ii)* 79/120, *(iii)* 11/120.

(b) A is one of six horses entered for a race and is to be ridden by one of the two jockeys B and C. It is 2 to 1 that B rides A, in which case all the horses are equally likely to win, with rider C, A's chance is trebled.

(i) Find the probability that A wins.

(ii) What are odds against A's winning?

[Shivaji Univ. B.Sc. (Stat. Hons.), 1992]

Hint. Probability of A's winning

$$= P(B \text{ rides } A \text{ and } A \text{ wins}) + P(C \text{ rides } A \text{ and } A \text{ wins})$$

$$= \frac{2}{3} \times \frac{1}{6} + \frac{1}{3} \times \frac{3}{6} = \frac{5}{18}$$

$\therefore$ Probability of A's losing $= 1 - 5/18 = 13/18$.

Hence odds against A's winning are : 13/18 : 5/18, *i.e.*, 13 : 5.

32. *(a)* Two-third of the students in a class are boys and the rest girls. It is known that the probability of a girl getting a first class is 0·25 and that of boy getting a first class is 0·28. Find the probability that a student chosen at random will get first class marks in the subject.

Ans. 0·27

(b) You need four eggs to make omelettes for breakfast. You find a dozen eggs in the refrigerator but do not realise that two of these are rotten. What is the probability that of the four eggs you choose at random

(i) none is rotten,

(ii) exactly one is rotten?

Ans. *(i)* 625/1296 *(ii)* 500/1296.

(c) The probability of occurrence of an event A is 0·7, the probability of non-occurrence of another event B is 0·5 and that of at least one of A or B not occurring is 0·6. Find the probability that at least one of A or B occurs.

[Mysore Univ. B.Sc., 1991]

33. *(a)* The odds against A solving a certain problem are 4 to 3 and odds in favour of B solving the same problem are 7 to 5. What is the probability that the problem is solved if they both try independently? **[Gujarat Univ.B.Sc., 1987]**

Ans. 16/21

(b) A certain drug manufactured by a company is tested chemically for its toxic nature. Let the event 'the drug is toxic' be denoted by E and the event 'the chemical test reveals that the drug is toxic' be denoted by F. Let $P(E) = \theta$, $P(F \mid E) = P(\bar{F} \mid \bar{E}) = 1 - \theta$. Then show that probability that the drug is not toxic given that the chemical test reveals that it is toxic is free from θ.

Ans. 1/2 **[M.S. Baroda Univ. B.Sc., 1992]**

34. A bag contains 6 white and 9 black balls. Four balls are drawn at a time. Find the probability for the first draw to give 4 white and the second draw to give 4 black balls in each of the following cases :

(i) The balls are replaced before the second draw.

(ii) The balls are not replaced before the second draw.

[Jammu Univ. B.Sc., 1992]

Ans. *(i)* $\dfrac{{}^6C_4}{{}^{15}C_4} \times \dfrac{{}^9C_4}{{}^{15}C_4}$ *(ii)* $\dfrac{{}^6C_4}{{}^{15}C_4} \times \dfrac{{}^9C_4}{{}^{11}C_4}$

35. The chances that doctor A will diagnose a disease X correctly is 60%. The chances that a patient will die by his treatment after correct diagnosis is 40% and the chance of death by wrong diagnosis is 70%. A patient of doctor A, who had disease X, died. What is the chance that his disease was diagnosed correctly?

Hint. Let us define the following events:

E_1 : Disease X is diagnosed correctly by doctor A.

E_2 : A patient (of doctor A) who has disease X dies.

Then we are given :

$$P(E_1) = 0{\cdot}6 \quad \Rightarrow \quad P(\bar{E}_1) = 1 - 0{\cdot}6 = 0{\cdot}4$$

and $P(E_2 \mid E_1) = 0{\cdot}4$ and $P(E_2 \mid \bar{E}_1) = 0{\cdot}7$

$$\text{We want } P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{P(E_1 \cap E_2)}{P(E_1 \cap E_2) + P(\bar{E}_1 \cap E_2)} = \frac{6}{13}$$

36. The probability that at least 2 of 3 people A, B and C will survive for 10 years is 247/315. The probability that A alone will survive for 10 years is 4/105 and the probability that C alone will die within 10 years is 2/21. Assuming that the events of the survival of A, B and C can be regarded as independent, calculate the

probability of surviving 10 years for each person.

Ans. 3/5, 5/7, 7/9.

37. A and B throw alternately a pair of unbiased dice, A beginning. A wins if he throws 7 before B throws 6, and B wins if he throws 6 before A throws 7. If A and B respectively denote the events that A wins and B wins the series, and a and b respectively denote the events that it is A's and B's turn to throw the dice, show that

$$(i)\ P(A \mid a) = \frac{1}{6} + \frac{5}{6} P(A \mid b),\ (ii)\ P(A \mid b) = \frac{31}{36} P(A \mid a),$$

$$(iii)\ P(B \mid a) = \frac{5}{6} P(B \mid b), \text{ and } (iv)\ P(B \mid b) = \frac{5}{36} + \frac{13}{36} P(B \mid a),$$

Hence or otherwise, find $P(A \mid a)$ and $P(B \mid a)$. Also comment on the result that $P(A \mid a) + P(B \mid a) = 1$.

38. A bag contains an assortment of blue and red balls. If two balls are drawn at randon, the probability of drawing two red balls is five times the probability of drawing two blue balls. Furthermore, the probability of drawing one ball of each colour is six times the probabililty of drawing two blue balls. How many red and blue balls are there in the bag?

Hint. Let number of red and blue balls in the bag be r and b respectively. Then

$$p_1 = \text{Prob. of drawing two red balls} = \frac{r(r-1)}{(r+b)(r+b-1)}$$

$$p_2 = \text{Prob. of drawing two blue balls} = \frac{b(b-1)}{(r+b)(r+b-1)}$$

$$p_3 = \text{Prob. of drawing one red and one blue ball} = \left[\frac{2br}{(r+b)(r+b-1)}\right].$$

Now $p_1 = 5p_2$ and $p_3 = 6p_2$

$\therefore$ $r(r-1) = 5b(b-1)$ and $2br = 6b(b-1)$

Hence $b = 3$ and $r = 6$.

39. Three newspapers A, B and C are published in a certain city. It is estimated from a survey that 20% read A; 16% read B, 14% read C, 8% read A and B, 5% read A and C, 4% read B and C and 2% read all the three newspapers. What is the probability that a normally chosen person

(i) does not read any paper, *(ii)* does not read C

(iii) reads A but not B, *(iv)* reads only one of these papers, and

(v) reads only two of these papers.

Ans. *(i)* 0·65, *(ii)* 0·86, *(iii)* 0·12, *(iv)* 0·22, *(v)* 0·11.

40. *(a)* A die is thrown twice, the event space S consisting of the 36 possible pairs of outcomes (a,b) each assigned probability 1/36. Let A, B and C denote the following events :

$A = \{(a,b) \mid a \text{ is odd}\}$, $B = \{(a,b) \mid b \text{ is odd}\}$, $C = \{(a,b) \mid a + b \text{ is odd.}\}$

Check whether A, B and C are independent or independent in pairs only.

[Calcutta Univ. B.Sc. Hons., 1985]

(b) Eight tickets numbered 111, 121, 122, 122, 211, 212, 212, 221 are placed in a hat and stirred. One of them is then drawn at random. Show that the event A : "the first digit on the ticket drawn will be 1", B : "the second digit on the ticket drawn will be 1," and C : "the third digit on the ticket drawn will be 1", are not pairwise independent although

$$P(A \cap B \cap C) = P(A)\,P(B)\,P(C)$$

41. *(a)* Four identical marbles marked 1, 2, 3 and 123 respectively are put in an urn and one is drawn at random. Let A_i, $(i = 1, 2, 3)$, denote the event that the number i appears on the drawn marble. Prove that the events A_1, A_2 and A_3 are pairwise independent but not mutually independent.

[Gauhati Univ. B.Sc. (Hons.), 1988]

Hint. $P(A_1) = \frac{1}{2} = P(A_2) = P(A_3)$; $P(A_1 A_2) = P(A_1 A_3) = P(A_2 A_3) = \frac{1}{4}$

$$P(A_1 A_2 A_3) = \frac{1}{4}.$$

(b) Two fair dice are thrown independently. Define the following events :

A : Even number on the first dice

B : Even number on the second dice.

C : Same number on both dice.

Discuss the independence of the events A, B and C.

(c) A die is of the shape of a regular tetrahedron whose faces bear the numbers 111, 112, 121, 122. A_1, A_2, A_3 are respectively the events that the first two, the last two and the extreme two digits are the same, when the die is tossed at random. Find whether or not the events A_1, A_2, A_3 are *(i)* pairwise independent, *(ii)* mutually (*i.e.* completely) independent. Determine $P(A_1 \mid A_2 A_3)$) and explain its value by by argument. **[Civil Services (Main), 1983]**

42. *(a)* For two events A and B we have the following probabilities:

$$P(A) = P(A \mid B) = \frac{1}{4} \text{ and } P(B \mid A) = \frac{1}{2}.$$

Check whether the following statements are true or false :

(i) A and B are mutually exclusive, *(ii)* A and B are independent, *(iii)* A is a subevent of B, and *(iv)* $P(\bar{A} \mid B) = \frac{3}{4}$

Ans. *(i)* False, *(ii)* True, *(iii)* False, and *(iv)* True.

(b) Consider two events A and B such that $P(A) = 1/4$, $P(B \mid A) = 1/2$, $P(A \mid B) = 1/4$. For each of the following statements, ascertain whether it is true or false :

(i) A is a sub-event of B, *(ii)* $P(\bar{A} \mid \bar{B}) = 3/4$,

(iii) $P(A \mid B) + P(A \mid \bar{B}) = 1$

43. *(a)* Let A and B be two events such that $P(A) = \frac{3}{4}$ and $P(B) = \frac{5}{8}$.

Show that

(*i*) $P(A \cup B) \geq \frac{3}{4}$, (*ii*) $\frac{3}{8} \leq P(A \cap B) \leq \frac{5}{8}$, and (*iii*) $\frac{1}{8} \leq P(A \cap \bar{B}) \leq \frac{3}{8}$.

[Coimbatore Univ. B.E., Nov. 1990; Delhi Univ. B.Sc.(Stat. Hons.),1986]

(b) Given two events A and B. If the odds against A are 2 to 1 and those in favour of $A \cup B$ are 3 to 1, show that

$$\frac{5}{12} \leq P(B) \leq \frac{3}{4}$$

Give an example in which $P(B) = 3/4$ and one in which $P(B) = 5/12$.

44. Let A and B be events, neither of which has probability zero. Prove or disprove the following events :

(i) If A and B are disjoint, A and B are independent,

(ii) If A and B are independent, A and B are disjoint.

45. *(a)* It is given that $P(A_1 \cup A_2) = \frac{5}{6}$, $P(A_1 \cap A_2) = \frac{1}{3}$ and $P(\bar{A}_2) = \frac{1}{2}$, where $P(\bar{A}_2)$ stands for the probability that A_2 does not happen. Determine $P(A_1)$ and $P(A_2)$.

Hence show that A_1 and A_2 are independent.

Ans. $P(A_1) = \frac{2}{3}$, $P(A_2) = \frac{1}{2}$

(b) A and B are events such that

$$P(A \cup B) = \frac{3}{4},\ P(A \cap B) = \frac{1}{4},\ \text{and}\ P(\bar{A}) = \frac{2}{3}.$$

Find *(i)* $P(A)$, *(ii)* $P(B)$ and *(iii)* $P(A \cap \bar{B})$.

(Madras Univ. B.E., 1989)

Ans. *(i)* 1/3, *(ii)* 2/3 *(iii)* 1/12.

46. A thief has a bunch of n keys, exactly one of which fits a lock. If the thief tries to open the lock by trying the keys at random, what is the probability that he requires exactly k attempts, if he rejects the keys already tried? Find the same probability if he does not reject the keys already tried.

(Aligarh Univ. B.Sc., 1991)

Ans. (*i*) $\frac{1}{n}$, (*ii*) $\left(\frac{n-1}{n}\right)^{k-1} \cdot \frac{1}{n}$

(b) There are M urns numbered 1 to M and M balls numbered 1 to M. The balls are inserted randomly in the urns with one ball in each urn. If a ball is put into the urn bearing the same number as the ball, a match is said to have occurred. Find the probability that no match has occurred. **[Civil Services (Main), 1984]**

Hint. See Example 4·54 page 4·97.

47. If n letters are placed at random in n correctly addressed envelopes, find the probability that

(i) none of the letters is placed in the correct envelope,

(ii) At least one letter goes to the correct envelope,

(iii) All letters go to the correct envelopes.

[Delhi Univ. B.Sc. (Stat Hons.), 1987, 1984]

48. An urn contains n white and m black balls, a second urn contains N white and M black balls. A ball is randomly transferred from the first to the second urn and then from the second to the first urn. If a ball is now selected randomly from the first urn, prove that the probability that it is white is

$$\frac{n}{n+m} + \frac{mN - nM}{(n+m)^2 (N+M+1)}$$

[Delhi Univ. B.Sc. (Stat.Hons.) 1986]

Hint. Let us define the following events :

B_i : Drawing of a black ball from the ith urn, $i = 1, 2$.

W_i : Drawing of a white ball from the ith urn , $i = 1, 2$.

The four distinct possibilities for the first two exchanges are $B_1 W_2$, $B_1 B_2$, $W_1 B_2$, $W_1 W_2$. Hence if E denotes the event of drawing a white ball from the first urn after the exchanges, then

$$P(E) = P(B_1 W_2 E) + P(B_1 B_2 E) + P(W_1 B_2 E) + P(W_1 W_2 E) \quad ...(*)$$

We have :

$$P(B_1 W_2 E) = P(B_1) \cdot P(W_2 \mid B_1) P(E \mid B_1 W_2) = \frac{m}{m+n} \times \frac{N}{M+N+1} \times \frac{n+1}{m+n}$$

$$P(B_1 B_2 E) = P(B_1) \cdot P(B_2 \mid B_1) \cdot P(E \mid B_1 B_2) = \frac{m}{m+n} \times \frac{M+1}{M+N+1} \times \frac{n}{m+n}$$

$$P(W_1 B_2 E) = P(W_1) \cdot P(B_2 \mid W_1) \cdot P(E \mid W_1 B_2) = \frac{n}{m+n} \times \frac{M}{M+N+1} \times \frac{n-1}{m+n}$$

$$P(W_1 W_2 E) = P(W_1) \cdot P(W_2 \mid W_1) \cdot P(E \mid W_1 W_2) = \frac{n}{m+n} \times \frac{N+1}{M+N+1} \times \frac{n}{m+n}$$

Substituting in (*) and simplifying we get the result.

49. A particular machine is prone to three similar types of faults A_1, A_2 and A_3. Past records on breakdowns of the machine show the following : the probability of a breakdown (*i.e.*, at least one fault) is 0·1; for each i, the probability that fault A_i occurs and the others do not is 0·02 ; for each pair i, j the probability that A_i and A_j occur but the third fault does not is 0·012. Determine the probabilities of

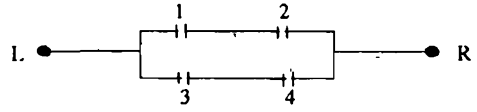
(a) the fault of type A_1 occurring irrespective of whether the other faults occur or not,

(b) a fault of type A_1 given that A_2 has occurred,

(c) faults of type A_1 and A_2 given that A_3 has occurred.

[London U. B.Sc. 1976]

50. The probability of the closing of each relay of the circuit shown below is given by p. If all the relays function independently, what is the probability that a circuit exists between the terminals L and R?



Ans. $p^2(2-p^2)$.

4·9. Bayes Theorem. *If $E_1, E_2, ..., E_n$ are mutually disjoint events with $P(E_i) \neq 0, (i = 1, 2, ..., n)$ then for any orbitrary event A which is a subset of $\bigcup_{i=1}^{n} E_i$ such that $P(A) > 0$, we have*

$$P(E_i \mid A) = \frac{P(E_i)\, P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)}, \quad i = 1, 2, ..., n. \qquad ...(4·12)$$

Proof. Since $A \subset \bigcup_{i=1}^{n} E_i$, we have

$$A = A \cap \left(\bigcup_{i=1}^{n} E_i\right) = \bigcup_{i=1}^{n} (A \cap E_i) \qquad \text{[By distributive law]}$$

Since $(A \cap E_i) \subset E_i$, $(i = 1, 2, ..., n)$ are mutually disjoint events, we have by addition theorem of probability (or Axiom 3 of probability)

$$P(A) = P\left[\bigcup_{i=1}^{n} (A \cap E_i)\right] = \sum_{i=1}^{n} P(A \cap E_i) = \sum_{i=1}^{n} P(E_i)\, P(A \mid E_i), \qquad ...(*)$$

by compound theorem of probability.

Also we have

$$P(A \cap E_i) = P(A)\, P(E_i \mid A)$$

$$P(E_i \mid A) = \frac{P(A \cap E_i)}{P(A)} = \frac{P(E_i)\, P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)} \qquad \text{[From (*)]}$$

Remarks. 1. The probabilities $P(E_1), P(E_2), ..., P(E_n)$ are termed as the *'a priori probabilities'* because they exist before we gain any information from the experiment itself.

2. The probabilities $P(A \mid E_i), i = 1, 2, ..., n$ are called *'likelihoods'* because they indicate how likely the event A under consideration is to occur, given each and every *a priori* probability.

3. The probabilities $P(E_i \mid A), i = 1, 2, ..., n$ are called *'posterior probabilities'* because they are determined after the results of the experiment are known.

4. From (*) we get the following important result:

"If the events $E_1, E_2, ..., E_n$ constitute a partition of the sample space S and $P(E_i) \neq 0, i = 1, 2, ..., n$, then for any event A in S we have

$$P(A) = \sum_{i=1}^{n} P(A \cap E_i) = \sum_{i=1}^{n} P(E_i)\, P(A \mid E_i) \qquad \ldots(4\cdot12\,a)$$

Cor. *(Bayes theorem for future events)*

The probability of the materialisation of another event C, given $P(C \mid A \cap E_1), P(C \mid A \cap E_2), \ldots, P(C \mid A \cap E_n)$ *is*

$$P(C \mid A) = \frac{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)\, P(C \mid E_i \cap A)}{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)} \qquad \ldots(4\cdot12\,b)$$

Proof. Since the occurrence of event A implies the occurrence of one and only one of the events $E_1, E_2, \ldots, E_n$, the event C (granted that A has occurred) can occu in the following mutually exclusive ways:

$$C \cap E_1, C \cap E_2, \ldots, C \cap E_n$$

$$i.e., \quad C = (C \cap E_1) \cup (C \cap E_2) \cup \ldots \cup (C \cap E_n)$$

$$\Rightarrow \quad C \mid A = [(C \cap E_1) \mid A] \cup [(C \cap E_2) \mid A] \cup \ldots \cup [(C \cap E_n) \mid A]$$

$$\therefore \quad P(C \mid A) = P[(C \cap E_1) \mid A] + P[(C \cap E_2) \mid A] + \ldots + P[(C \cap E_n) \mid A]$$

$$= \sum_{i=1}^{n} P[(C \cap E_i) \mid A]$$

$$= \sum_{i=1}^{n} P(E_i \mid A)\; P[C \mid (E_i \cap A)]$$

Substituting the value of $P(E_i \mid A)$ from (*), we get

$$P(C \mid A) = \frac{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)\; P(C \mid E_i \cap A)}{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)}$$

Remark. It may happen that the materialisation of the event E_i makes C independent of A, then we have

$$P(C \mid E_i \cap A) = P(C \mid E_i),$$

and the above formula reduces to

$$P(C \mid A) = \frac{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)\; P(C \mid E_i)}{\sum_{i=1}^{n} P(E_i)\, P(A \mid E_i)} \qquad \ldots(4\cdot12\,c)$$

The event C can be considered in regard to A as Future Event.

Example 4·30. *In 1989 there were three candidates for the position of principal – Mr. Chatterji, Mr. Ayangar and Dr. Singh – whose chances of getting the appointment are in the proportion 4:2:3 respectively. The probability that Mr. Chatterji if selected would introduce co-education in the college is 0·3. The probabilities of Mr. Ayangar and Dr. Singh doing the same are respectively 0·5 and 0·8. What is the probability that there was co-education in the college in 1990?*

[Delhi Univ. B.Sc.(Stat. Hons.), 1992; Gorakhpur Univ. B.Sc., 1992]

Solution. Let the events and probabilities be defined as follows:

A : Introduction of co-education

E_1 : Mr. Chatterji is selected as principal

E_2 : Mr. Ayangar is selected as principal

E_3 : Dr. Singh is selected as principal.

Then

$$P(E_1) = \frac{4}{9}, \quad P(E_2) = \frac{2}{9} \quad \text{and} \quad P(E_3) = \frac{3}{9}$$

$$P(A \mid E_1) = \frac{3}{10}, \quad P(A \mid E_2) = \frac{5}{10} \quad \text{and} \quad P(A \mid E_3) = \frac{8}{10}$$

$$\therefore \quad P(A) = P[(A \cap E_1) \cup (A \cap E_2) \cup (A \cap E_3)]$$

$$= P(A \cap E_1) + P(A \cap E_2) + P(A \cap E_3)$$

$$= P(E_1) P(A \mid E_1) + P(E_2) P(A \mid E_2) + P(E_3) P(A \mid E_3)$$

$$= \frac{4}{9} \cdot \frac{3}{10} + \frac{2}{9} \cdot \frac{5}{10} + \frac{3}{9} \cdot \frac{8}{10} = \frac{23}{45}$$

Example. 4·31. *The contents of urns I, II and III are as follows:*

1 white, 2 black and 3 red balls,

2 white, 1 black and 1 red balls, and

4 white, 5 black and 3 red balls.

One urn is chosen at random and two balls drawn. They happen to be white and red. What is the probability that they come from urns I, II or III ?

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

Solution. Let E_1, E_2, and E_3 denote the events that the urn I, II and III is chosen, respectively, and let A be the event that the two balls taken from the selected urn are white and red. Then

$$P(E_1) = P(E_2) = P(E_3) = \frac{1}{3}$$

$$P(A \mid E_1) = \frac{1 \times 3}{{}^6C_2} = \frac{1}{5}, \quad P(A \mid E_2) = \frac{2 \times 1}{{}^4C_2} = \frac{1}{3},$$

and $$P(A \mid E_3) = \frac{4 \times 3}{{}^{12}C_2} = \frac{2}{11}$$

Hence

$$P(E_2 \mid A) = \frac{P(E_2)\, P(A \mid E_2)}{\sum_{i=1}^{3} P(E_i)\, P(A \mid E_i)}$$

$$= \frac{\frac{1}{3} \times \frac{1}{3}}{\frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{2}{11}} = \frac{55}{118}$$

Similarly

$$P(E_3 \mid A) = \frac{\frac{1}{3} \times \frac{2}{11}}{\frac{1}{3} \times \frac{1}{5} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{11}} = \frac{30}{118}$$

$$\therefore \quad P(E_1 \mid A) = 1 - \frac{55}{118} - \frac{30}{118} = \frac{33}{118}$$

Example. 4·32. *In answering a question on a multiple choice test a student either knows the answer or he guesses. Let p be the probability that he knows the answer and 1–p the probability that he guesses. Assume that a student who guesses at the answer will be correct with probability 1/5, where 5 is the number of multiple-choice alternatives. What is the conditional probability that a student knew the answer to a question given that he answered it correctly?*

[Delhi Univ. B.Sc. (Maths Hons.), 1985]

Solution. Let us define the following events:

E_1 : The student knew the right answer.

E_2 : The student guesses the right answer.

A : The student gets the right answer.

Then we are given

$$P(E_1) = p, \quad P(E_2) = 1 - p, \quad P(A \mid E_2) = 1/5$$

$P(A \mid E_1) = P$ [student gets the right answer given that he knew the right answer] = 1

We want $P(E_1 \mid A)$.

Using Bayes' rule, we get :

$$P(E_1 \mid A) = \frac{P(E_1)\, P(A \mid E_1)}{P(E_1)\, P(A \mid E_1) + P(E_2)\, P(A \mid E_2)} = \frac{p \times 1}{p \times 1 + (1-p) \times \frac{1}{5}} = \frac{5p}{4p+1}$$

Example 4·33. *In a bolt factory machines A, B and C manufacture respectively 25%, 35% and 40% of the total. Of their output 5, 4, 2 per cent are defective bolts. A bolt is drawn at random from the product and is found to be defective. What are the probabilities that it was manufactured by machines A, B and C?*

Solution. Let E_1, E_2 and E_3 denote the events that a bolt selected at random is manufactured by the machines A, B and C respectively and let E denote the event of its being defective. Then we have

$$P(E_1) = 0{\cdot}25,\ P(E_2) = 0{\cdot}35,\ P(E_3) = 0{\cdot}40$$

The probability of drawing a defective bolt manufactured by machine A is $P(E \mid E_1) = 0{\cdot}05$.

Similarly, we have

$$P(E \mid E_2) = 0{\cdot}04, \text{ and } P(E \mid E_3) = 0{\cdot}02$$

Hence the probability that a defective bolt selected at random is manufactured by machine A is given by

$$P(E_1 \mid E) = \frac{P(E_1)\,P(E \mid E_1)}{\sum_{i=1}^{3} P(E_i)\,P(E \mid E_i)}$$

$$= \frac{0{\cdot}25 \times 0{\cdot}05}{0{\cdot}25 \times 0{\cdot}05 + 0{\cdot}35 \times 0{\cdot}04 + 0{\cdot}40 \times 0{\cdot}02} = \frac{125}{345} = \frac{25}{69}$$

Similarly

$$P(E_2 \mid E) = \frac{0{\cdot}35 \times 0{\cdot}04}{0{\cdot}25 \times 0{\cdot}05 + 0{\cdot}35 \times 0{\cdot}04 + 0{\cdot}40 \times 0{\cdot}02} = \frac{140}{345} = \frac{28}{69}$$

and

$$P(E_3 \mid E) = 1 - [P(E_1 \mid E) + P(E_2 \mid E)] = 1 - \frac{25}{69} - \frac{28}{69} = \frac{16}{69}$$

This example illustrates one of the chief applications of Bayes Theorem.

EXERCISE 4 (d)

1. *(a)* State and prove Baye's Theorem.

(b) The set of events A_k, $(k = 1, 2, ..., n)$ are *(i)* exhaustive and *(ii)* pairwise mutually exclusive. If for all k the probabilities $P(A_k)$ and $P(E \mid A_k)$ are known, calculate $P(A_k \mid E)$, where E is an arbitrary event. Indicate where conditions *(i)* and *(ii)* are used.

(c) The events $E_1, E_2, ..., E_n$ are mutually exclusive and $E = E_1 \cup E_2 \cup ... \cup E_n$. Show that if $P(A \mid E_i) = P(B \mid E_i)$; $i = 1, 2, ..., n$, then $P(A \mid E) = P(B \mid E)$. Is this conclusion true if the events E_i are not mutually exclusive?

[Calcutta Univ. B.Sc. (Maths Hons.), 1990]

(d) What are the criticisms against the use of Bayes theorem in probability theory. **[Sri. Venketeswara Univ. B.Sc., 1991]**

(e) Using the fundamental addition and multiplication rules of probability, show that

$$P(B \mid A) = \frac{P(B)\,P(A \mid B)}{P(B)\,P(A \mid B) + P(\bar{B})\,P(A \mid \bar{B})}$$

where $\bar{B}$ is the event complementary to the event B.

[Delhi Univ. M.A. (Econ.), 1987]

2. *(a)* Two groups are competing for the positions on the Board of Directors of a corporation. The probabilities that the first and second groups will win are 0·6 and 0·4 respectively. Furthermore, if the first group wins the probability of introducing a new product is 0·8 and the corresponding probability if the second group wins is 0·3. What is the probability that the new product will be introduced?

Ans. $0{\cdot}6 \times 0{\cdot}8 + 0{\cdot}4 \times 0{\cdot}3 = 0{\cdot}6$

(b) The chances of X, Y, Z becoming managers of a certain company are 4:2:3. The probabilities that bonus scheme will be introduced if X, Y, Z become managers, are 0·3, 0·5 and 0·8 respectively. If the bonus scheme has been introduced, what is the probability that X is appointed as the manager.

Ans. 0·51

(c) A restaurant serves two special dishes, A and B to its customers consisting of 60% men and 40% women. 80% of men order dish A and the rest B. 70% of women order dish B and the rest A. In what ratio of A to B should the restaurant prepare the two dishes? **(Bangalore Univ. B.Sc., 1991)**

Ans. $P(A) = P[(A \cap M) \cup (A \cap W)] = 0{\cdot}6 \times 0{\cdot}8 + 0{\cdot}4 \times 0{\cdot}3 = 0{\cdot}6$

Similarly $P(B) = 0{\cdot}4$. Required ratio $= 0{\cdot}6 : 0{\cdot}4 = 3 : 2$.

3. *(a)* There are three urns having the following compositions of black and white balls.

Urn 1 : 7 white, 3 black balls
Urn 2 : 4 white, 6 black balls
Urn 3 : 2 white, 8 black balls.

One of these urns is chosen at random with probabilities 0·20, 0·60 and 0·20 respectively. From the chosen urn two balls are drawn at random without replacement. Calculate the probability that both these balls are white.

Ans. 8/45. **(Madurai Univ. B.Sc., 1991)**

(b) Bowl I contain 3 red chips and 7 blue chips, bowl II contain 6 red chips and 4 blue chips. A bowl is selected at random and then 1 chip is drawn from this bowl. *(i)* Compute the probability that this chip is red, *(ii)* Relative to the hypothesis that the chip is red, find the conditional probability that it is drawn from bowl II.

[Delhi Univ. B.Sc. (Maths Hons.)1987]

(c) In a factory machines A and B are producing springs of the same type. Of this production, machines A and B produce 5% and 10% defective springs, respectively. Machines A and B produce 40% and 60% of the total output of the factory. One spring is selected at random and it is found to be defective. What is the possibility that this defective spring was produced by machine A ?

[Delhi Univ. M.A. (Econ.),1986]

(d) Urn A contains 2 white, 1 black and 3 red balls, urn B contains 3 white, 2 black and 4 red balls and urn C contains 4 white, 3 black and 2 red balls. One urn is chosen at random and 2 balls are drawn. They happen to be red and black. What

is the probability that both balls came from urn 'B' ?

[Madras U. B.Sc. April; 1989]

(e) Urn X_1, X_2, X_3, each contains 5 red and 3 white balls. Urns Y_1, Y_2, each contain 2 red and 4 white balls. An urn is selected at random and a ball is drawn. It is found to be red. Find the probability that the ball comes out of the urns of the first type. **[Bombay U. B.Sc., April 1992]**

(f) Two shipments of parts are received. The first shipment contains 1000 parts with 10% defectives and the second shipment contains 2000 parts with 5% defectives. One shipment is selected at random. Two parts are tested and found good. Find the probability (*a posterior*) that the tested parts were selected from the first shipment. **[Burdwan Univ. B.Sc. (Hons.), 1988]**

(g) There are three machines producing 10,000 ; 20,000 and 30,000 bullets per hour respectively. These machines are known to produce 5%, 4% and 2% defective bullets respectively. One bullet is taken at random from an hour's production of the three machines. What is the probability that it is defective? If the drawn bullet is defective, what is the probability that this was produced by the second machine? **[Delhi Univ. B.Sc. (Stat. Hons.), 1991]**

4. *(a)* Three urns are given each containing red and white chips as indicated.

Urn 1 : 6 red and 4 white.
Urn 2 : 2 red and 6 white.
Urn 3 : 1 red and 8 white.

(i) An urn is chosen at random and a ball is drawn from this urn. The ball is red. Find the probability that the urn chosen was urn I .

(ii) An urn is chosen at random and two balls are drawn without replacement from this urn. If both balls are red, find the probability that urn I was chosen. Under these conditions, what is the probability that urn III was chosen.

Ans. 108/173, 112/12, 0 **[Gauhati Univ. B.Sc., 1990]**

(b) There are ten urns of which each of three contains 1 white and 9 black balls, each of other three contains 9 white and 1 black ball, and of the remaining four each contains 5 white and 5 black balls. One of the urns is selected at random and a ball taken blindly from it turns out to be white. What is the probability that an urn containing 1 white and 9 black balls was selected? **(Agra Univ. B.Sc., 1991)**

Hint : $P(E_1) = \frac{3}{10}$, $P(E_2) = \frac{3}{10}$ and $P(E_3) = \frac{4}{10}$.

Let A be the event of drawing a white ball.

$$P(A) = \frac{3}{10} \times \frac{1}{10} + \frac{3}{10} \times \frac{9}{10} + \frac{4}{10} \times \frac{5}{10} = \frac{1}{2}$$

$$P(A \mid E_1) = \frac{1}{10} \text{ and } P(E_1 \mid A) = \frac{3}{50}$$

(c) It is known that an urn containing altogether 10 balls was filled in the following manner: A coin was tossed 10 times, and according as it showed heads or tails, one white or one black ball was put into the urn. Balls are drawn from this

urn one at a time, 10 times in succession (with replacement) and every one turns out to be white. Find the chance that the urn contains nothing but white balls.

Ans. 0·0702.

5. (*a*) From a vessel containing 3 white and 5 black balls, 4 balls are transferred into an empty vessel. From this vessel a ball is drawn and is found to be white. What is the probability that out of four balls transferred, 3 are white and 1 black. **[Delhi Uni. B.Sc. (Stat. Hons.), 1985]**

Hint. Let the five mutually exclusive events for the four balls transferred be E_0, E_1, E_2, E_3, and E_4, where E_i denotes the event that i white balls are transferred and let A be the event of drawing a white ball from the new vessel.

Then $P(E_0) = \frac{{}^5C_4}{{}^8C_4}$, $P(E_1) = \frac{{}^3C_1 \times {}^5C_3}{{}^8C_4}$, $P(E_2) = \frac{{}^3C_2 \times {}^5C_2}{{}^8C_4}$

$P(E_3) = \frac{{}^3C_3 \times {}^5C_1}{{}^8C_4}$ and $P(E_4) = 0$

Also $P(A \mid E_0) = 0$, $P(A \mid E_1) = \frac{1}{4}$, $P(A \mid E_2) = \frac{2}{4}$, $(A \mid E_3) = \frac{3}{4}$,

and $P(A \mid E_4) = 1$. Hence $P(E_3 \mid A) = \frac{1}{7}$.

(*b*) The contents of the urns 1 and 2 are as follows :

Urn 1 : 4 white and 5 black balls.

Urn 2 : 3 white and 6 black balls.

One urn is chosen at random and a ball is drawn and its colour noted and replaced back to the urn. Again a ball is drawn from the same urn, colour noted and replaced. The process is repeated 4 times and as a result one ball of white colour and three balls of black colour are obtained. What is the probability that the urn chosen was the urn 1 ? **(Poona Univ. B.E., 1989)**

Hint. $P(E_1) = P(E_2) = 1/2$,

$P(A \mid E_1) = 4/9$, $\quad 1 - P(A \mid E_1) = 5/9$

$P(A \mid E_2) = 1/3$, $\quad 1 - P(A \mid E_2) = 2/3$

The probability that the urn chosen was the urn 1

$$= \frac{\frac{1}{2} \cdot \frac{4}{9} \left(\frac{5}{9}\right)^3}{\frac{1}{2} \cdot \frac{4}{9} \cdot \left(\frac{5}{9}\right)^3 + \frac{1}{2} \cdot \frac{1}{3} \cdot \left(\frac{2}{3}\right)^3}$$

(*c*) There are five urns numbered 1 to 5. Each urn contains 10 balls. The ith urn has i defective balls and $10 - i$ non-defective balls; $i = 1, 2, \ldots 5$. An urn is chosen at random and then a ball is selected at random from that urn. (*i*) What is the probability that a defective ball is selected ?

(*ii*) If the selected ball is defective, find the probability that it came from urn i, ($i = 1, 2, \ldots, 5$). **[Delhi Univ. B.Sc. (Maths Hons.), 1987]**

Hint.: Define the following events :

E_i : ith urn is selected at random.

A : Defective ball is selected.

$P(E_i) = 1/5;\ i = 1, 2, ..., 5.$

$P(A \mid E_i) = P$ [Defective ball from ith urn] $= i/10, (i = 1, 2, ..., 5)$

$$P(E_i) \cdot P(A \mid E_i) = \frac{1}{5} \times \frac{i}{10} = \frac{i}{50}, \quad (i = 1, 2, ..., 5).$$

(i) $$P(A) = \sum_{i=1}^{5} P(E_i)\, P(A \mid E_i) = \sum_{i=1}^{5} \left(\frac{i}{50}\right) = \frac{1+2+3+4+5}{50} = \frac{3}{10}$$

(ii) $$P(E_i \mid A) = \frac{P(E_i)\, P(A \mid E_i)}{\sum_i P(E_i)\, P(A \mid E_i)} = \frac{i/50}{3/10} = \frac{i}{15}\ ;\ i = 1, 2, ..., 5.$$

For example, the probability that the defective ball came from 5th urn $= (5/15) = 1/3$.

6. *(a)* A bag contains six balls of different colours and a ball is drawn from it at random. A speaks truth thrice out of 4 times and B speaks truth 7 times out of 10 times. If both A and B say that a red ball was drawn, find the probability of their joint statement being true.

[Delhi Univ. B.Sc. (Stat. Hons.),1987; Kerala Univ. B.Sc.1988]

(b) A and B are two very weak students of Statistics and their chances of solving a problem correctly are 1/8 and 1/12 respectively. If the probability of their making a common mistake is 1/1001 and they obtain the same answer, find the chance that their answer is correct. **[Poona Univ. B.Sc., 1989]**

Ans. Reqd. Probability $= \dfrac{1/8 \times 1/12}{1/8 \times 1/12 + (1 - 1/8) \cdot (1 - 1/12) \cdot 1/1001} = \dfrac{13}{14}$

7. *(a)* Three boxes, practically indistinguishable in appearance, have two drawers each. Box I contains a gold coin in one and a silver coin in the other drawer, box II contains a gold coin in each drawer and box III contains a silver coin in each drawer. One box is chosen at random and one of its drawers is opened at random and a gold coin found. What is the probability that the other drawer contains a coin of silver? **(Gujarat Univ. B.Sc., 1992)**

Ans. 1/3, 1/3.

(b) Two cannons No. 1 and 2 fire at the same target. Cannon No. 1 gives on an average 9 shots in the time in which Cannon No. 2 fires 10 projectiles. But on an average 8 out of 10 projectiles from Cannon No. 1 and 7 out of 10 from Cannon No. 2 strike the target. In the course of shooting, the target is struck by one projectile. What is the probability of a projectile which has struck the target belonging to Cannon No. 2 ? **(Lucknow Univ. B.Sc., 1991)**

Ans. 0·493

(c) Suppose 5 men out of 100 and 25 women out of 10,000 are colour blind. A colour blind person is chosen at random. What is the probability of his being male? (Assume males and females to be in equal number.)

Hint. E_1 = Person is a male, E_2 = Person is a female.

A = Person is colour blind.

Then $P(E_1) = P(E_2) = \frac{1}{2}$, $P(A \mid E_1) = 0{\cdot}05$, $P(A \mid E_2) = 0{\cdot}0025$.

Hence find $P(E_1 \mid A)$.

8. *(a)* Three machines X, Y, Z with capacities proportional to 2:3:4 are producting bullets. The probabilities that the machines produce defective are 0·1, 0·2 and 0·1 respectively. A bullet is taken from a day's production and found to be defective What is the probability that it came from machine X?

[Madras Univ. B.Sc., 1988]

(b) In a factory 2 machines M_1 and M_2 are used for manufacturing screws which may be uniquely classified as good or bad. M_1 produces per day n_1 boxes of screws, of which on the average, $p_1\%$ are bad while the corresponding numbers for M_2 are n_2 and p_2. From the total production of both M_1 and M_2 for a certain day, a box is chosen at random, a screw taken out of it and it is found to be bad. Find the chance that the selected box is manufactured *(i)* by M_1, *(ii)* M_2.

Ans. *(i)* $n_1 p_1/(n_1 p_1 + n_2 p_2)$, *(ii)* $n_2 p_2/(n_1 p_1 + n_2 p_2)$.

9. *(a)* A man is equally likley to choose any one of three routes A, B, C from his house to the railway station, and his choice of route is not influenced by the weather. If the weather is dry, the probabilities of missing the train by routes A, B, C are respectively 1/20, 1/10, 1/5. He sets out on a dry day and misses the train. What is the probability that the route chosen was C?

On a wet day, the respective probabilities of missing the train by routes A, B, C are 1/20, 1/5, 1/2 respectively. On the average, one day in four is wet. If he misses the train, what is the probability that the day was wet?

[Allahabad Univ. B.Sc., 1991]

(b) A doctor is to visit the patient and from past experience it is known that the probabilities that he will come by train, bus or scooter are respectively 3/10, 1/5, and 1/10, the probabililty that he will use some other means of transport being, therefore, 2/5. If he comes by train, the probability that he will be late is 1/4, if by bus 1/3 and if by scooter 1/12, if he uses some other means of transport it can be assumed that he will not be late. When he arrives he is late. What is the probability that *(i)* he comes by train *(ii)* he is not late?

[Burdwan Univ. B.Sc. (Hons.), 1990]

Ans. *(i)* 1/2, *(ii)* 9/34

10. State and prove Bayes rule and expalin why, in spite of its easy deductibility from the postulates of probability, it has been the subject of such extensive controversy.

In the chest X-ray tests, it is found that the probability of detection when a person has actually T.B. is 0·95 and probabililty of diagnosing incorrectly as having T.B. is 0·002. In a certain city 0·1% of the adult population is suspected to be suffering from T.B. If an adult is selected at random and is diagnosed as having

T.B. on the basis of the X-ray test, what is the probability of his actually having a T.B.? **(Nagpur Univ. B.E., 1991)**

Ans. 0·97

11. A certain transistor is manufactured at three factories at Barnsley, Bradford and Bristol. It is known that the Barnsley factory produces twice as many transistors as the Bradford one, which produces the same number as the Bristol one (during the same period). Experience also shows that 0·2% of the transistors produced at Barnsley and Bradford are faulty and so are 0·4% of those produced at Bristol.

A service engineer, while maintaining an electronic equipment, finds a defective transistor. What is the probability that the Bradford factory is to blame?

(Bangalore Univ. B.E., Oct. 1992)

12. The sample space consists of integers from 1 to $2n$ which are assigned probabilities proportional to their logarithms. Find the probabilities and show that the conditional probability of the integer 2, given that an even integer occurs, is

$$\frac{\log 2}{[\, n \log 2 + \log (n\,!)\,]}$$ **(Lucknow Univ. M.A., 1992)**

[Hint. Let E_i : the event that the integer $2i$ is drawn, $(i = 1, 2, 3, ..., n)$.

A : the event of drawing an even integer.

$$\Rightarrow \qquad A = E_1 \cup E_2 \cup \ldots \cup E_n \qquad \Rightarrow \qquad P(A) = \sum_{i=1}^{n} P(E_i)$$

But $\quad P(E_i) = k \log (2i) \quad$ (Given)

$$\therefore \qquad P(A) = k \sum_{i=1}^{n} \log (2i) = k \log \prod_{i=1}^{n} (2i) = k\,[\, n \log 2 + \log (n\,!)\,]$$

$$\therefore \qquad P(E_i \mid A) = \frac{\log (2i)}{[\, n \log 2 + \log (n\,!)\,]}$$

13. In answering a question on a multiple choice test, an examinee either knows the answer (with probability p), or he guesses (with probability $1 - p$). Assume that the probability of answering a question correctly is unity for an examinee who knows the answer and $1/m$ for the examinee who guesses, where m is the number of multiple choice alternatives. Supposing an examinee answers a question correctly, what is the probability that he really knows the answer?

[Delhi Univ. M.C.A., 1990; M.Sc. (Stat.), 1989]

Hint. Let E_1 = The examinee knows the answer,

E_2 = The examinee guesses the answer,

and A = The examinee answers correctly.

Then $P(E_1) = p$, $P(E_2) = 1 - p$, $P(A \mid E_1) = 1$ and $P(A \mid E_2) = 1/m$

Now use Bayes theorem to prove

$$P(E_1 \mid A) = \frac{mp}{1 + (m-1)\,p}$$

14. Die A has four red and two white faces whereas die B has two red and four white faces. A biased coin is flipped once. If it falls heads, the game continues by

throwing die A, if it falls tails die B is to be used.

(i) Show that the probability of getting a red face at any throw is 1/2.

(ii) If the first two throws resulted in red faces, what is the probability of red face at the 3rd throw?

(iii) If red face turns up at the first n throws, what is the probability that die A is being used?

Ans. *(ii)* 3/5 *(iii)* $\frac{2^n}{2^n+1}$

15. A manufacturing firm produces steel pipes in three plants with daily production volumes of 500, 1,000 and 2,000 units respectively. According to past experience it is known that the fraction of defective outputs produced by the three plants are respectively 0.005, 0.008 and 0.010. If a pipe is selected at random from a day's total production and found to be defective, from which plant does that pipe come?

Ans. Third plant.

16. A piece of mechanism consists of 11 components, 5 of type A, 3 of type B, 2 of type C and 1 of type D. The probability that any particular component will function for a period of 24 hours from the commencement of operations without breaking down is independent of whether or not any other component breaks down during that period and can be obtained from the following table:

Component type: A B C D

Probability: 0·6 0·7 0·3 0·2

(i) Calculate the probability that 2 components chosen at random from the 11 components will both function for a period of 24 hours from the commencement of operations without breaking down.

(ii) If at the end of 24 hours of operations neither of the 2 components chosen in *(i)* has broken down, what is the probability that they are both type C components.

Hint.

(i) Required probability $= \frac{1}{{}^{11}C_2}[{}^5C_2 \times (0{\cdot}6)^2 + {}^3C_2\,(0{\cdot}7)^2 + {}^2C_2\,(0{\cdot}3)^2$

$+ {}^5C_1 \times {}^3C_1 \times 0{\cdot}6 \times 0{\cdot}7 + {}^5C_1 \times {}^2C_1 \times (0{\cdot}6) \times (0{\cdot}3)$

$+ {}^5C_1 \times {}^1C_1 \times (0{\cdot}6) \times (0{\cdot}2) + {}^3C_1 \times {}^2C_1 \times 0{\cdot}7 \times 0{\cdot}3$

$+ {}^3C_1 \times {}^1C_1 \times 0{\cdot}7 \times 0{\cdot}2 + {}^2C_1 \times {}^1C_1 \times 0{\cdot}3 \times 0{\cdot}2]$

$= p$ (Say).

(ii) Required probability (By Bayes theorem)

$$= \frac{{}^2C_2 \times (0{\cdot}3)^2}{p} = \frac{0{\cdot}09}{p}$$

4·10. Geometric probability. In remark 3, § 4·3·1 it was pointed out that the classical definition of probability fails if the total number of outcomes of an experiment is infinite. Thus, for example, if we are interested in finding the

probability that a point selected at random in a given region will lie in a specified part of it, the classical definition of probability is modified and extended to what is called *geometrical probability or probability in continuum*. In this case, the general expression for probability 'p' is given by

$$p = \frac{\textit{Measure of specified part of the region}}{\textit{Measure of the whole region}}$$

where 'measure' refers to the length, area or volume of the region if we are dealing with one, two or three dimensional space respectively.

Example 4·34. *Two points are taken at random on the given straight line of length a. Prove that the probability of their distance exceeding a given length* $c\ (< a)$ *is equal to* $\left(1 - \frac{c}{a}\right)^2$.

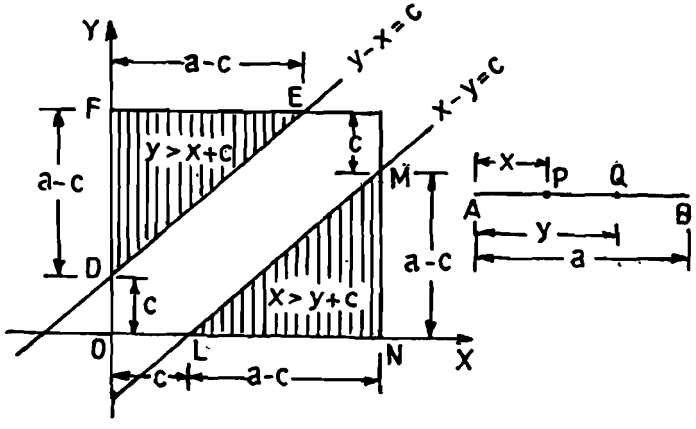
[Burdwan Univ. B.Sc. (Hons.), 1992; Delhi Univ. M.A. (Econ.), 1987]

Solution. Let P and Q be any two points taken at random on the given straight line AB of length 'a'. Let $AP = x$ and $AQ = y$,

$$(0 \le x \le a, 0 \le y \le a).$$

Then we want $P\{|x - y| > c\}$.

The probability can be easily calculated geometrically. Plotting the lines $x - y = c$ and $y - x = c$ along the co-ordinate axes, we get the following diagram:



Since $0 \le x \le a, 0 \le y \le a$, total area $= a \cdot a = a^2$.

Area favourable to the event $|x - y| > c$ is given by

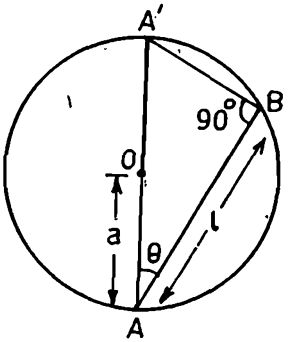
$$\Delta LMN + \Delta DEF = \frac{1}{2} LN \cdot MN + \frac{1}{2} EF \cdot DF$$

$$= \frac{1}{2}(a-c)^2 + \frac{1}{2}(a-c)^2 = (a-c)^2$$

$$P(|x - y| > c) = \frac{(a-c)^2}{a^2} = \left(1 - \frac{c}{a}\right)^2$$

Example 4·35. (Bertrand's Problem). *If a chord is taken at random in a circle, what is the chance that its length l is not less than 'a' , the radius of the circle?*

Solution. Let the chord AB make an angle θ with the diameter AOA' of the circle with centre O and radius $OA=a$. Obviously θ lies betwseen $-\pi/2$ and $\pi/2$. Since all the positions of the chord AB and consequently all the values of θ are equally likely, θ may be regarded as a random variable which is uniformly distributed *c.f.* § 8·1 over $(-\pi/2, \pi/2)$ with probability density function



$$f(\theta) = \frac{1}{\pi} \; ; \; -\pi/2 < \theta \leq \pi/2$$

$\angle ABA'$, being the angle in a semi-circle, is a right angle. From $\Delta ABA'$ we have

$$\frac{AB}{AA'} = \cos\theta$$

$$\Rightarrow \qquad l = 2a\cos\theta$$

The required probability 'p' is given by

$$p = P(l \geq a) = P(2a\cos\theta \geq a)$$
$$= P(\cos\theta \geq 1/2) = P(|\theta| \leq \pi/3)$$
$$= \int_{-\pi/3}^{\pi/3} f(\theta)\, d\theta = \frac{1}{\pi}\int_{-\pi/3}^{\pi/3} d\theta = \frac{2}{3}$$

Example 4·36. *A rod of length 'a' is broken into three parts at random. What is the probability that a triangle can be formed from these parts?*

Solution. Let the lengths of the three parts of the rod be x, y and $a-(x+y)$. Obviously, we have

$$x>0\,;\; y>0 \text{ and } x+y<a \quad \Rightarrow \quad y<a-x \qquad \ldots(*)$$

In order that these three parts form the sides of a triangle, we should have

$$\left.\begin{aligned} x+y &> a-(x+y) \quad \Rightarrow \quad y>\frac{a}{2}-x \\ \text{and} \qquad x+a-(x+y) &> y \quad \Rightarrow \quad y<\frac{a}{2} \\ y+a-(x+y) &> x \quad \Rightarrow \quad y<\frac{a}{2} \end{aligned}\right\} \qquad \ldots(**)$$

since in a triangle, the sum of any two sides is greater than the third. Equivalently, (**) can be written as

$$\frac{a}{2}-x<y<\frac{a}{2} \quad \wedge \quad 0<x<\frac{a}{2} \qquad \ldots(***)$$

Hence, on using (*) and (***), the required probability is given by

$$\frac{\int_0^{a/2}\int_{(a/2)-x}^{a/2} dy\,dx}{\int_0^{a}\int_{0}^{a-x} dy\,dx} = \frac{\int_0^{a/2}\left[\frac{a}{2}-\left(\frac{a}{2}-x\right)\right]dx}{\int_0^{a}(a-x)\,dx}$$

$$= \frac{\left|\frac{x^2}{2}\right|_0^{a/2}}{\left|\frac{-(a-x)^2}{2}\right|_0^{a}} = \frac{a^2/8}{a^2/2} = \frac{1}{4}$$

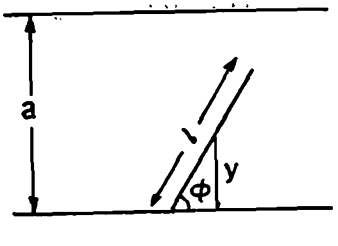
Example 4·37. (Buffon's Needle Problem). *A vertical board is ruled with horizontal parallel lines at constant distance 'a' apart. A needle of length l ($< a$) is thrown at random on the table. Find the probability that it will intersect one of the lines.*

Solution. Let y denote the distance from the centre of the needle to the nearest parallel and ϕ be angle formed by the needle with this parallel. The quantities y and ϕ fully determine the position of the needle. Obviously y ranges from 0 to $a/2$ (since $l < a$) and ϕ from 0 to π.

Since the needle is dropped randomly, all possible values of y and ϕ may be regarded as equally likely and consequently the joint probability density function $f(y, \phi)$ of y and ϕ is given by the uniform distribution (c.f. § 8.1) by

$$f(y, \phi) = k; \quad 0 \le \phi \le \pi, \quad 0 \le y \le a/2, \quad \ldots(*)$$

where k is a constant.



The needle will intersect one of the lines if the distance of its centre from the line is less than $\frac{1}{2} l \sin\phi$, *i.e.*, the required event can be represented by the inequality $0 < y < \frac{1}{2} l \sin\phi$. Hence the required probability p is given by

$$p = \frac{\int_0^{\pi}\int_0^{(l\sin\phi)/2} f(y, \phi)\,dy\,d\phi}{\int_0^{\pi}\int_0^{a/2} f(y, \phi)\,dy\,d\phi}$$

$$= \frac{\frac{l}{2}\int_0^{\pi} \sin\phi\,d\phi}{(a/2)\cdot\pi}$$

$$= \frac{l}{a\pi}\left|-\cos\phi\right|_0^{\pi} = \frac{2l}{a\pi}$$

EXERCISE 4 (e)

1. Two points are selected at random in a line AC of length 'a' so as to lie on the opposite sides of its mid-point O. Find the probability that the distance between them is less than $a/3$.

2. *(a)* Two points are selected at random on a line of length a. What is the probability that none of three sections in which the line is thus divided is less than $a/4$?

Ans. 1/16.

(b) A rectilinear segment AB is divided by a point C into two parts $AC=a$, $CB=b$. Points X and Y are taken at random on AC and CB respectively. What is the probability that AX, XY and BY can form a triangle?

(c) ABG is a straight line such that AB is 6 inches and BG is 5 inches. A point Y is chosen at random on the BG part of the line. If C lies between B and G in such a way that $AC=t$ inches, find

(i) the probability that Y will lie in BC.

(ii) the probability that Y will lie in CG.

What can you say about the sum of these probabilities?

(d) The sides of a rectangle are taken at random each less than a and all lengths are equally likely. Find the chance that the diagonal is less than a.

3. *(a)* Three points are taken at random on the circumference of a circle. Find the chance that they lie on the same semi- circle.

(b) A chord is drawn at random in a given circle. What is the probability that it is greater than the side of an equilateral triangle inscribed in that circle?

(c) Show that the probability of choosing two points randomly from a line segment of length 2 inches and their being at a distance of at least 1 inch from each other is 1/4. **[Delhi Univ. M.A. (Econ.), 1985]**

4. A point is selected at random inside a circle. Find the probability that the point is closer to the centre of the circle than to its circumference.

5. One takes at random two points P and Q on a segment AB of length a

(i) What is the probability for the distance PQ being less than b ($<a$)?

(ii) Find the chance that the distance between them is greater than a given length b.

6. Two persons A and B, make an appointment to meet on a certain day at a certain place, but without fixing the time further than that it is to be between 2 p.m. and 3 p.m and that each is to wait not longer than ten minutes for the other. Assuming that each is independently equally likely to arrive at any time during the hour, find the probability that they meet.

Third person C, is to be at the same place from 2·10 p.m. until 2·40 p.m. on the same day. Find the probabilities of C being present when A and B are there together *(i)* When A and B remain after they meet, *(ii)* When A and B leave as soon as they meet.

Hint. Denote the times of arrival of A by x and of B by y. For the meeting to take place it is necessary and sufficient that

$$| x - y | < 10$$

We depict x and y as Cartesian coordinates in the plane; for the scale unit we take one minute. All possible outcomes can be described as points of a aquare with side 60. We shall finally get [*c.f.* Example 4·34, with $a = 60$, $c = 10$]

$$P[|x - y| < 10] = 1 - (5/6)^2 = 11/36$$

7. The outcome of an experiment are represented by points in the square bounded by $x = 0$, $x = 2$ and $y = 2$ in the xy-plane. If the probability is distributed uniformly, determine the probability that $x^2 + y^2 > 1$

Hint.

Required probability $P(E) = \int_E \frac{1}{4}\, dx\, dy = 1 - \int_{E'} \frac{1}{4}\, dx\, dy$

where E is the region for which $x^2 + y^2 > 1$ and E' is the region for which $x^2 + y^2 \leq 1$.

$$\therefore \quad 4P(E) = 4 - \int_0^1 \int_0^1 dx\, dy = 3 \qquad \Rightarrow \qquad P(E) = \frac{3}{4}$$

8. A floor is paved with tiles, each tile being a parallelogram such that the distance between pairs of opposite sides are a and b respectively, the length of the diagonal being l. A stick of length c falls on the floor parallel to the diagonal. Show that the probability that it will lie entirely on one tile is

$$\left(1 - \frac{c}{l}\right)^2$$

If a circle of diameter d is thrown on the floor, show that the probability that it will lie on one tile is

$$\left(1 - \frac{d}{a}\right)\left(1 - \frac{d}{b}\right)$$

9. Circular discs of radius r are thrown at random on to a plane circular table of radius R which is surrounded by a border of uniform width r lying in the same plane as the table. If the discs are thrown independently and at random, and N stay on the table, show that the probability that a fixed point on the table but not on the border, will be covered is

$$1 - \left(1 - \frac{r^2}{(R + r)^2}\right)^N$$

SOME MISCELLANEOUS EXAMPLES

Example 4·38. *A die is loaded in such a manner that for $n = 1, 2, 3, 4, 5, 6$ the probability of the face marked n, landing on top when the die is rolled is proportional to n. Find the probability that an odd number will appear on tossing the die.*

[Madras Univ. B.Sc. (Stat. Main), 1987]

Solution. Here we are given

$P(n) \propto n$ or $P(n) = kn$, where k is the constant of proportionality.

Also $P(1) + P(2) + ...P(6) = 1 \Rightarrow k(1+2+3+4+5+6) = 1$ or $k = 1/21$

Required Probability $= P(1) + P(3) + P(5) = \dfrac{1+3+5}{21} = \dfrac{3}{7}$

Example 4·39. *In terms of probability :*

$p_1 = P(A)$, $p_2 = P(B)$, $p_3 = P(A \cap B)$, $(p_1, p_2, p_3 > 0)$

Express the following in terms of p_1, p_2, p_3

(a) $P(\overline{A \cup B})$, *(b)* $P(\bar{A} \cup \bar{B})$, *(c)* $P(\bar{A} \cap B)$, *(d)* $P(\bar{A} \cup B)$, *(e)* $P(\bar{A} \cap \bar{B})$

(f) $P(A \cap \bar{B})$, *(g)* $P(A \mid B)$, *(h)* $P(B \mid \bar{A})$, *(i)* $P[\bar{A} \cap (A \cup B)]$

Solution.

(a) $P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - [P(A) + P(B) - P(AB)]$.

$= 1 - p_1 - p_2 + p_3$.

(b) $P(\bar{A} \cup \bar{B}) = P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - p_3$

(c) $P(\bar{A} \cap B) = P(B - AB) = P(B) - P(A \cap B) = p_2 - p_3$

(d) $P(\bar{A} \cup B) = P(\bar{A}) + P(B) - P(\bar{A} \cap B) = 1 - p_1 + p_2 - (p_2 - p_3)$

$= 1 - p_1 + p_3$

(e) $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - p_1 - p_2 + p_3$. [Part *(a)*]

(f) $P(A \cap \bar{B}) = P(A - A \cap B) = P(A) - P(A \cap B) = p_1 - p_3$

(g) $P(A \mid B) = P(A \cap B)/P(B) = p_3/p_2$

(h) $P(B \mid \bar{A}) = P(\bar{A} \cap B)/P(\bar{A}) = (p_2 - p_3)/(1 - p_1)$

(i) $P[\bar{A} \cap (A \cup B)] = P[(\bar{A} \cap A) \cup (\bar{A} \cap B)]$

$= P(\bar{A} \cap B) = p_2 - p_3$ $[\because A \cap \bar{A} = \phi]$

Example 4·40. *Let $P(A) = p$, $P(A \mid B) = q$, $P(B \mid A) = r$. Find relations between the numbers p, q, r for the following cases :*

(a) Events A and B are mutually exclusive.

(b) A and B are mutually exclusive and collectively exhaustive.

(c) A is a subevent of B; B is a subevent of A.

(d) $\bar{A}$ and $\bar{B}$ are mutually exclusive.

[Delhi Univ. B.Sc. (Maths Hons.) 1985]

Solution. From given data : $P(A) = p$, $P(A \cap B) = P(A)\,P(B \mid A) = rp$

$$\therefore \quad P(B) = \frac{P(A \cap B)}{P(A \mid B)} = \frac{rp}{q}$$

(a) $P(A \cap B) = 0 \quad \Rightarrow \quad rp = 0.$

(b) $P(A \cap B) = 0$ and $P(A) + P(B) = 1$

$\Rightarrow p(q + r) = q$; $rp = 0 \quad \Rightarrow \quad pq = q \quad \Rightarrow \quad p = 1 \vee q = 0.$

(c) $A \subseteq B \Rightarrow A \cap B = A$ or $P(A \cap B) = P(A) \Rightarrow rp = p \Rightarrow r = 1 \vee p = 0.$

$B \subseteq A \Rightarrow A \cap B = B$ or $P(A \cap B) = P(B)$

$\Rightarrow \quad rp = (rp/q)$ or $rp(q - 1) = 0 \quad \Rightarrow \quad q = 1$

(d) $P(\bar{A} \cap \bar{B}) = 1 - P(A \cup B) \quad \Rightarrow \quad 0 = 1 - [P(A) + P(B) - P(A \cap B)]$

So $P(A) + P(B) = 1 + P(A \cap B) \Rightarrow p[1 + (r/q)] = 1 + rp$

$\therefore \quad p(q + r) = q(1 + pr).$

Example 4·41. *(a) Twelve balls are distributed at random among three boxes. What is the probability that the first box will contain 3 balls?*

(b) If n biscuits be distributed among N persons, find the chance that a particular person receives r (< n) biscuits. **[Marathwada Univ. B.Sc. 1992]**

Solution. *(a)* Since each ball can go to any one of the three boxes, there are 3 ways in which a ball can go to any one of the three boxes. Hence there are 3^{12} ways in which 12 balls can be placed in the three boxes.

Number of ways in which 3 balls out of 12 can go to the first box is $^{12}C_3$. Now the remaining 9 balls are to be placed in 2 boxes and this can be done in 2^9 ways. Hence the total number of favourable cases $= {}^{12}C_3 \times 2^9$.

$\therefore$ Required probability $= \dfrac{{}^{12}C_3 \times 2^9}{3^{12}}$

(b) Take any one biscuit. This can be given to any one of the N beggars so that there are N ways of distributing any one biscuit. Hence the total number of ways in which n biscuit can be distributed at random among N beggars

$$= N \,.\, N \ldots N \,(n \text{ times}) = N^n.$$

'r' biscuits can be given to any particular beggar in ^nC_r ways. Now we are left with $(n - r)$ biscuits which are to be distributed among the remaining $(N - 1)$ beggars and this can be done in $(N - 1)^{n-r}$ ways.

$\therefore$ Number of favourable cases $= {}^nC_r \,.\, (N - 1)^{n-r}$

Hence, required probability $= \dfrac{{}^nC_r\,(N-1)^{n-r}}{N^n}$

Example 4·42. *A car is parked among N cars in a row, not at either end. On his return the owner finds that exactly r of the N places are still occupied. What is the probability that both neighbouring places are empty?*

Solution. Since the owner finds on return that exactly r of the N places (including owner's car) are occupied, the exhaustive number of cases for such an arrangement is $^{N-1}C_{r-1}$ [since the remaining $r - 1$ cars are to be parked in the remaining $N - 1$ places and this can be done in $^{N-1}C_{r-1}$ ways].

Let A denote the event that both the neighbouring places to owner's car are empty. This requires the remaining $(r - 1)$ cars to be parked in the remaining $N - 3$ places and hence the number of cases favourable to A is $^{N-3}C_{r-1}$. Hence

$$P(A) = \frac{{}^{N-3}C_{r-1}}{{}^{N-1}C_{r-1}} = \frac{(N-r)(N-r-1)}{(N-1)(N-2)}$$

Example 4·43. *What is the probability that at least two out of n people have the same birthday? Assume 365 days in a year and that all days are equally likely.*

Solution. Since the birthday of any person can fall on any one of the 365 days, the exhaustive number of cases for the birthdays of n persons is 365^n.

If the birthdays of all the n persons fall on different days, then the number of favourable cases is

$$365\,(365-1)\,(365-2)\,....\,[365-(n-1)],$$

because in this case the birthday of the first person can fall on any one of 365 days, the birthday of the second person can fall on any one of the remaining 364 days and so on.

Hence the probability (p) that birthdays of all the n persons are different is given by :

$$p = \frac{365\,(365-1)\,(365-2)\,...\,[365-(n-1)]}{365^n}$$

$$= \left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\left(1-\frac{3}{365}\right)...\left(1-\frac{n-1}{365}\right)$$

Hence the required probability that at least two persons have the same birthday is

$$1-p = 1-\left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\left(1-\frac{3}{365}\right)...\left(1-\frac{n-1}{365}\right)$$

Example 4·44. *A five-figure number is formed by the digits 0, 1, 2, 3, 4 (without repetition). Find the probability that the number formed is divisible by 4.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1990]

Solution. The total number of ways in which the five digits 0, 1, 2, 3, 4 can be arranged among themselves is 5!. Out of these, the number of arrangements which begin with 0 (and, therefore, will give only 4-digited numbers) is 4!. Hence the total number of five digited numbers that can be formed from the digits 0, 1, 2, 3, 4 is

$$5\,! - 4\,! = 120 - 24 = 96$$

The number formed will be divisible by 4 if the number formed by the two digits on extreme right (*i.e.*, the digits in the unit and tens places) is divisible by 4. Such numbers are :

$$04\,,\,12\,,\,20\,,\,24\,,\,32\,,\text{ and }40$$

If the numbers end in 04, the remaining three digits, *viz.*,1, 2 and 3 can be arranged among themselves in 3 ! ways. Similarly, the number of arrangements of the numbers ending with 20 and 40 is 3 ! in each case.

If the numbers end with 12, the remaining three digits 0 , 3 , 4 can be arranged in 3 ! ways. Out of these we shall reject those numbers which start with 0 (*i.e.*, have 0 as the first digit). There are $(3-1)\,! = 2\,!$ such cases. Hence, the number of five digited numbers ending with 12 is

$$3\,! - 2\,! = 6 - 2 = 4$$

Similarly the number of 5 digited numbers ending with 24 and 32 each is 4. Hence the total number of favourable cases is

$$3 \times 3! + 3 \times 4 = 18 + 12 = 30$$

Hence required probability $= \frac{30}{96} = \frac{5}{16}$

Example 4·45. *(Huyghen's problem). A and B throw alternately with a pair of ordinary dice. A wins if he throws 6 before B throws 7, and B wins if he throws 7 before A throws 6. If A begins, show that his chance of winning is 30 / 61*

[Delhi Univ. B.Sc. (Stat. Hons.), 1991; Delhi Univ. B.Sc.,1987]

Solution. Let E_1 denote the event of A's throwing '6' and E_2 the event of B's throwing '7 with a pair of dice. Then $\bar{E}_1$ and $\bar{E}_2$ are the complementary events.

'6' can be obtained with two dice in the following ways:

(1, 5), (5, 1), (2, 4), (4, 2), (3, 3), *i.e.*, in 5 distinct ways.

$$\therefore \quad P(E_1) = \frac{5}{36} \quad \text{and} \quad P(\bar{E}_1) = 1 - \frac{5}{36} = \frac{31}{36}$$

'7' can be obtained with two dice as follows:

(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3), *i.e.*, in 6 distinct ways.

$$\therefore \quad P(E_2) = \frac{6}{36} = \frac{1}{6} \quad \text{and} \quad P(\bar{E}_2) = 1 - \frac{1}{6} = \frac{5}{6}$$

If A starts the game, he will win in the following mutually exclusive ways:

(i) E_1 happens *(ii)* $\bar{E}_1 \cap \bar{E}_2 \cap E_1$ happens

(iii) $\bar{E}_1 \cap \bar{E}_2 \cap \bar{E}_1 \cap \bar{E}_2 \cap E_1$ happens, and so on.

Hence by addition theorem of probability, the required probability of A's winning, (say), $P(A)$ is given by

$$P(A) = P(i) + P(ii) + P(iii) + \ldots$$

$$= P(E_1) + P(\bar{E}_1 \cap \bar{E}_2 \cap E_1) + P(\bar{E}_1 \cap \bar{E}_2 \cap \bar{E}_1 \cap \bar{E}_2 \cap E_1) + \ldots.$$

$$= P(E_1) + P(\bar{E}_1)\,P(\bar{E}_2)\,P(E_1) + P(\bar{E}_1)\,P(\bar{E}_2)\,P(\bar{E}_1)\,P(\bar{E}_2)\,P(E_1) + \ldots$$

(By compound probability theorem)

$$= \frac{5}{36} + \frac{31}{36} \times \frac{5}{6} \times \frac{5}{36} + \frac{31}{36} \times \frac{5}{6} \times \frac{31}{36} \times \frac{5}{6} \times \frac{5}{36} + \ldots$$

$$= \frac{5/36}{1 - \frac{31}{36} \times \frac{5}{6}} = \frac{30}{61}$$

Example 4·46. *A player tosses a coin and is to score one point for every head and two points for every tail turned up. He is to play on until his score reaches or passes n. If p_n is the chance of attaining exactly n score, show that*

$$p_n = \frac{1}{2}[p_{n-1} + p_{n-2}],$$

and hence find the value of p_n. **[Delhi Univ. B.Sc. (Stat. Hons.),1992]**

Solution. The score n can be reached in the following two mutually exclusive ways:

(i) By throwing a tail when score is $(n-2)$, and

(ii) By throwing a head when score is $(n-1)$.

Hence by addition theorem of probability, we get

$$p_n = P(i) + P(ii) = \frac{1}{2} \cdot p_{n-2} + \frac{1}{2} \cdot p_{n-1} = \frac{1}{2}(p_{n-1} + p_{n-2}) \quad \ldots(*)$$

To find p_n explicitly, (*) may be re-written as

$$\begin{aligned} p_n + \frac{1}{2} p_{n-1} &= p_{n-1} + \frac{1}{2} p_{n-2} \\ &= p_{n-2} + \frac{1}{2} p_{n-3} \\ &\quad \ldots \quad \ldots \\ &\quad \ldots \quad \ldots \\ &= p_2 + \frac{1}{2} p_1 \end{aligned} \quad \ldots(**)$$

Since the score 2 can be obtained as

(i) Head in first throw and head in 2nd throw,

(ii) Tail in the first throw, we have

$$p_2 = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \text{ and obviously } p_1 = \frac{1}{2}$$

Hence, from (**), we get

$$p_n + \frac{1}{2} p_{n-1} = \frac{3}{4} + \frac{1}{2} \cdot \frac{1}{2} = 1 = \frac{2}{3} + \frac{1}{3} = \frac{2}{3} + \frac{1}{2} \cdot \frac{2}{3}$$

$$\left.\begin{aligned} p_n - \tfrac{2}{3} &= (-\tfrac{1}{2})(p_{n-1} - \tfrac{2}{3}) \\ p_{n-1} - \tfrac{2}{3} &= (-\tfrac{1}{2})(p_{n-2} - \tfrac{2}{3}) \\ &\vdots \qquad\qquad \vdots \\ p_2 - \tfrac{2}{3} &= (-\tfrac{1}{2})(p_1 - \tfrac{2}{3}) \end{aligned}\right\}$$

Multiplying all the above equations, we get

$$\begin{aligned} p_n - \tfrac{2}{3} &= (-\tfrac{1}{2})^{n-1} (p_1 - \tfrac{2}{3}) \\ &= (-\tfrac{1}{2})^{n-1} (\tfrac{1}{2} - \tfrac{2}{3}) = (-1)^n \cdot \frac{1}{2^n} \cdot \frac{1}{3} \\ \Rightarrow \quad p_n &= \frac{2}{3} + (-1)^n \frac{1}{2^n} \cdot \frac{1}{3} \\ &= \frac{1}{3}\left[2 + (-1)^n \frac{1}{2^n}\right] \end{aligned}$$

Example 4·47. *A coin is tossed (m+n) times, (mn). Show that the probability of at least m consecutive heads is* $\dfrac{n+2}{2^{m+1}}$.

[Kurukshetra Univ. M.Sc. 1990; Calcutta Univ. B.Sc.(Hons.),1986]

Solution. Since $m > n$, only one sequence of m consecutive heads is possible. This sequence may start either with the first toss or second toss or third toss, and so on, the last one will be starting with $(n+1)th$ toss.

Let E_i denote the event that the sequence of m consecutive heads starts with ith toss. Then the required probability is

$$P(E_1) + P(E_2) + \ldots + P(E_{n+1}) \qquad \ldots(*)$$

Now $P(E_1) = P$ [Consecutive heads in first m tosses and head or tail in the rest]

$$= \left(\frac{1}{2}\right)^m$$

$P(E_2) = P$ [Tail in the first toss, followed by m consecutive heads and head or tail in the next]

$$= \frac{1}{2}\left(\frac{1}{2}\right)^m = \frac{1}{2^{m+1}}$$

In general,

$P(E_r) = P$ [tail in the $(r-1)$th trial followed by m consecutive heads and head or tail in the next]

$$= \frac{1}{2}\left(\frac{1}{2}\right)^m = \frac{1}{2^{m+1}}, \quad \forall \quad r = 2, 3, \ldots, n+1.$$

Substituting in (*),

$$\text{Required probability} = \frac{1}{2^m} + \frac{n}{2^{m+1}} = \frac{2+n}{2^{m+1}}$$

Example 4·48. *Cards are dealt one by one from a well-shuffled pack until an ace appears. Show that the probability that exactly n cards are dealt before the first ace appears is*

$$\frac{4(51-n)(50-n)(49-n)}{52.51.50.49}$$

[Delhi Univ. B.Sc. 1992]

Solution. Let E_i denote the event that an ace appears when the ith card is dealt. Then the required probability 'p' is given by

$p = P$ [Exactly n cards are dealt before the first ace appears]

$= P$ [The first ace appears at the $(n+1)$th dealing]

$$= P(\bar{E}_1 \cap \bar{E}_2 \cap \bar{E}_3 \cap \ldots \cap \bar{E}_{n-1} \cap \bar{E}_n \cap E_{n+1})$$

$$= P(\bar{E}_1)\, P(\bar{E}_2 \mid \bar{E}_1)\, P(\bar{E}_3 \mid \bar{E}_1 \cap \bar{E}_2) \ldots \times P(\bar{E}_n \mid \bar{E}_1 \cap \bar{E}_2 \cap \ldots \cap \bar{E}_{n-1}) \times P(E_{n+1} \mid \bar{E}_1 \cap \bar{E}_2 \cap \ldots \cap \bar{E}_n) \qquad \ldots(*)$$

Now

$$P(E_1) = \frac{4}{52} \quad \Rightarrow \quad P(\bar{E}_1) = \frac{48}{52}$$

$$P(E_2 \mid \bar{E}_1) = \frac{4}{51} \quad \Rightarrow \quad P(\bar{E}_2 \mid \bar{E}_1) = \frac{47}{51}$$

$$P(E_3 \mid \bar{E}_1 \cap \bar{E}_2) = \frac{4}{50} \quad \Rightarrow \quad P(\bar{E}_3 \mid \bar{E}_1 \cap \bar{E}_2) = \frac{46}{50}$$

$$\vdots \qquad\qquad \vdots$$

$$P(E_{n-1} \mid \bar{E}_1 \cap \bar{E}_2 \cap \ldots \cap \bar{E}_{n-2}) = \frac{4}{52-(n-2)}$$

$$\therefore \quad P(\bar{E}_{n-1} \mid \bar{E}_1 \cap \bar{E}_2 \cap \ldots \cap \bar{E}_{n-2}) = \frac{50-n}{52-(n-2)}$$

$$P(E_n \mid \bar{E}_1 \cap \bar{E}_2 \cap \ldots \cap \bar{E}_{n-1}) = \frac{4}{52-(n-1)}$$

$$\therefore \quad P(\bar{E}_n \mid \bar{E}_1 \cap \bar{E}_2 \cap \ldots \cap \bar{E}_{n-1}) = \frac{49-n}{52-(n-1)}$$

and $\quad P(E_{n+1} \mid \bar{E}_1 \cap \bar{E}_2 \cap \ldots \cap \bar{E}_n) = \frac{4}{52-n}$

Hence, from (*) we get

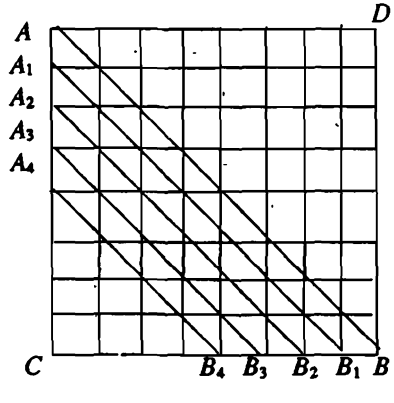
$$p = \left[\frac{48}{52} \times \frac{47}{51} \times \frac{46}{50} \times \frac{45}{49} \times \frac{44}{48} \times \frac{43}{47} \times \ldots \times \frac{52-n}{52-(n-4)} \times \frac{51-n}{52-(n-3)} \times \frac{50-n}{52-(n-2)} \times \frac{49-n}{52-(n-1)} \times \frac{4}{52-n}\right]$$

$$= \frac{(51-n)(50-n)(49-n)\,4}{52 \times 51 \times 50 \times 49}$$

Example 4·49. *If four squares are chosen at random on a chess-board, find the chance that they should be in a diagonal line.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

Solution. In a chess-board there are $8 \times 8 = 64$ squares as shown in the following diagram.



Let us consider the number of ways in which the 4 squares selected at random are in a diagonal line parallel to AB. Consider the ΔABC. Number of ways in which 4 selected squares are along the lines A_4B_4, A_3B_3, A_2B_2, A_1B_1 and AB are ${}^4C_4$, ${}^5C_4$, ${}^6C_4$, ${}^7C_4$ and ${}^8C_4$ respectively.

Similarly, in ΔABD there are an equal number of ways of selecting 4 squares in a diagonal line parallel to AB.

Hence, total number of ways in which the 4 selected squares are in a diagonal line parallel to AB are $2({}^4C_4 + {}^5C_4 + {}^6C_4 + {}^7C_4) + {}^8C_4$.

Since there is an equal number of ways in which 4 selected squares are in a diagonal line parallel to CD, the required number of favourable cases is given by

$$2\,[\,2(\,{}^4C_4 + {}^5C_4 + {}^6C_4 + {}^7C_4) + {}^8C_4\,]$$

Since 4 squares can be selected out of 64 in ${}^{64}C_4$ ways, the required probability is

$$= \frac{2\,[\,2(\,{}^4C_4 + {}^5C_4 + {}^6C_4 + {}^7C_4) + {}^8C_4]}{{}^{64}C_4}$$

$$= \frac{[4\,(\,1 + 5 + 15 + 35) + 140\,] \times 4!}{64 \times 63 \times 62 \times 61} = \frac{91}{158844}$$

Example 4·50. *An urn contains four tickets marked with numbers 112, 121, 211, 222 and one ticket is drawn at random. Let A_i, $(i=1, 2, 3)$ be the event that ith digit of the number of the ticket drawn is 1. Discuss the independence of the events A_1, A_2 and A_3.* **[Delhi Univ. B.Sc.(Stat. Hons.),1987; Poona Univ. B.Sc.,1986]**

Solution. We have

$$P(A_1) = \frac{2}{4} = \frac{1}{2} = P(A_2) = P(A_3)$$

$A_1 \cap A_2$ is the event that the first two digits in the number which the selected ticket bears are each equal to unity and the only favourable case is ticket with number 112.

$$\therefore \qquad P(A_1 \cap A_2) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$$

$$= P(A_1)\,P(A_2)$$

Similarly,

$$P(A_2 \cap A_3) = \frac{1}{4} = P(A_2)\,P(A_3)$$

and $$P(A_3 \cap A_1) = \frac{1}{4} = P(A_3)\,P(A_1)$$

Thus we conclude that the events A_1, A_2 and A_3 are pairwise independent.

Now $P(A_1 \cap A_3 \cap A_3)$ = P {all the three digits in the number are 1's}

$$= P(\phi)$$

$$= 0 \neq P(A_1)\,P(A_2)\,P(A_3)$$

Hence A_1, A_2 and A_3 though pairwise independent are not mutually independent.

Example 4·51. *Two fair dice are thrown independently. Three events A, B and C are defined as follows:*

A : Odd face with first dice

B : Odd face with second dice

C : Sum of points on two dice is odd.

Are the events A, B and C mutually independent?

[Delhi Univ. B.Sc. (Stat. Hons.) 1983; M.S. Baroda Univ. B.Sc.1987]

Solution. Since each of the two dice can show any one of the six faces 1, 2, 3, 4, 5, 6, we get :

$$P(A) = \frac{3 \times 6}{36} = \frac{1}{2} \qquad [\because A = \{1, 3, 5\} \times \{1, 2, 3, 4, 5, 6\}]$$

$$P(B) = \frac{3 \times 6}{36} = \frac{1}{2} \qquad [\because B = \{1, 2, 3, 4, 5, 6\} \times \{1, 3, 5\}]$$

The sum of points on two dice will be odd if one shows odd number and the other shows even number. Hence favourable cases for C are :

(1, 2), (1, 4), (1, 6) ;	(4, 1), (4, 3), (4, 5)
(2, 1), (2, 3), (2, 5) ;	(5, 2), (5, 4), (5, 6)
(3, 2), (3, 4), (3, 6) ;	(6, 1), (6, 3), (6, 5)

i.e., 18 cases in all.

Hence $P(C) = \frac{18}{36} = \frac{1}{2}$.

Cases favourable to the events $A \cap B, A \cap C, B \cap C$ and $A \cap B \cap C$ are given below :

Event	*Favourable cases*
$A \cap B$	(1,1), (1, 3), (1, 5), (3, 1), (3, 3), (3, 5), (5, 1) (5, 3) (5, 5), *i.e.*, 9 in all.
$A \cap C$	(1, 2), (1, 4), (1, 6), (3, 2), (3, 4), (3, 6), (5, 2), (5, 4) (5, 6), *i.e.*, 9 in all.
$B \cap C$	(2, 1), (4, 1), (6, 1) (2, 3), (4, 3), (6, 3), (2, 5), (4, 5), (6, 5), *i.e.*, 9 in all
$A \cap B \cap C$	Nil, because $A \cap B$ implies that sum of points on two dice is even and hence $(A \cap B) \cap C = \phi$

$$\therefore \quad P(A \cap B) = \frac{9}{36} = \frac{1}{4} = P(A) \cdot P(B)$$

$$P(A \cap C) = \frac{9}{36} = \frac{1}{4} = P(A)\, P(C)$$

$$P(B \cap C) = \frac{9}{36} = \frac{1}{4} = P(B)\, P(C)$$

and $$P(A \cap B \cap C) = P(\phi) = 0 \neq P(A)\, P(B)\, P(C)$$

Hence the events A, B and C are pairwise independent but not mutually independent.

Example 4·52. *Let $A_1, A_2, ..., A_n$ be independent events and $P\{A_k\} = p_k$. Further, let p be the probability that none of the events occurs; then show that*

$$p \leq e^{-\Sigma p_k}$$

[Agra Univ. M.Sc., 1987]

Solution. We have

$$p = P(\bar{A}_1 \cap \bar{A}_2 \cap \ldots \cap \bar{A}_n)$$

$$= \prod_{i=1}^{n} P(\bar{A}_i) = \prod_{i=1}^{n} [1 - P(A_i)] = \prod_{i=1}^{n} (1 - p_i)$$

[since A_i's are independent]

$$\leq \prod_{i=1}^{n} e^{-p_i} \qquad [\because 1 - x \leq e^{-x} \text{ for } 0 \leq x \leq 1 \text{ and } 0 \leq p_i \leq 1]$$

$$\Rightarrow \quad p \leq \exp\left[-\sum_{i=1}^{n} p_i\right],$$

as desired.

Remark. We have

$$1 - x \leq e^{-x} \text{ for } 0 \leq x \leq 1 \qquad \ldots(*)$$

Proof. The inequality (*) is obvious for $x = 0$ and $x = 1$. Consider $0 < x < 1$. Then

$$\log (1 - x)^{-1} = -\log (1 - x)$$

$$= \left[x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \ldots\right], \qquad \ldots(**)$$

the expansion being valid since $0 < x < 1$. Further since $x > 0$, we get from (**)

$$\log (1 - x)^{-1} > x$$

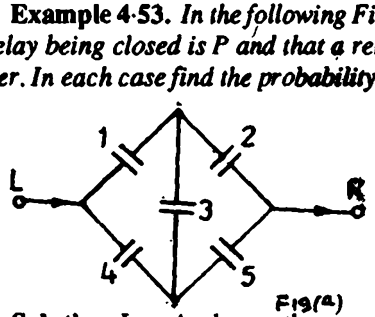
$$\Rightarrow \quad -\log (1 - x) > x$$

$$\Rightarrow \quad \log (1 - x) < -x$$

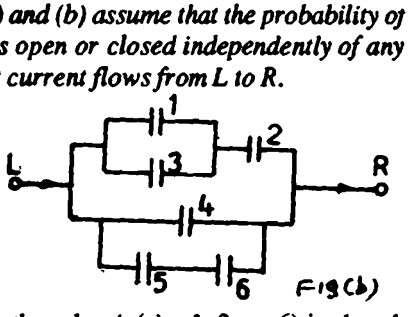
$$\Rightarrow \quad 1 - x < e^{-x},$$

as desired.

Example 4·53. *In the following Fig.(a) and (b) assume that the probability of a relay being closed is P and that a relay is open or closed independently of any other. In each case find the probability that current flows from L to R.*



Fig(a)



Fig(b)

Solution. Let A_i denote the event that the relay i, ($i = 1, 2, \ldots, 6$) is closed. Let E be the event that current flows from L to R.

In Fig. (a) the current will flow from L to R if at least one of the circuits from L to R is closed. Thus for the current to flow from L to R we have the following favourable cases:

$(i)\ A_1 \cap A_2 = B_1, \quad (ii)\ A_4 \cap A_5 = B_2,$

$(iii)\ A_1 \cap A_3 \cap A_5 = B_3, \quad (iv)\ A_4 \cap A_3 \cap A_2 = B_4,$

The probability p_1 that current flows from L to R is given by

$$p_1 = P(B_1 \cup B_2 \cup B_3 \cup B_4) = \sum_i P(B_i) - \sum_{i<j} P(B_i \cap B_j) + \sum_{i<j<k} P(B_i \cap B_j \cap B_k) - P(B_1 \cap B_2 \cap B_3 \cap B_4) \quad ...(*)$$

Since the relays operate independently of each other, we have

$$P(B_1) = P(A_1 \cap A_2) = P(A_1) \cdot P(A_2) = p \cdot p = p^2$$
$$P(B_2) = P(A_4 \cap A_5) = P(A_4) \cdot P(A_5) = p \cdot p = p^2$$
$$P(B_3) = P(A_1)\, P(A_3)\, P(A_5) = p^3$$
$$P(B_4) = P(A_4)\, P(A_3)\, P(A_2) = p^3$$

Similarly

$$P(B_1 \cap B_2) = P(A_1 \cap A_2 \cap A_4 \cap A_5) = P(A_1)\, P(A_2)\, P(A_4)\, P(A_5) = p^4$$

$$P(B_1 \cap B_2 \cap B_3) = P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = p^5$$

and so on. Finally, substituting in (*), we get

$$p_1 = (p^2 + p^2 + p^3 + p^3) - (p^4 + p^4 + p^4 + p^4 + p^4 + p^5) + (p^5 + p^5 + p^5 + p^5) - p^5$$
$$= 2p^2 + 2p^3 - 5p^4 + 2p^5$$

In Fig. (b). Arguing as in the above case, the required probability p_2 that the current flows from L to R is given by

$$p_2 = P(E_1 \cup E_2 \cup E_3 \cup E_4)$$

where

$$E_1 = A_1 \cap A_2,\ E_2 = A_3 \cap A_2,\ E_3 = A_4,\ E_4 = A_5 \cap A_6$$

$$\therefore \quad p_2 = \sum_i P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \sum_{i<j<k} P(E_i \cap E_j \cap E_k) - P(E_1 \cap E_2 \cap E_3 \cap E_4)$$
$$= (p^2 + p^2 + p + p^2) - (p^3 + p^3 + p^4 + p^3 + p^4 + p^3) + (p^4 + p^5 + p^5 + p^5) - p^6$$
$$= p + 3p^2 - 4p^3 - p^4 + 3p^5 - p^6$$

Matching Problem. Let us have n letters corresponding to which there exist n envelopes bearing different addresses. Considering various letters being put in various envelopes, *a match is said to occur* if a letter goes into the right envelope. (Alternatively, if in a party there are n persons with n different hats, *a match is said to occur* if in the process of selecting hats *at random*, the ith person rightly gets the ith hat.)

A match at the kth position for k=1, 2, ..., n. Let us first consider the event A_k when a match occurs at the kth place. For better understanding let us put the envelopes bearing numbners 1, 2, ..., n in ascending order. When A_k occurs, k th

letter goes to the kth envelope but $(n-1)$ letters can go to the remaining $(n-1)$ envelopes in $(n-1)!$ ways.

Hence $P(A_k) = \dfrac{(n-1)!}{n!} = \dfrac{1}{n}$,

where $P(A_k)$ denotes the probability of the kth match. It is interesting to see that $P(A_k)$ *does not depend on* k.

Example 4·54. (a) *'n' different objects 1, 2, ..., n are distributed at random in n places marked 1, 2, ..., n. Find the probability that none of the objects occupies the place corresponding to its number.* **[Calcutta Univ. B.A.(Stat.Hons.)1986; Delhi Univ. B.Sc.(Maths Hons.), 1990; B.Sc.(Stat.Hons.) 1988]**

(b) If n letters are randomly placed in correctly addressed envelopes,prove that the probability that exactly r leters are placed in correct envelopes is given by

$$\frac{1}{r!}\sum_{k=0}^{n-r}(-1)^k\frac{1}{k!};\ r=1,2,\ldots,n$$

[Bangalore Univ. B.Sc., 1987]

Solution (Probability of no match). Let E_i, $(i = 1, 2, ..., n)$ denote the event that the ith object occupies the place corresponding to its number so that $\overline{E}_i$, is the complementary event. Then the probability 'p' that none of the objects occupies the place corresponding to its number is given by

$$\begin{aligned}
p &= P(\overline{E}_1 \cap \overline{E}_2 \cap \overline{E}_3 \cap \ldots \overline{E}_n)\\
&= 1 - P\ \{\text{at least one of the objects occupies the place corresponding to its number}\}\\
&= 1 - P(E_1 \cup E_2 \cup E_3 \cup \ldots \cup E_n)\\
&= 1 - [\sum_{i=1}^{n} P(E_i) - \sum_{\substack{i,j=1\\ i<j}}^{n} P(E_i \cap E_j) + \sum_{\substack{i,j,k=1\\ i<j<k}}^{n} P(E_i \cap E_j \cap E_k) - \ldots\\
&\qquad + (-1)^{n-1}\ P(E_1 \cap E_2 \cap \ldots \cap E_n)] \qquad \ldots(*)
\end{aligned}$$

Now $P(E_i) = \dfrac{1}{n}$, $\forall\ i$

$$\begin{aligned}
P(E_i \cap E_j) &= P(E_i)\,P(E_j \mid E_i)\\
&= \frac{1}{n}\cdot\frac{1}{n-1},\ \forall\ \ i,j\ (i<j)\\
P(E_i \cap E_j \cap E_k) &= P(E_i)\,P(E_j \mid E_i)\,P(E_k \mid E_i \cap E_j)\\
&= \frac{1}{n}\cdot\frac{1}{n-1}\cdot\frac{1}{n-2},\ \forall\ \ i,j,k\ (i<j<k)
\end{aligned}$$

and so on. Finally,

$$P(E_1 \cap E_2 \cap E_3 \cap \ldots \cap E_n) = \frac{1}{n}\cdot\frac{1}{n-1}\cdot\frac{1}{n-2}\cdots\frac{1}{2}\cdot 1$$

Substituting in (*), we get

$$p = 1 - \left[{}^nC_1 \frac{1}{n} - {}^nC_2 \frac{1}{n(n-1)} + {}^nC_3 \frac{1}{n(n-1)(n-2)} - \ldots + (-1)^{n-1} \frac{1}{n(n-1)\ldots 3 \cdot 2 \cdot 1} \right]$$

$$= 1 - \left[1 - \frac{1}{2!} + \frac{1}{3!} - \ldots + (-1)^{n-1} \frac{1}{n!} \right]$$

$$= \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \ldots + (-1)^n \frac{1}{n!}$$

$$= \sum_{k=0}^{n} \frac{(-1)^k}{k!}$$

Remark. For large n,

$$p = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \ldots$$

$$= e^{-1} = 0{\cdot}36787$$

Hence the *probability of at least one match is*

$$1 - p = 1 - \frac{1}{2!} + \frac{1}{3!} - \ldots + \frac{(-1)^n}{n!}$$

$$= 1 - \frac{1}{e}, \text{ (for large } n)$$

(b) **[Probability of exactly r matches $\{r \leq (n-2)\}$]** Let A_i, $(i = 1, 2, \ldots, n)$ denote the event that ith letter goes to the correct envelope. Then the probability that none of the n letters goes to the correct envelope is

$$P(\bar{A}_1 \cap \bar{A}_2 \cap \ldots \cap \bar{A}_n) = \sum_{k=0}^{n} (-1)^k / k! \qquad \ldots(**) \text{ [(c.f. part (a)]}$$

The probability that each of the 'r' letters is in the right envelope is $\frac{1}{n(n-1)(n-2)\ldots(n-r+1)}$, and the probability that none of the remaining $(n-r)$ letters goes in the correct envelope is obtained by replacing n by $(n-r)$ in (**) and is thus given by $\sum_{k=0}^{n-r} \frac{(-1)^k}{k!}$. Hence by compound probability theorem, the probability that out of n letters exactly r letters go to correct envelopes, (in a specified order), is

$$\frac{1}{n(n-1)(n-2)\ldots(n-r+1)} \sum_{k=0}^{n-r} \frac{(-1)^k}{k!} \; ; \quad r \leq n-2.$$

Since r letters can go to n envelopes in ${}^nC_r$ mutually exclusive ways, the required probability of exactly r letters going to correct envelopes, (in any order, whatsoever), is given by

$$^nC_r \times \frac{1}{n(n-1)(n-2)\ldots(n-r+1)} \sum_{k=0}^{n-r} \frac{(-1)^k}{k!} = \frac{1}{r!} \sum_{k=0}^{n-r} (-1)^k \frac{1}{k!}$$

Example 4·55. *Each of the n urns contains 'a' white balls and 'b' black balls. One ball is transferred from the first urn to the second, then one ball from the latter into the third, and so on. If p_k is the probability of drawing a white ball from the kth urn, show that*

$$p_{k+1} = \frac{a+1}{a+b+1} p_k + \frac{a}{a+b+1}(1-p_k)$$

Hence for the last urn, prove that

$$p_n = \frac{a}{a+b}$$ **[Punjab Univ. B.Sc.(Maths Hons.),1988]**

Solution. The event of drawing a white ball from the kth urn can materialise in the following two ways:

(i) The ball transferred from the $(k-1)$th urn is white and then a white ball is drawn from the kth urn.

(ii) The ball transferred from the $(k-1)$th urn is black and then a white ball is drawn from the kth urn.

The probability of case *(i)* is $p_{k-1} \times \frac{a+1}{a+b+1}$,

since the probability of drawing a white ball from the $(k-1)$th urn is p_{k-1} and then the probability of drawing white ball from the kth urn is

$$\frac{a+1}{a+b+1}.$$

Since the probability of drawing a black ball from the $(k-1)$th urn is $[1-p_{k-1}]$ and then the probability of drawing a white ball from the kth urn is

$$\frac{a}{a+b+1},$$

the probability of case *(ii)* is given by

$$\frac{a}{a+b+1}[1-p_{k-1}]$$

Since the cases *(i)* and *(ii)* are mutually exclusive, we have by addition theorem of probability

$$p_k = \frac{a+1}{a+b+1} p_{k-1} + \frac{a}{a+b+1}[1-p_{k-1}] \quad \ldots(*)$$

$$\therefore \quad p_k = \frac{1}{a+b+1} p_{k-1} + \frac{a}{a+b+1} \quad \ldots(1)$$

Replacing k by $k+1$ in (*) we get the required result.

Changing k to $k-1, k-2, \ldots$ and so on, we get

$$p_{k-1} = \frac{1}{a+b+1} p_{k-2} + \frac{a}{a+b+1} \quad \ldots(2)$$

$$p_{k-2} = \frac{1}{a+b+1}\, p_{k-3} + \frac{a}{a+b+1} \qquad \ldots(3)$$

. . .

. . .

. . .

$$p_2 = \frac{1}{a+b+1}\, p_1 + \frac{a}{a+b+1} \qquad \ldots(k-1)$$

But p_1 = Probability of drawing a white ball from the first urn $= \frac{a}{a+b}$.

Multiplying (1) by 1, (2) by $\frac{1}{a+b+1}$, (3) by $\left(\frac{1}{a+b+1}\right)^2$, ..., and $(k-1)$th equation by $\left(\frac{1}{a+b+1}\right)^{k-2}$ and adding, we get

$$p_k = \left(\frac{1}{a+b+1}\right)^{k-1} p_1 + \frac{a}{a+b+1}\left[1 + \frac{1}{a+b+1} + \frac{1}{(a+b+1)^2} + \ldots + \left(\frac{1}{a+b+1}\right)^{k-2}\right]$$

$$= \left(\frac{1}{a+b+1}\right)^{k-1} \times \frac{a}{(a+b)} + \frac{a}{a+b+1}\left[\frac{1-\left(\frac{1}{a+b+1}\right)^{k-1}}{\left(1-\frac{1}{a+b+1}\right)}\right]$$

$$= \frac{a}{a+b}\left(\frac{1}{a+b+1}\right)^{k-1} + \frac{a}{a+b}\left[1-\left(\frac{1}{a+b+1}\right)^{k-1}\right]$$

$$= \frac{a}{a+b}\left[\left(\frac{1}{a+b+1}\right)^{k-1} + \left\{1-\left(\frac{1}{a+b+1}\right)^{k-1}\right\}\right]$$

$$= \frac{a}{a+b}, \quad (k = 1, 2, \ldots, n)$$

Since the probability of drawing a white ball from the kth urn is independent of k, we have

$$p_n = \frac{a}{a+b}.$$

Example 4·56. *(i) Let the probability p_n that a family has exactly n children be αp^n when $n \geq 1$ and $p_0 = 1 - \alpha p\,(1 + p + p^2 + \ldots)$. Suppose that all sex distributions of n children have the same probability. Show that for $k \geq 1$, the probability that a family contains exactly k boys is $2\alpha \cdot p^k/(2-p)^{k+1}$.*

(ii) Given that a family includes at least one boy, show that the probability that there are two or more boys is $p/(2-p)$.

Solution. We are given

$$p_n = P \text{ [that a family has exactly } n \text{ children]} = \alpha p^n, \; n \geq 1$$

and $p_0 = 1 - \alpha p (1 + p + p^2 + \ldots)$

Let E_j be the event that the number of children in a family is j and let A be the event that a family contains exactly k boys. Then

$$P(E_j) = p_j; \; j = 0, 1, 2, \ldots$$

Now, since each child can have any of the two sex distributions (either boy or girl), the total number of possible distributions for a family to have 'j' children is 2^j.

$$\therefore \qquad P(A \mid E_j) = \frac{{}^jC_k}{2^j}, \; j \geq k$$

and
$$P(A) = \sum_{j=k}^{\infty} P(E_j) P(A \mid E_j) = \sum_{j=k}^{\infty} p_j P(A \mid E_j)$$

$$= \sum_{j=k}^{\infty} \alpha p^j \left[\frac{{}^jC_k}{2^j} \right], \quad j \geq k \geq 1$$

$$= \alpha \sum_{j=k}^{\infty} \left(\frac{p}{2} \right)^j {}^jC_k$$

$$= \alpha \sum_{r=0}^{\infty} {}^{k+r}C_k \left(\frac{p}{2} \right)^{k+r} \qquad [\text{Put } j - k = r.]$$

$$= \alpha \left(\frac{p}{2} \right)^k \sum_{r=0}^{\infty} {}^{k+r}C_r \left(\frac{p}{2} \right)^r \qquad [\because {}^nC_r = {}^nC_{n-r}]$$

We know that

$${}^{-n}C_r = (-1)^r \cdot {}^{n+r-1}C_r \quad \Rightarrow \quad (-1)^r \cdot {}^{-n}C_r = {}^{n+r-1}C_r$$

$$\therefore \qquad (-1)^r \cdot {}^{-(k+1)}C_r = {}^{k+r}C_r$$

Hence

$$P(A) = \alpha \left(\frac{p}{2} \right)^k \sum_{r=0}^{\infty} (-1)^r \cdot {}^{-(k+1)}C_r \left(\frac{p}{2} \right)^r$$

$$= \alpha \left(\frac{p}{2} \right)^k \sum_{r=0}^{\infty} {}^{-(k+1)}C_r \left(-\frac{p}{2} \right)^r$$

$$= \alpha \left(\frac{p}{2} \right)^k \left(1 - \frac{p}{2} \right)^{-(k+1)}$$

$$= \alpha \left(\frac{p}{2} \right)^k \frac{2^{k+1}}{(2-p)^{k+1}} = \frac{2\alpha p^k}{(2-p)^{k+1}}$$

(b) Let B denote the event that a family includes at least one boy and C denote the event that a family has two or more boys. Then

$$P(B) = \sum_{k=1}^{\infty} P[\text{family has exactly } k \text{ boys}]$$

$$= \sum_{k=1}^{\infty} \frac{2\alpha p^k}{(2-p)^{k+1}} = \frac{2\alpha}{2-p} \sum_{k=1}^{\infty} \left(\frac{p}{2-p}\right)^k$$

$$= \frac{2\alpha}{2-p} \times \frac{p/(2-p)}{1-[p/(2-p)]} = \frac{\alpha p}{(1-p)(2-p)}$$

$$P(C) = \sum_{k=2}^{\infty} P[\text{family has exactly } k \text{ boys}]$$

$$= \sum_{k=2}^{\infty} \frac{2\alpha p^k}{(2-p)^{k+1}} = \frac{2\alpha}{2-p} \sum_{k=2}^{\infty} \left(\frac{p}{2-p}\right)^k$$

$$= \frac{2\alpha}{2-p} \cdot \frac{[p/(2-p)]^2}{1-[p/(2-p)]} = \frac{\alpha p^2}{(2-p)^2(1-p)}$$

Since $C \subset B$ and $B \cap C = C, P(B \cap C) = P(C) \Rightarrow P(B) P(C \mid B) = P(C)$

Therefore,

$$P(C \mid B) = \frac{P(C)}{P(B)} = \frac{\alpha p^2}{(2-p)^2(1-p)} \times \frac{(1-p)(2-p)}{\alpha p} = \frac{p}{2-p}$$

Example 4·57. *A slip of paper is given to person A who marks it either with a plus sign or a minus sign; the probability of his writing a plus sign is 1/3. A passes the slip to B, who may either leave it alone or change the sign before passing it to C. Next C passes the slip to D after perhaps changing the sign. Finally D passes it to a referee after perhaps changing the sign. The referee sees a plus sign on the slip. It is known that B, C and D each change the sign with probability 2/3. Find the probability that A originally wrote a plus.*

Solution. Let us define the following events :

E_1 : A wrote a plus sign; E_2 : A wrote a minus sign

E : The referee observes a plus sign on the slip.

We are given : $P(E_1) = 1/3$, $P(E_2) = 1 - 1/3 = 2/3$

We want $P(E_1 \mid E)$, which by Bayes rule is given by :

$$P(E_1 \mid E) = \frac{P(E_1) P(E \mid E_1)}{P(E_1) P(E \mid E_1) + P(E_2) P(E \mid E_2)} \quad \ldots(i)$$

$P(E \mid E_1) = P$ [Referee observes the plus sign given that 'A' wrote the plus sign an the slip]

$= P$ [(Plus sign was not changed at all) $\cup$ (Plus sign was changed exactly twice in passing from 'A' to referee through B, C and D)]

$= P(\hat{E}_3 \cup \hat{E}_4)$, (say).

$= P(\hat{E}_3) + P(\hat{E}_4)$ $\quad \ldots(ii)$

Let A_1, A_2 and A_3 respectively denote the events that B, C and D change the sign on the slip. Then we are given

$$P(A_1) = P(A_2) = P(A_3) = 2/3 \quad ; \quad P(\bar{A}_1) = P(\bar{A}_2) = P(\bar{A}_3) = 1/3$$

We have

$$P(E_3) = P(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3) = P(\bar{A}_1)\,P(\bar{A}_2)\,P(\bar{A}_3) = (1/3)^3 = 1/27$$

$$P(E_4) = P[(A_1 A_2 \bar{A}_3) \cup (A_1 \bar{A}_2 A_3) \cup (\bar{A}_1 A_2 A_3)]$$

$$= P(A_1 A_2 \bar{A}_3) + P(A_1 \bar{A}_2 A_3) + P(\bar{A}_1 A_2 A_3)$$

$$= P(A_1)\,P(A_2)\,P(\bar{A}_3) + P(A_1)\,P(\bar{A}_2)\,P(A_3) + P(\bar{A}_1)\,P(A_2)\,P(A_3)$$

$$= \frac{2}{3}\cdot\frac{2}{3}\cdot\frac{1}{3} + \frac{2}{3}\cdot\frac{1}{3}\cdot\frac{2}{3} + \frac{1}{3}\cdot\frac{2}{3}\cdot\frac{2}{3} = \frac{4}{9}$$

Substituting in *(ii)* we get

$$P(E \mid E_1) = \frac{1}{27} + \frac{4}{9} = \frac{13}{27} \qquad \ldots(iii)$$

Similarly,

$P(E \mid E_2) = P$ [Referee observes the plus sign given that 'A' wrote minus sign on the slip]

$= P$ [(Minus sign was changed exactly once) $\cup$ (Minus sign was changed thrice)]

$= P(E_5 \cup E_6)$, (say),

$$= P(E_5) + P(E_6) \qquad \ldots(iv)$$

$$P(E_5) = P[(A_1 \bar{A}_2 \bar{A}_3) \cup (\bar{A}_1 A_2 \bar{A}_3) \cup (\bar{A}_1 \bar{A}_2 A_3)]$$

$$= P(A_1)\,P(\bar{A}_2)\,P(\bar{A}_3) + P(\bar{A}_1)\,P(A_2)\,P(\bar{A}_3) + P(\bar{A}_1)\,P(\bar{A}_2)\,P(A_3)$$

$$= \frac{2}{3}\cdot\frac{1}{3}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{2}{3}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{1}{3}\cdot\frac{2}{3} = \frac{2}{9}$$

$$P(E_6) = P(A_1 A_2 A_3) = P(A_1)\,P(A_2)\,P(A_3) = \frac{2}{3}\cdot\frac{2}{3}\cdot\frac{2}{3} = \frac{8}{27}$$

Substituting in *(iv)* we get :

$$P(E \mid E_2) = \frac{2}{9} + \frac{8}{27} = \frac{14}{27}. \qquad \ldots(v)$$

Substituting from *(iii)* and *(v)* in *(i)* we get :

$$P(E_1 \mid E) = \frac{\frac{1}{3}\times\frac{13}{27}}{\frac{1}{3}\times\frac{13}{27} + \frac{2}{3}\times\frac{14}{27}} = \frac{13}{13 + 28} = \frac{13}{41}$$

Example 4·58. *Three urns of the same appearance have the following proportion of balls.*

First urn	:	*2 black*	*1 white*
Second Urn	:	*1 black*	*2 white*
Third urn	:	*2 black*	*2 white*

One of the urns is selected and one ball is drawn. It turns out to be white. What is the probability of drawing a white ball again, the first one not having been returned?

Solution. Let us define the events:

E_i = The event of selection of ith urn, $(i = 1,2,3)$

and A = The event of drawing a white ball.

Then

$$P(E_1) = P(E_2) = P(E_3) = 1/3$$

and $P(A \mid E_1) = 1/3, P(A \mid E_2) = 2/3$ and $P(A \mid E_3) = 1/2$

Let C denote the future event of drawing another white ball from the urns. Then

$$P(C \mid E_1 \cap A) = 0, P(C \mid E_2 \cap A) = 1/2, \text{ and } P(C \mid E_3 \cap A) = 1/3$$

$$\therefore \quad P(C \mid A) = \frac{\sum_{i=1}^{3} P(E_i) P(A \mid E_i) P(C \mid E_i \cap A)}{\sum_{i=1}^{3} P(E_i) P(A \mid E_i)}$$

$$= \frac{\frac{1}{3} \cdot \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3} + \frac{1}{3} \cdot \frac{1}{2}} = \frac{1}{3}$$

MISCELLANEOUS EXERCISE ON CHAPTER IV

1. Probabilities of occurrence of n independent events $E_1, E_2, ..., E_n$ are $p_1, p_2, ..., p_n$ respectively. Find the probability of occurrence of the compound event in which $E_1, E_2, ..., E_r$ occur and $E_{r+1}, E_{r+2}, ..., E_n$ do not occur.

Ans. $\prod_{i=1}^{r} p_i \times \prod_{i=r+1}^{n} (1 - p_i)$

2. Prove that for any integer $m \geq 1$,

(a) $P\left(\bigcap_{i=1}^{m} A_i\right) \leq P(A_i) \leq P\left(\bigcup_{i=1}^{m} A_i\right) \leq \sum_{i=1}^{m} P(A_i)$

(b) $P\left(\bigcap_{i=1}^{m} A_i\right) \geq 1 - \sum_{i=1}^{m} P(\bar{A}_i)$

3. Establish the inequalities:

$$P(A \cap B \cap C) \leq P(A \cap B) \leq P(A \cup B) \leq P(A \cup B \cup C) \leq P(A) + P(B) + P(C)$$

4. Let $A_1, A_2, ..., A_n$ be mutually independent events with $P(A_k) = p_k$, $k = 1, 2, ..., n$.

Let p be the probability that none of the events $A_1, A_2, ..., A_n$ occurs. Show that

$$p = \prod_{k=1}^{n} (1 - p_k) \leq \exp\left\{-\sum_{k=1}^{n} p_k\right\}$$

Use the above relation to compute the probability that in six tosses of a fair die, no "aces are obtained". Compare this with the upper bound given above. Show that if each p_k is small compared with n, the upper bound is a good approximation.

5. A and B play a match, the winner being the one who first wins two games in succession, no games being drawn. Their respective chances of winning a particular game are $p : q$. Find

(i) A's initial chance of winning.

(ii) A's chance of winning after having won the first game.

6. A carpenter has a tool chest with two compartments, each one having a lock. He has two keys for each lock, and he keeps all four keys in the same ring. His habitual procedure in opening a compartment is to select a key at random and try it. If it fails, he selects one of the remaining three and tries it and so on. Show that the probability that he succeeds on the first, second and third try is 1/2,1/3,1/6 respectively. **(Lucknow Univ. B.Sc., 1990)**

7. Three players A, B and C agree to play a series of games observing the following rules : two players participate in each game, while third is idle, and the game is to be won by one of them. The loser in each game quits and his place in the next game is taken by the player who was idle. The player who succeeds in winning over both of his opponents without interruption wins the whole series of games.

Supposing the probability for each player to win a single game is 1/2, and that the first game is played by A and B, find the probability for A, B and C respectively to win the whole series if the number of games is unlimited.

Ans. 5/14, 5/14, 2/7

8. In a certain group of mathematicians, 60 per cent have insufficient background of modern Algebra, 50 per cent have inadequate knowledge of Mathematical Statistics and 80 per cent are in either one or both of the two categories. What is the percentage of people who know Mathematical Statistics among those who have a sufficient background of Modern Algebra? **(Ans. 0·50)**

9. *(a)* If A has $(n+1)$ and B has n fair coins, which they flip, show that the probability that A gets more heads than B is $\frac{1}{2}$.

(b) A student is given a column of 10 dates and column of 10 events and is asked to match the correct date to each event. He is not allowed to use any item more than once. Consider the case where the student knows how to match four of the items but he is very doubtful of the remaining six. He decides to match these at random. Find the probabilities that he will correctly match *(i)* all the items, *(ii)* at least seven of the items, and *(iii)* at least five.

Ans. *(a)* $\frac{1}{6!}$, *(b)* $\frac{10}{6!}$, *(c)* $1-\frac{1}{6!}$

10. An astrologer claims that he can predict before birth the sex of a baby just to be born. Suppose that the astrologer has no real power but he tosses a coin just

once before every birth and if the head turns up he predicts a boy for that birth and if the tail turns up he predicts a girl. Let p be the probability of the event that at a certain birth a male child is born, and p' the probability of a head turning up in a single toss with astrologer's coin. Find the probability of a correct prediction and that of at least one correct prediction in n predictions.

11. From a pack of 52 cards an even number of cards is drawn. Show that the probability of half of these cards being red is

$$[52\,!/(26\,!)^2 - 1] / (2^{51} - 1)$$

12. A sportsman's chance of shooting an animal at a distance $r\ (> a)$ is a^2/r^2. He fires when $r = 2a$, and if he misses he reloads and fires when $r = 3a, 4a...$ If he misses at distance na, the animal escapes. Find the odds against the sportsman.

Ans. $n + 1 : n - 1$

Hint. P [Sportsman shoots at a distance ia] $= \dfrac{a^2}{(ia)^2} = \dfrac{1}{i^2}$

$\Rightarrow$ P [Sportsman misses the shot at a distance ia] $= 1 - \dfrac{1}{i^2}$

$$\therefore \quad P \text{ [Animal escapes]} = \prod_{i=2}^{n}\left(1 - \frac{1}{i^2}\right) = \prod_{i=2}^{n}\left[\left(\frac{i-1}{i}\right)\left(\frac{i+1}{i}\right)\right]$$

$$= \prod_{i=2}^{n}\left(\frac{i-1}{i}\right)\prod_{i=2}^{n}\left(\frac{i+1}{i}\right) = \frac{n+1}{2n}$$

$$\text{Required ratio} = \frac{n+1}{2n} : \left(1 - \frac{n+1}{2n}\right) = (n+1) : (n-1)$$

13. *(a)* Pataudi, the captain of the Indian team, is reported to have observed the rule of calling 'heads' every time the toss was made during the five matches of the Test series with the Australian team. What is the probability of his winning the toss in all the five matches?

Ans. $(1/2)^5$

How will the probability be affected if

(i) he had made a rule of tossing a coin privately to decide whether to call "heads" or "tails" on each occasion.

(ii) the factors determining his choice were not predetermined but he called out whatever occurred to him on the spur of the moment?

(b) A lot contains 50 defective and 50 non-defective bulbs. Two bulbs are drawn at random one at a time, with replacement. The events A, B, C are defined as

$A =$ {The first bulb is defective}

$B =$ {The second bulb is non-defective}

$C =$ {The two bulbs are both defective or both non-defective}

Determine whether

(i) A, B, C are pairwise independent,

(ii) A, B, C are independent.

14. A, B and C are three urns which contain 2 white, 1 black, 3 white, 2 black and 2 white and 2 black balls, respectively. One ball is drawn from urn A and put into the urn B; then a ball is drawn from urn B and put into the urn C. Then a ball is drawn from urn C. Find the probability that the ball drawn is white.

Ans. 4/15.

15. An urn contains a white and b black balls and a series of drawings of one ball at a time is made, the ball removed being returned to the urn immediately after the next drawing is made. If p_n denotes the probability that the nth ball drawn is black, show that

$$p_n = (b - p_{n-1}) / (a + b - 1).$$

Hence find p_n.

16. A person is to be tested to see whether he can differentiate between the taste of two brands of cigarettes. If he cannot differentiate, it is assumed that the probability is one-half that he will identify a cigarette correctly. Under which of the following two procedures is there less chance that he will make all correct identifications when he actually cannot differentiate between the two brands?

(i) The subject is given four pairs each containing both brands of cigarettes (this is known to the subject), he must identify for each pair which cigarette represents each brand.

(ii) The subject is given eight cigarettes and is told that the first four are of one brand and the last four of the other brand.

How do you explain the difference in results despite the fact that eight cigarettes are tested in each case?

Ans. *(i)* 1/16 *(ii)* 1/2

17. (Sampling with replacement). A sample of size r is taken from a population of n people. Find the probability U_r that N given people will be included in the sample.

Ans. $U_r = \sum_{m=0}^{N} (-1)^m \binom{N}{m} \left(1 - \frac{m}{n}\right)^r$

18. In a lottery m tickest are drawn at a time out of the total number of n tickets, and returned before the next drawing is made. Show that the chance that in k drawings, each of the numbers $1, 2, 3, \ldots, n$ will appear at least once is given by

$$P_k = 1 - \binom{n}{1}\left(1 - \frac{m}{n}\right)^k + \binom{n}{2}\left(1 - \frac{m}{n}\right)^k \left(1 - \frac{m}{n-1}\right)^k - \ldots$$

[Nagpur Univ. M.Sc. 1987]

19. In a certain book of N pages, no page contains more than four errors, n_1 of them contain one error, n_2 contain two errors, n_3 contain three errors and n_4 contain four errors. Two copies of the book are opened at any two given pages. Show the probability that the number of errors in these two pages shall not exceed five is

$$1 - \frac{1}{N^2}\,(n_3^2 + n_4^2 + 2n_2 n_4 + 2n_3 n_4)$$

Hint. Let E_i I : the event that a page of first book contains i errors.

and E_i II : the event that a page of second book contains i errors.

P (No. of errors in the two pages shall not exceed 5)

$$= 1 - P\,[E_2\,\mathrm{I}\,E_4\,\mathrm{II} + E_3\,\mathrm{I}\,E_4\,\mathrm{II} + E_4\,\mathrm{I}\,E_4\,\mathrm{II} + E_3\,\mathrm{I}\,E_3\,\mathrm{II} + E_4\,\mathrm{I}\,E_3\,\mathrm{II} + E_4\,\mathrm{I}\,E_2\,\mathrm{II}\,]$$

20. *(a)* Of three independent events, the chance that the first only should happens is a, the chance of the second only is b and the chance of the third only is c. Show that the independent chances of the three events are respectively

$$\frac{a}{a+x},\ \frac{b}{b+x},\ \frac{c}{c+x}$$

where x is the root of the equation

$$(a+x)(b+x)(c+x) = x^2$$

Hint. $P(E_1 \cap \bar{E}_2 \cap \bar{E}_3) = P(E_1)\,[1 - P(E_2)]\,[1 - P(\bar{E}_3)] = a$ ...(*)

$P(\bar{E}_1 \cap E_2 \cap \bar{E}_3) = [1 - P(E_1)]\,P(E_2)\,[1 - P(E_3)] = b$ ...(**)

$P(\bar{E}_1 \cap \bar{E}_2 \cap E_3) = [1 - P(E_1)]\,[1 - P(E_2)]\,P(E_3) = c$ ...(***)

Multiplying (*), (**) and (***), we get

$$P(E_1)\,P(E_2)\,P(E_3)\,x^{-2} = abc,$$

where $x = [1 - P(E_1)]\,[1 - P(E_2)]\,[1 - P(E_3)]$

Multiplying (*) by $[1 - P(E_1)]$, we get

$$P(E_1) = \frac{a}{a+x}, \text{ and so on.}$$

(b) Of three independent events, the probability that the first only should happens is 1/4, the probability that the second only should happen is 1/8, and the probability that the third only should happen is 1/12. Obtain the unconditional probabilities of the three events.

Ans. 1/2, 1/3, 1/4.

(c) A total of n shells are fired at a target. The probability of the ith shell hitting the target is p_i; $i = 1, 2, 3, ..., n$. Assuming that the n firings are n mutually independent events, find the probability that at least two shells out of n hit the target. **[Calcutta Univ. B.Sc.(Maths Hons.), 1988]**

(d) An urn contains M balls numbered 1 to M, where the first K balls are defective and the remaining $M - K$ are non-defective. A sample of n balls is drawn from the urn. Let A_k be the event that the sample of n balls contains exactly k defectives. Find $P(A_k)$ when the sample is drawn *(i)* with replacement and, *(ii)* without replacement. **[Delhi Univ. B.Sc. (Maths Hons.), 1989]**

21. For three independent events A, B and C, the probability for A to occur is a, the probability that A, B and C will not occur is b, and the probability that at least one of the three events will not occur is c. If p denotes the probability that C occurs but neither A nor B occurs, prove that p satisfies the quadratic equation

$$ap^2 + [ab - (1-a)(a+c-1)]\,p + b(1-a)(1-c) = 0$$

and hence deduce that $c > \dfrac{(1-a)^2 + ab}{(1-a)}$

Further show that the probability of occurrence of C is $p/(p+b)$, and that of B's happening is $(1-c)(p+b)/ap$.

Hint. Let $P(A) = x$, $P(B) = y$ and $P(C) = z$

Then $x = a$, $(1-x)(1-y)(1-z) = b$, $1 - xyz = c$

and $p = z(1-x)(1-y)$

Elimination of x, y and z gives quadratic equation in p.

22. *(a)* The chance of success in each trial is p. If p_k is the probability that there are even number of successes in k trials, prove that

$$p_k = p + p_{k-1}(1-2p)$$

Deduce that $p_k = \frac{1}{2}[1 + (1-2p)^k]$

(b) If a day is dry, the conditional probability that the following day will also be dry is p; if a day is wet, the conditional probability that the following day will be dry is p'. If u_n is the probability that the nth day will be dry, prove that

$$u_n - (p - p')\,u_{n-1} - p' = 0 \;;\; n \geq 2$$

If the first day is dry, $p = 3/4$ and $p' = 1/4$, find u_n.

23. There are n similar biased dice such that the probability of obtaining a 6 with each one of them is the same and equal to p. If all the dice are rolled once, show that p_n, the probability that an odd number of 6's is obtained satisfies the difference equation

$$p_n + (2p-1)\,p_{n-1} = p$$

and hence derive an explicit expression for p_n.

Ans. $p_n = \frac{1}{2}[1 + (1-2p)^n]$

24. Suppose that each day the weather can be uniquely classified as 'fine' or 'bad'. Suppose further that the probability of having fine weather on the last day of a certain year is P_0 and we have the probability p that the weather on an arbitrary day will be of the same kind as on the preceding day. Let the probability of having fine weather on the nth day of the following year be P_n. Show that

$$P_n = (2p-1)\,P_{n-1} + (1-p)$$

Deduce that

$$P_3 = (2p-1)^3\left(P_0 - \frac{1}{2}\right) + \frac{1}{2}$$

25. A closet contains n pairs of shoes. If $2r$ shoes are chosen at random (with $2r < n$), what is the probability that there will be *(i)* no complete pair,

(ii) exactly one complete pair, *(iii)* exactly two complete pairs among them?

Hint. *(i)* P(no complete pair)$= \binom{n}{2r} 2^{2r} \div \binom{2n}{2r}$

(ii) P(exactly one complete pair)$= n\binom{n-1}{2r-2} 2^{2r-2} \div \binom{2n}{2r}$

and *(iii)* P(exactly two complete pairs)$= \binom{n}{2}\binom{n-2}{2r-4} 2^{2r-4} \div \binom{2n}{2r}$

26. Show that the probability of getting no right pair out of n, when the left foot shoes are paired randomly with the rigth foot shoes; is the sum of the first $(n+1)$ terms in the expansion of e^{-1}.

27. *(a)* In a town consisting of $(n+1)$ inhabitants, a person narrates a rumour to a second person, who in turn narrates it to a third person, and so on. At each step the recipient of the rumour is chosen at random from the n available persons, excluding the narrator himself. Find the probability that the rumour will be told r times without:

(i) returning to the originator,

(ii) being narrated to any person more than once.

(b) Do the above problem when, at each step the rumour is told by one person to a gathering of N randomly chosen people.

Ans. *(a)* *(i)* $\dfrac{n(n-1)^{r-1}}{n^r} = \left(1-\dfrac{1}{n}\right)^{r-1}$; *(ii)* $\dfrac{n(n-1)(n-2)\ldots(n-r+1)}{n^r}$

(b) *(i)* $\left(1-\dfrac{N}{n}\right)^{r-1}$; *(ii)* $\dfrac{\binom{n}{rN}}{\left[\binom{n}{N}\right]^r}$

28. What is the probability that *(i)* the birthdays of twelve people will fall in twelve different calendar months (assume equal probabilities for the twelve months) and *(ii)* the birthdays of six people will fall in exactly two calendar months?

Hint. *(i)* The birthday of the first person, for instance, can fall in 12 different ways and so for the second, and so on.

$\therefore$ The total number of cases $= 12^{12}$.

Now there are 12 months in which the birthday of one person can fall and 11 months in which the birthday of the second person can fall and 10 months for another third person, and so on.

$\therefore$ The total number of favourable cases $= 12.11.10\ldots3.2.1$

Hence the required probability $= \dfrac{12!}{12^{12}}$

(ii) The total number of ways in which the birthdays of 6 persons can fall in any of the month $= 12^6$.

$\therefore$ The required probability $= \dfrac{\binom{12}{2}\left(2^6-2\right)}{12^6}$

29. An elevator starts with 7 passengers and stops at 10 floors. What is the probability p that no two passengers leave at the same floor?

[Delhi Univ. M.C.A., 1988]

30. A bridge player knows that his two opponents have exactly five hearts between two of them. Each opponent has thirteen cards. What is the probability that there is three-two split on the hearts (that is one player has three hearts and the other two)? **[Delhi Univ. B.Sc.(Maths Hons.), 1988]**

31. An urn contains 2 white and 2 black balls. A ball is drawn at random. If it is white, it is not replaced into the urn. Otherwise it is replaced along with another ball of the same colour. The process is repeated. Find the probability that the third ball drawn is black. **[Burdwan Univ. B.Sc. (Hons.), 1990]**

Ans. $\dfrac{23}{30}$

32. There is a series of n urns. In the ith urn there are i white and $(n-i)$ black balls, $i = 1, 2, 3, ..., k$. One urn is chosen at random and 2 balls are drawn from it. Both turn out to be white. What is the probability that the jth urn was chosen, where j is a particular number between 3 and n.

Hint. Let E_j denote the event of selection of jth urn, $j = 3, 4, ..., n$ and A denote the event of drawing of 2 white balls, then

$$P(A \mid E_j) = \left(\frac{j}{n}\right)\left(\frac{j-1}{n-1}\right),\ P(E_j) = \frac{1}{n},\ P(A) = \sum_{i=1}^{n} \frac{1}{n}\left(\frac{i}{n}\right)\left(\frac{i-1}{n-1}\right)$$

$$\therefore \qquad P(E_j \mid A) = \frac{\frac{1}{n}\left(\frac{j}{n}\right)\left(\frac{j-1}{n-1}\right)}{\sum_{i=1}^{n}\left(\frac{1}{n}\right)\left(\frac{i}{n}\right)\left(\frac{i-1}{n-1}\right)}$$

33. There are $(N+1)$ identical urns marked $0, 1, 2, ..., N$ each of which contains N white and red balls. The kth urn contains k red and $N-k$ white balls, $(k = 0, 1, 2, ... N)$. An urn is chosen at random and n random drawings of a ball are made from it, the ball drawn being replaced after each draw. If the balls drawn are all red, show that the probability that the next drawing will also yield a red ball is approximately $(n+1)(n+2)$ when N is large.

34. A printing machine can print n letters, say $\alpha_1, \alpha_2, ..., \alpha_n$. It is operated by electrical impulses, each letter being produced by a different impulse. Assume that p is the constant probability of printing the correct letter and the impulses are independent. One of the n impulses, chosen at random, was fed into the machine twice and both times the letter α_1 was printed. Compute the probability that the impulse chosen was meant to print α_1. **[Delhi Univ. M.Sc.(Stat.), 1981]**

Ans. $(n-1)p^2/(np^2 - 2p + 1)$

35. Two players A and B agree to contest a match consisting of a series of games, the match to be won by the player who first wins three games, with the provision that if the players win two games each, the match is to continue until it

is won by one player winning two games more than his opponent. The probability of A winning any given game is p, and the games cannot be drawn.

(i) Prove that $f(p)$, the initial probability of A winning the match is given by:

$$f(p) = p^3(4 - 5p + 2p^2)/(1 - 2p + 2p^2)$$

(ii) Show that the equation $f(p) = p$ has five real roots, of which three are admissible values of p. Find these three roots and explain their significance.

[Civil Services (Main), 1986]

36. Two players A and B start playing a series of games with Rs. a and b respectively. The stake is Re. 1 on a game and no game can be drawn. If the probability of A winning any game is a constant p, find the initial probability of his exhausting the funds of B or his own. Also show that if the resources of B are unlimited then

(i) A is certain to be ruined if $p = \frac{1}{2}$, and

(ii) A has an even chance of escaping ruin if $p = 2^{1/a}/(1 + 2^{1/a})$.

Hint. Let u_n be the probability of A's final win when he has Rs. n.

Then $u_n = pu_{n+1} + (1 - p)u_{n-1}$ where $u_0 = 0$ and $u_{a+b} = 1$

$$\therefore \quad u_{n+1} - u_n = \left(\frac{1-p}{p}\right)(u_n - u_{n-1})$$

Hence $u_{n+1} - u_n = \left(\frac{1-p}{p}\right)^n u_1$, by repeated application,

so that $$u_n = u_1\left[1 - \left(\frac{1-p}{p}\right)^n\right] \Big/ \left[1 - \left(\frac{1-p}{p}\right)\right]$$

Hence using $u_{a+b} = 1$, $$u_n = \left[1 - \left(\frac{1-p}{p}\right)^n\right] \Big/ \left[1 - \left(\frac{1-p}{p}\right)^{a+b}\right]$$

$\therefore$ Initial probability of A's win is $u_a = \dfrac{p^a - (1-p)^a}{p^{a+b} - (1-p)^{a+b}} \cdot p^b$

Probability of A's ruin $= 1 - u_a$.

For $p = \frac{1}{2}$, $u_a = \dfrac{a}{a+b} \to 0$ as $b \to \infty$ and for $p \neq \frac{1}{2}$, $u_a = \frac{1}{2}$ if $p = 2^{1/a}/(1 + 2^{1/a})$.

37. In a game of skill a player has probability 1/3, 5/12 and 1/4 of scoring 0, 1 and 2 points respectively at each trial, the game terminating on the first realization of a zero score at a trial. Assuming that the trials are independent, prove that the probability of the player obtaining a total score of n points is

$$u_n = \frac{3}{13}\left(\frac{3}{4}\right)^n + \frac{4}{39}\left(-\frac{1}{3}\right)^n$$

Hint. Event can materialize in the two mutually exclusive ways:

(i) at the $(n-1)$th trial, a score of $(n-1)$ points is obtained and a score of 1 point is obtained at the nth trial.

(ii) at the $(n-2)$th trial, a score of $(n-2)$ points is obtained and a score of 2 points is obtained at the last two trials.

Hence $u_n = \frac{5}{12}u_{n-1} + \frac{1}{4}u_{n-2}$ where $u_0 = \frac{1}{3}$, $u_1 = \frac{1}{3}\cdot\frac{5}{12} = \frac{5}{36}$

Also $u_n = \left(\frac{3}{4} - \frac{1}{3}\right)u_{n-1} + \frac{1}{4}u_{n-2} \Rightarrow u_n + \frac{1}{3}u_{n-1} = \frac{3}{4}\left(u_{n-1} + \frac{1}{3}u_{n-2}\right)$

This equation can be solved as a homogeneous difference equation of second order with the initial conditions

$$u_0 = \frac{1}{3},\ u_1 = \frac{1}{3}\cdot\frac{5}{12} = \frac{5}{36}$$

38. The following weather forecasting is used by an amateur forecaster. Each day is classified as 'dry' or 'wet' and the probability that any given day is same as the preceding one is assumed to be a constant p, $(0 < p < 1)$. Based on past records, it is supposed that January 1 has a probability β of being dry. Letting

β_n = Probability that nth day of the year is dry, obtain an expression for β_n in terms of β and p. Also evaluate $\lim_{n\to\infty} \beta_n$.

Hint. $\beta_n = p\,.\,\beta_{n-1} + (1-p)(1-\beta_{n-1})$

$\Rightarrow$ $\beta_n = (2p-1)\beta_{n-1} + (1-p)\ ;\ n = 2, 3, 4, \ldots$

Ans. $\beta_n = (2p-1)^{n-1}(\beta - \frac{1}{2}) + \frac{1}{2}\ ;\ \lim_{n\to\infty} \beta_n = \frac{1}{2}$

39. Two urns contain respectively 'a white and b black' and 'b white and a black' balls. A series of drawings is made according to the following rules:

(i) Each time only one ball is drawn and immediately returned to the same urn it came from.

(ii) If the ball drawn is white, the next drawing is made from the first urn.

(iii) If it is black, the next drawing is made from the second urn.

(iv) The first ball drawn comes from the first urn.

What is the probability that nth ball drawn will be white?

Hint. $p_r = P$ [Drawing a white ball at the rth draw].

$$p_r = \frac{a}{a+b}p_{r-1} + \frac{b}{a+b}\cdot\left(1 - p_{r-1}\right)$$

$\Rightarrow$ $$p_r = \frac{a-b}{a+b}\cdot p_{r-1} + \frac{b}{a+b}$$

Ans. $$p_n = \frac{1}{2} + \frac{1}{2}\left(\frac{a-b}{a+b}\right)^n$$

40. If a coin is tossed repeatedly, show that the probability of getting m heads before n tails is :

$$\frac{1}{2^{m+n-1}} \sum_{i=m}^{m+n-1} {}^{m+n-1}C_i.$$

[Burdwan Univ. (Maths Hons.), 1991]

OBJECTIVE TYPE QUESTIONS

I. Find out the correct answer from group Y for each item of group X.

Group X	*Group Y*
(a) At least one of the events A or B occurs.	*(i)* $(\bar{A} \cap B) \cup (A \cap \bar{B}) \cup (\bar{A} \cap \bar{B})$
(b) Neither A nor B occurs.	*(ii)* $(A \cup B) - (A \cap B)$
(c) Exactly one of the events A or B occurs.	*(iii)* $A \subset B$
(d) If event A occurs, so does B.	*(iv)* $B \subset A$
(e) Not more than one of the events A or B occur.	*(v)* $[A - (A \cap B)] \cup [B - (A \cap B)]$
	(vi) $A \cap \bar{B}$
	(vii) $1 - (A \cup \bar{B})$
	(viii) $A \cup B$
	(ix) $1 - (A \cup B)$

II. Match the correct expression of probabilities on the left :

(a) $P(\phi)$, where ϕ is null set	*(i)* $1 - P(A)$
(b) $P(A \mid B) P(B)$	*(ii)* $P(A \cap B)$
(c) $P(\bar{A})$	*(iii)* $P(A) - P(A \cap B)$
(d) $P(\bar{A} \cap \bar{B})$	*(iv)* 0
(e) $P(A \sim B)$	*(v)* $1 - P(A) - P(B) + P(A \cap B)$
	(vi) $P(A) + P(B) - P(A \cap B)$.

III. Given that A, B and C are mutually exclusive events, explain why the following are not permissible assignments of probabilities:

(i) $P(A) = 0{\cdot}24$, $P(B) = 0{\cdot}4$ and $P(A \cup C) = 0{\cdot}2$

(ii) $P(A) = 0{\cdot}4$, $P(B) = 0{\cdot}61$

(iii) $P(A) = 0{\cdot}6$, $P(A \cap \bar{B}) = 0{\cdot}5$

IV. In each of the following, indicate whether events A and B are : *(i)* independent, *(ii)* mutually exclusive, *(iii)* dependent but not mutually exclusive.

(a) $P(A \cap B) = 0$ *(b)* $P(A \cap B) = 0{\cdot}3$, $P(A) = 0{\cdot}45$

(c) $P(A \cup B) = 0{\cdot}85$, $P(A) = 0{\cdot}3$, $P(B) = 0{\cdot}6$

(d) $P(A \cup B) = 0{\cdot}70$, $P(A) = 0{\cdot}5$, $P(B) = 0{\cdot}4$

(e) $P(A \cup B) = 0{\cdot}90$, $P(A \mid B) = 0{\cdot}8$, $P(B) = 0{\cdot}5$.

V. Give the correct label as answer like *a* or *b* etc., for the following questions:

(i) The probability of drawing any one spade card from a pack of cards is

(a) $\frac{1}{52}$ *(b)* $\frac{1}{13}$ *(c)* $\frac{4}{13}$ *(d)* $\frac{1}{4}$

(ii) The probability of drawing one white ball from a bag containing 6 red, 8 black, 10 yellow and 1 green balls is

(a) $\frac{1}{25}$ *(b)* 0 *(c)* 1 *(d)* $\frac{24}{25}$ *(e)* $\frac{15}{20}$

(iii) A coin is tossed three times in succession, the number of sample points in sample space is

(a) 6 *(b)* 8 *(c)* 3

(iv) In the simultaneous tossing of two perfect coins, the probability of having at least one head is

(a) $\frac{1}{2}$ *(b)* $\frac{1}{4}$ *(c)* $\frac{3}{4}$ *(d)* 1

(v) In the simultaneous tossing of two perfect dice, the probability of obtaining 4 as the sum of the resultant faces is

(a) $\frac{4}{12}$ *(b)* $\frac{1}{12}$ *(c)* $\frac{3}{12}$ *(d)* $\frac{2}{12}$

(vi) A single letter is selected at random from the word 'probability'. The probability that it is a vowel is

(a) $\frac{3}{11}$ *(b)* $\frac{2}{11}$ *(c)* $\frac{4}{11}$ *(d)* 0

(vii) An urn contains 9 balls, two of which are red, three blue and four black. Three balls are drawn at random. The chance that they are of the same colour is

(a) $\frac{5}{84}$ *(b)* $\frac{3}{9}$ *(c)* $\frac{3}{7}$ *(d)* $\frac{7}{17}$

(viii) A number is chosen at random among the first 120 natural numbers. The probability of the number chosen being a multiple of 5 or 15 is

(a) $\frac{1}{5}$ *(b)* $\frac{1}{8}$ *(c)* $\frac{1}{16}$

(ix) If A and B are mutually exclusive events, then

(a) $P(A \cup B) = P(A) \cdot P(B)$

(b) $P(A \cup B) = P(A) + P(B)$, *(c)* $P(A \cup B) = 0$.

(x) If A and B are two independent events, the probability that both A and B occur is $\frac{1}{8}$ and the probability that neither of them occurs is $\frac{3}{8}$. The probability of the occurrence of A is :

(a) $\frac{1}{2}$, *(b)* $\frac{1}{3}$, *(c)* $\frac{1}{4}$, *(d)* $\frac{1}{5}$.

VI. Fill in the blanks :

(i) Two events are said to be equally likely if .....

(ii) A set of events is said to be independent if .....

(iii) If $P(A) \cdot P(B) \cdot P(C) = P(A \cap B \cap C)$, then the events A, B, C are .....

(iv) Two events A and B are mutually exclusive if $P(A \cap B) = \ldots$ and are independent if $P(A \cap B) = \ldots$

(v) The probability of getting a multiple of 2 in a throw of a dice is 1/2 and of getting a multiple of 3 is 1/3. Hence probability of getting a multiple of 2 or 3 is .....

(vi) Let A and B be independent events and suppose the event C has probability 0 or 1. Then A, B and C are ..... events.

(vii) If A, B, C are pairwise independent and A is independent of $B \cup C$, then A, B, C are ..... independent.

(viii) A man has tossed 2 fair dice. The conditional probability that he has tossed two sixes, given that he has tossed at least one six is .....

(ix) Let A and B be two events such that $P(A) = 0{\cdot}3$ and $P(A \cup B) = 0{\cdot}8$. If A and B are independent events then $P(B) = \ldots$

VII. Each of following statements is either true or false. If it is true prove it, otherwise, give a counter example to show that it is false.

(i) The probability of occurrence of at least one of two events is the sum of the probability of each of the two events.

(ii) Mutually exclusive events are independent.

(iii) For any two events A and B, $P(A \cap B)$ cannot be less than either $P(A)$ or $P(B)$.

(iv) The conditional probability of A given B is always greater than $P(A)$.

(v) If the occurrence of an event A implies the occurrence of another event B then $P(A)$ cannot exceed $P(B)$.

(vi) For any two events A and B, $P(A \cup B)$ cannot be greater then either $P(A)$ or $P(B)$.

(vii) Mutually exclusive events are not independent.

(viii) Pairwise independence does not necessarily imply mutual independence.

(ix) Let A and B be events neither of which has probability zero. Then if A and B are disjoint, A and B are independent.

(x) The probability of any event is always a proper fraction.

(xi) If $0 < P(B) < 1$ so that $P(A \mid B)$ and $P(A \mid \overline{B})$ are both defined, then $P(A) = P(B)\,P(A \mid B) + P(\overline{B})\,P(A \mid \overline{B})$.

(xii) For two events A and B if

$P(A) = P(A \mid B) = 1/4$ and $P(A \mid \overline{B}) = 1/2$, then

(a) A and B are mutually exclusive.

(b) A and B are independent.

(c) A is a sub-event of B.

(d) $P(\overline{A} \mid B) = 3/4$. **[Delhi Univ. B.Sc.(Stat. Hons.), 1992]**

(xiii) Two events can be independent and mutually exclusive simultaneously.

(xiv) Let A and B be events, neither of which has probability zero. Prove or disprove the following :

(a) If A and B are disjoint, A and B are independent.

(b) If A and B are independent, A and B are disjoint.

(xv) If $P(A) = 0$, then $A = \phi$.

Random Variables — Distribution Functions

5·1. Random Variable. Intuitively by a *random variable (r.v)* we mean a real number X connected with the outcome of a random experiment E. For example, if E consists of two tosses of a coin, we may consider the random variable which is the number of heads (0, 1 or 2).

Outcome :	*HH*	*HT*	*TH*	*TT*
Value of X :	2	1	1	0

Thus to each outcome ω, there corresponds a real number $X(\omega)$. Since the points of the sample space S correspond to outcomes, this means that a real number, which we denote by $X(\omega)$, is defined for each $\omega \in S$. From this standpoint, we define random variable to be a real function on S as follows:

" Let S be the sample space associated with a given random experiment. A real-valued function defined on S and taking values in R $(-\infty, \infty)$ is called a one-dimensional random variable. If the function values are ordered pairs of real numbers (i.e., vectors in two-space) the function is said to be a two- dimensional random variable. More generally, an n-dimensional random variable is simply a function whose domain is S and whose range is a collection of n-tuples of real numbers (vectors in n- space)."

For a mathematical and rigorous definition of the random variable, let us consider the probability space, the triplet (**S, B, P**), where S is the sample space, *viz.*, space of outcomes, **B** is the σ-field of subsets in S, and **P** is a probability function on **B**.

Def. A random variable (r.v.) is a function $X(\omega)$ with domain S and range $(-\infty, \infty)$ such that for every real number a, the event $[\omega : X(\omega) \leq a] \in \mathbf{B}$.

Remarks: 1. The refinement above is the same as saying that the function $X(\omega)$ is *measurable* real function on $(S, \mathbf{B})$.

2. We shall need to make probability statements about a random variable X such as $P\{X \leq a\}$. For the simple example given above we should write $P\{X \leq 1\} = P\{HH, HT, TH\} = 3/4$. That is, $P(X \leq a)$ is simply the probability of the set of outcomes ω for which $X(\omega) \leq a$ or

$$P(X \leq a) = P\{\omega : X(\omega) \leq a\}$$

Since **P** is a measure on $(S,\mathbf{B})$ *i.e.*, **P** is defined on subsets of **B**, the above probability will be defined *only* if $\{\omega : X(\omega) \leq a\} \in \mathbf{B}$, which implies that $X(\omega)$ is a measurable function on $(S,\mathbf{B})$.

3. One-dimensional random variables will be denoted by capital letters, $X, Y, Z, \ldots$etc. A typical outcome of the experiment (*i.e.*, a typical element of the sample space) will be denoted by ω or e. Thus $X(\omega)$ represents the real number which the random variable X associates with the outcome ω. The values which $X, Y, Z, \ldots$ etc., can assume are denoted by lower case letters *viz.*, $x, y, z, \ldots$ etc.

4. *Notations.* If x is a real number, the set of all ω in S such that $X(\omega) = x$ is denoted briefly by writing $X = x$. Thus

$$P(X = x) = P\{\omega : X(\omega) = x\}$$

Similarly $$P(X \leq a) = P\{\omega : X(\omega) \in [-\infty, a]\}$$

and $$P[a < X \leq b) = P(\omega : X(\omega) \in (a, b])$$

Analogous meanings are given to

$$P(X = a \text{ or } X = b) = P\{(X = a) \cup (X = b)\},$$

$$P(X = a \text{ and } X = b) = P\{(X = a) \cap (X = b)\}, \quad \text{etc.}$$

Illustrations : 1. If a coin is tossed, then

$$S = \{\omega_1, \omega_2\} \text{ where } \omega_1 = H, \ \omega_2 = T$$

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = H \\ 0, & \text{if } \omega = T \end{cases}$$

$X(\omega)$ is a Bernoulli random variable. Here $X(\omega)$ takes only two values. A random variable which takes only a finite number of values is called *single*.

2. An experiment consists of rolling a die and reading the number of points on the upturned face. The most natural random variable X to consider is

$$X(\omega) = \omega \ ; \omega = 1, 2, ..., 6$$

If we are interested in whether the number of points is even or odd, we consider a random variable Y defined as follows :

$$Y(\omega) = \begin{cases} 0, & \text{if } \omega \text{ is even} \\ 1, & \text{if } \omega \text{ is odd} \end{cases}$$

3. If a dart is thrown at a circular target, the sample space S is the set of all points w on the target. By imagining a coordinate system placed on the target with the origin at the centre, we can assign various random variables to this experiment. A natural one is the two dimensional random variable which assigns to the point ω, its rectangular coordinates (x,y). Another is that which assigns ω its polar coordinates (r, θ). A one dimensional random variable assigns to each ω only one of the coordinates x or y (for cartesian system), r or θ (for polar system). The event **E**, "that the dart will land in the first quadrant" can be described by a random variable which assigns to each point w its polar coordinate θ so that $X(\omega) = \theta$ and then $\mathbf{E} = \{\omega : 0 \leq X(\omega) \leq \pi/2\}$.

4. If a pair of fair dice is tossed then $S = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$ and $n(S) = 36$. Let X be a random variable with image set

$$X(S) = \{1,2,3,4,5,6\}$$

$$P(X = 1) = P\{1,1\} = 1/36$$

$$P(X = 2) = P\{(2,1),(2,2),(1,2)\} = 3/36$$

$$P(X = 3) = P\{(3,1),(3,2),(3,3),(2,3),(1,3)\} = 5/36$$

$P(X = 4) = P[(4,1), (4,2), (4,3), (4,4), (3,4), (2,4), (1,4)] = 7/36$

Similarly $P(X = 5) = 9/36$ and $P(X = 6) = 11/36$

Some theorems on Random Variables. Here we shall state (without proof) some of the fundamental results and theorems on random variables.

Theorem 5·1. A function $X(\omega)$ from S to R $(-\infty, \infty)$ is a random variable if and only if

$$\{\omega : X(\omega) < a\} \in \mathbf{B}$$

Theorem 5·2. If X_1 and X_2 are random variables and C is a constant then CX_1, $X_1 + X_2$, X_1X_2 are also random variables.

Remark. It will follow that $C_1X_1 + C_2X_2$ is a random variable for constants C_1 and C_2. In particular $X_1 - X_2$ is a r.v.

Theorem 5·3. If $\{X_n(\omega), n \geq 1\}$ are random variables then

$\sup_n X_n(\omega)$, $\inf_n X_n(\omega)$, $\lim_{n\to\infty} \sup X_n(\omega)$ and $\lim_{n\to\infty} \inf X_n(\omega)$ are all random variables, whenever they are finite for all ω.

Theorem 5·4. If X is a random variable then

(i) $\frac{1}{X}$ *where* $\left(\frac{1}{X}\right)(\omega) = \infty$ if $X(\omega) = 0$

(ii) $X_+(\omega) = \max [0, X(\omega)]$

(iii) $X_-(\omega) = -\min [0, X(\omega)]$

(iv) $|X|$

are random variables.

Theorem 5·5. If X_1 and X_2 are random variables then

(i) max $[X_1, X_2]$ and *(ii)* min $[X_1, X_2]$ are also random variables.

Theorem 5·6. If X is a r.v. and $f(\cdot)$ is a continuous function, then $f(X)$ is a r.v.

Theorem 5·7. If X is a r.v. and $f(\cdot)$ is an increasing function, then $f(X)$ *is a r.v.*

Corollary. If f is a function of bounded variations on every finite interval [a,b], and X is a r.v. then $f(X)$ is a r.v.

(proofs of the above theorems are beyond the scope of this book)

EXERCISE 5 (a)

1. Let X be a one dimensional random variable. *(i)* If $a < b$, show that the two events $a < X \leq b$ and $X \leq a$ are disjoint, *(ii)* Determine the union of the two events in part *(i)*, *(iii)* show that $P(a < X \leq b) = P(X \leq b) - P(X \leq a)$.

2. Let a sample space S consist of three elements ω_1, ω_2, and ω_3. Let $P(\omega_1) = 1/4$, $P(\omega_2) = 1/2$ and $P(\omega_3) = 1/4$. If X is a random variable defined on S by $X(\omega_1) = -10$, $X(\omega_2) = -3$, $X(\omega_3) = 15$, find $P(-2 \leq X \leq 2)$.

3. Let $S = (e_1, e_2, ..., e_n)$ be the sample space of some experiment and let $E \subseteq S$ be some event associated with the experiment.

Define ψ_E, the *characteristic random variable* of E as follows :

$$\psi_E(e_i) = \begin{cases} 1 & \text{if } e_i \in E \\ 0 & \text{if } e_i \notin E. \end{cases}$$

In other words, ψ_E is equal to 1 if E occurs, and ψ_E is equal to 0 if E does not occur.

Verify the following properties of characteristic random variables :

(i) ψ_ϕ *is identically zero , i.e.,* $\psi_\phi(e_i) = 0$; $i = 1, 2, ..., n$

(ii) ψ_S *is identically one , i.e.,* $\psi_S(e_i) = 1$; $i = 1, 2, ..., n$

(iii) $E = F \Rightarrow \psi_E(e_i) = \psi_F(e_i)$; $i = 1, 2, ..., n$ and *conversely*

(iv) If $E \subseteq F$ *then* $\psi_E(e_i) \leq \psi_F(e_i)$; $i = 1, 2, ..., n$

(v) $\psi_E(e_i) + \psi_{\bar{E}}(e_i)$ *is identically* 1 : $i = 1, 2, ..., n$

(vi) $\psi_{E \cap F}(e_i) = \psi_E(e_i)\,\psi_F(e_i)$; $i = 1, 2, ..., n$

(vii) $\psi_{E \cup F}(e_i) = \psi_E(e_i) + \psi_F(e_i) - \psi_E(e_i)\,\psi_F(e_i)$, for $i = 1, 2, ..., n$.

5.2. Distribution Function. Let X be a r.v. on (S,B,P). Then the function :

$$F_X(x) = P(X \leq x) = P\{\omega : X(\omega) \leq x\}, \quad -\infty < x < \infty$$

is called the distribution function (d.f.) of X.

If clarity permits, we may write $F(x)$ instead of $F_X(x)$. ...(5·1)

5·2·1. Properties of Distribution Function. We now proceed to derive a number of properties common to all distribution functions.

Property 1. *If F is the d.f. of the r.v. X and if $a < b$, then*

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

Proof. The events '$a < X \leq b$' and '$X \leq a$' are disjoint and their union is the event '$X \leq b$'. Hence by addition theorem of probability

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

$$\Rightarrow P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \quad ...(5·2)$$

Cor. 1.

$$P(a \leq X \leq b) = P\{(X = a) \cup (a < X \leq b)\}$$

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

(using additive property of P)

$$= P(X = a) + [F(b) - F(a)] \quad ...(5·2\,a)$$

Similarly, we get

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

$$= F(b) - F(a) - P(X = b) \quad ...(5·2\,b)$$

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

$$= F(b) - F(a) - P(X = b) + P(X = a) \quad ...(5\cdot2\,c)$$

Remark. When $P(X = a) = 0$ and $P(X = b) = 0$, all four events $a \leq X \leq b$. $a < X < b$, $a \leq X < b$ and $a < X \leq b$ have the same probability $F(b) - F(a)$.

Property 2. *If F is the d.f. of one-dimensional r.v. X, then (i) $0 \leq F(x) \leq 1$, (ii) $F(x) \leq F(y)$ if $x < y$.*

In other words, all distribution functions are monotonically non-decreasing and lie between 0 and 1.

Proof. Using the axioms of certainty and non-negativity for the probability function **P**, part *(i)* follows triviality from the definition of $F(x)$.

For part *(ii)*, we have for $x < y$,

$$F(y) - F(x) = P(x < X \leq y) \geq 0 \qquad \text{(Property 1)}$$

$$\Rightarrow \quad F(y) \geq F(x)$$

$$\Rightarrow \quad F(x) \leq F(y) \text{ when } x < y \qquad ...(5\cdot3)$$

Property 3. *If F is d.f. of one-dimensional r.v. X, then*

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0$$

and

$$F(\infty) = \lim_{x \to \infty} F(x) = 1$$

Proof. Let us express the whole sample space S as a countable union of disjoint events as follows :

$$S = \left[\bigcup_{n=1}^{\infty} (-n < X \leq -n+1)\right] \cup \left[\bigcup_{n=0}^{\infty} (n < X \leq n+1)\right]$$

$$\Rightarrow P(S) = \sum_{n=1}^{\infty} P(-n < X \leq -n+1) + \sum_{n=0}^{\infty} P(n < X \leq n+1)$$

($\because$ **P** is additive)

$$\Rightarrow \quad 1 = \lim_{a \to \infty} \sum_{n=1}^{a} [F(-n+1) - F(-n)]$$

$$+ \lim_{b \to \infty} \sum_{n=0}^{b} [F(n+1) - F(n)]$$

$$= \lim_{a \to \infty} [F(0) - F(-a)] + \lim_{b \to \infty} [F(b+1) - F(0)]$$

$$= [F(0) - F(-\infty)] + [F(\infty) - F(0)]$$

$$\therefore \quad 1 = F(\infty) - F(-\infty) \qquad ...(*)$$

Since $-\infty < \infty$, $F(-\infty) \leq F(\infty)$. Also

$$F(-\infty) \geq 0 \text{ and } F(\infty) \leq 1 \qquad \text{(Property 2)}$$

$$\therefore \quad 0 \le F(-\infty) \le F(\infty) \le 1 \qquad (**)$$

(*) and (**) give $F(-\infty) = 0$ and $F(\infty) = 1$.

Remarks. 1. Discontinuities of $F(x)$ are at most countable.

2. $$F(a) - F(a-0) = \lim_{h \to 0} P(a-h \le X \le a), \; h > 0$$

$$\therefore \quad F(a) - F(a-0) = P(X = a)$$

and $$F(a+0) - F(a) = \lim_{h \to 0} P(a \le X \le a+h) = 0, \; h > 0$$

$$\Rightarrow \quad F(a+0) = F(a)$$

5·3. Discrete Random Variable. If a random variable takes at most a countable number of values, it is called a discrete random variable. *In other words, a real valued function defined on a discrete sample space is called a discrete random variable.*

5·3·1. Probability Mass Function (*and probability distribution of a discrete random variable*).

Suppose X is a one-dimensional discrete random variable taking at most a countably infinite number of values $x_1, x_2, \ldots$ With each possible outcome x_i, we associate a number $p_i = P(X = x_i) = p(x_i)$, called the probability of x_i. The numbers $p(x_i)$; $i = 1, 2, \ldots$ must satisfy the following conditions :

$$(i) \quad p(x_i) \ge 0 \;\; \forall \; i, \quad (ii) \quad \sum_{i=1}^{\infty} p(x_i) = 1$$

This function p is called the probability mass function of the random variable X and the set $\{x_i, p(x_i)\}$ is called the probability distribution (p.d.) of the r.v. X.

Remarks: 1. The set of values which X takes is called the *spectrum* of the random variable.

2. For discrete random variable, a knowledge of the probability mass function enables us to compute probabilities of arbitrary events. In fact, if E is a set of real numbers, we have

$$P(X \in E) = \sum_{x \in E \cap S} p(x), \text{ where } S \text{ is the sample space.}$$

Illustration. Toss of coin, $S = \{H, T\}$. Let X be the random variable defined by

$$X(H) = 1, \; i.e., \; X = 1, \text{ if 'Head' occurs.}$$
$$X(T) = 0, \; i.e., \; X = 0, \text{ if 'Tail' occurs.}$$

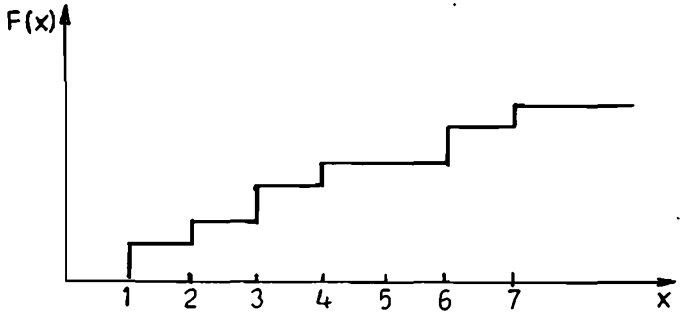
If the coin is 'fair' the probability function is given by

$$P(\{H\}) = P(\{T\}) = \tfrac{1}{2}$$

and we can speak of the probability distribution of the random variable X as

$$P(X = 1) = P(\{H\}) = \tfrac{1}{2},$$
$$P(X = 0) = P(\{T\}) = \tfrac{1}{2},$$

5·3·2. Discrete Distribution Function. In this case there are a countable number of points $x_1, x_2, x_3, \ldots$ and numbers $p_i \geq 0$, $\sum_{1}^{\infty} p_i = 1$ such that $F(X) = \sum_{(i : x_i \leq x)} p_i$. For example if x_i is just the integer i, $F(x)$ is a "*step function*" having jump p_i at i, and being constant between each pair of integers.



Theorem 5·5. $p(x_j) = P(X = x_j) = F(x_j) - F(x_{j-1})$, where F is the *d.f.* of X.

Proof. Let $x_1 < x_2 < \ldots$ We have

$$F(x_j) = P(X \leq x_j)$$

$$= \sum_{i=1}^{j} P(X = x_i) = \sum_{i=1}^{j} p(x_i)$$

and $$F(x_{j-1}) = P(X \leq x_{j-1}) = \sum_{i=1}^{j-1} p(x_i)$$

$$\therefore \quad F(x_j) - F(x_{j-1}) = p(x_j) \qquad \ldots(5·5)$$

Thus, given the distribution function of discrete random variable, we can compute its probability mass function.

Example 5·1. *An experiment consists of three independent tosses of a fair coin. Let*

X = The number of heads

Y = The number of head runs,

Z = The lenght of head runs,

a head run being defined as consecutive occurrence of at least two heads, its length then being the number of heads occurring together in three tosses of the coin.

Find the probability function of (i) X, (ii) Y, (iii) Z, (iv) X+Y and (v) XY and construct probability tables and draw their probability charts.

Solution.

Table 1

S. No.	*Elementary event*	*Random Variables*				
		X	*Y*	*Z*	*X+Y*	*XY*
1	*HHH*	3	1	3	4	3
2	*HHT*	2	1	2	3	2
3	*HTH*	2	0	0	2	0
4	*HTT*	1	0	0	1	0
5	*THH*	2	1	2	3	2
6	*THT*	1	0	0	1	0
7	*TTH*	1	0	0	1	0
8	*TTT*	0	0	0	0	0

Here sample space is.

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

(i) Obviously X is a r.v. which can take the values 0, 1, 2, and 3

$$p(3) = P(HHH) = (1/2)^3 = 1/8$$

$$p(2) = P[HHT \cup HTH \cup THH]$$

$$= P(HHT) + P(HTH) + P(THH) = 1/8 + 1/8 + 1/8 = 3/8$$

Similarly $p(1) = 3/8$ and $p(0) = 1/8$.

These probabilities could also be obtained directly from the above table 1.

Table 2

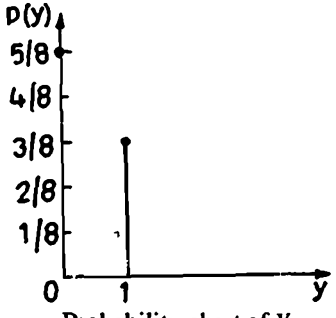
Probability table of X

Values of X (x)	0	1	2	3
$p(x)$	1/8	3/8	3/8	1/8

Table 3

(ii) **Probability Table of Y**

Values of Y, (y)	0	1
$p(y)$	5/8	3/8



Probability chart of Y

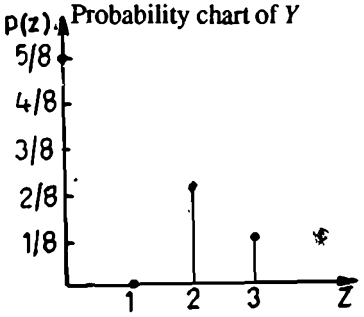
This is obvious from table 1.

(iii) From table 1 , we have

Table 4

Probability Table of

Values of Z, (z)	0	1	2	3
$p(z)$	5/8	0	2/8	1/8



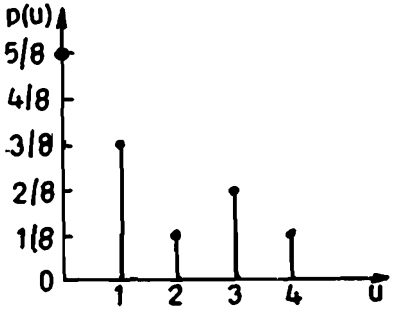
Probability chart of Z

(iv) Let $U = X + Y$. From table 1, we get

Table 5

Probability Table of U

Values of U, (u)	0	1	2	3	4
$p(u)$	1/8	3/8	1/8	2/8	1/8



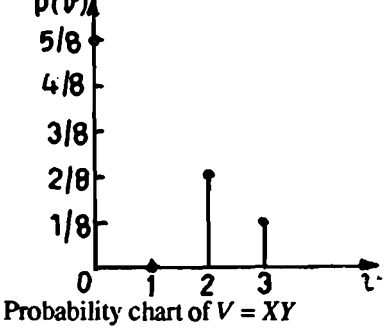
Probability chart of $U = X+Y$

(v) Let $V = XY$

Table 6

Probability Table of V

Values of V, (v)	0	1	2	3
$p(v)$	5/8	0	2/8	1/8



Probability chart of $V = XY$

Example 5·2. *A random variable X has the following probability distribution :*

x:	0	1	2	3	4	5	6	7
$p(x)$:	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$	$7k^2 + k$

(i) Find k, (ii) Evaluate $P(X < 6)$, $P(X \geq 6)$, *and* $P(0 < X < 5)$, *(iii) If* $P(X \leq c) > \frac{1}{2}$, *find the minimum value of c, and (iv) Determine the distribution function of X.* **[Madurai Univ. B.Sc., Oct. 1988]**

Solution. Since $\sum_{x=0}^{7} p(x) = 1$, we have

$$\Rightarrow \quad k + 2k + 2k + 3k + k^2 + 2k^2 + 7k^2 + k = 1$$

$$\Rightarrow \quad 10k^2 + 9k - 1 = 0$$

$$\Rightarrow \quad (10k - 1)(k + 1) = 0 \Rightarrow k = 1/10$$

[$\because k = -1$, is rejected, since probability canot be negative.]

(ii) $P(X < 6) = P(X = 0) + P(X = 1) + \ldots + P(X = 5)$

$$= \frac{1}{10} + \frac{2}{10} + \frac{2}{10} + \frac{3}{10} + \frac{1}{100} = \frac{81}{100}$$

$$P(X \geq 6) = 1 - P(X < 6) = \frac{19}{100}$$

$P(0 < X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 8k = 4/5$

(iii) $P(X \leq c) > \frac{1}{2}$. By trial, we get $c = 4$.

(iv)

X	$F_X(x) = P(X \leq x)$
0	0
1	$k = 1/10$
2	$3k = 3/10$
3	$5k = 5/10$
4	$8k = 4/5$
5	$8k + k^2 = 81/100$
6	$8k + 3k^2 = 83/100$
7	$9k + 10k^2 = 1$

EXERCISE 5 (b)

1. *(a)* A student is to match three historical events (Mahatma Gandhi's Birthday, India's freedom, and First World War) with three years (1947, 1914, 1896). If he guesses with no knowledge of the correct answers, what is the probability distribution of the number of answers he gets correctly ?

(b) From a lot of 10 items containing 3 defectives, a sample of 4 items is drawn at random. Let the random variable X denote the number of defective items in the sample. Answer the following when the sample is drawn without replacement.

(i) Find the probability distribution of X,

(ii) Find $P(X \le 1)$, $P(X < 1)$ and $P(0 < X < 2)$

Ans. *(a)*

x	0	1	2	3
$p(x)$	$\frac{1}{3}$	$\frac{1}{2}$	0	$\frac{1}{6}$

(b) *(i)*

x	0	1	2	3
$p(x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{30}$

(ii) 2/3, 5/6, 1/2

2. *(a)* A random variable X can take all non-negative integral values, and the probability that X takes the value r is proportional to α^r $(0 < \alpha < 1)$. Find $P(X = 0)$. **[Calcutta Univ. B.Sc. 1987]**

Ans. $P(X = r) = A\alpha^r$; $r = 0, 1, 2, \ldots$; $A = 1 - \alpha$; $P(X = 0) = A = 1 - \alpha$

(b) Suppose that the random variable X has possible values 1, 2, 3, ... and $P(X = j) = 1/2^j$, $j = 1, 2, \ldots$ *(i)* Compute $P(X$ *is even*$)$, *(ii)* Compute $P(X \ge 5)$, and *(iii)* Compute $P(X$ *is divisible by* $3)$.

Ans. *(i)* 1/3, *(ii)* 1/16, and *(iii)* 1/7

3. *(a)* Let X be a random variable such that

$$P(X = -2) = P(X = -1),\ P(X = 2) = P(X = 1) \text{ and}$$
$$P(X > 0) = P(X < 0) = P(X = 0).$$

Obtain the probability mass function of X and its distribution function.

Ans.

X :	−2	−1	0	1	2
$p(x)$:	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$
$F(x)$:	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

(b) A random variable X assumes the values −3, −2, −1, 0, 1, 2, 3 such that

$$P(X = -3) = P(X = -2) = P(X = -1),$$
$$P(X = 1) = P(X = 2) = P(X = 3),$$

and $$P(X = 0) = P(X > 0) = P(X < 0),$$

Obtain the probability mass function of X and its distribution function, and find further the probability mass function of $Y = 2X^2 + 3X + 4$.

[Poona Univ. B.Sc., March 1991]

Ans.

X :	−3	−2	−1	0	1	2	3
$p(x)$:	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
Y :	13	6	3	4	9	18	31
$p(y)$:	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$

4. *(a)* A random variable X has the following probability function :

Values of X, x :	-2	-1	0	1	2	3
$p(x)$:	0.1	k	0.2	$2k$	0.3	k

(i) Find the value of k, and calculate mean and variance.

(ii) Construct the c.d.f. $F(X)$ and draw its graph.

Ans. *(i)* 0·1, 0·8 and 2·16, *(ii)* $F(X) = 0.1, 0.2, 0.4, 0.6, 0.9, 1.0$

(b) Given the probability function

x	0	1	2	3
$p(x)$	0·1	0·3	0·5	0·1

Let $Y = X^2 + 2X$, then find *(i)* the probability function of Y, *(ii)* mean and variance of Y.

Ans. *(i)*

y	0	3	8	15
$p(y)$	0·1	0·3	0·5	0·1

, *(ii)* 6·4 ,16·24

5. A random variable X has the following probability distribution :

Values of X, x	0	1	2	3	4	5	6	7	8
$p(x)$	a	$3a$	$5a$	$7a$	$9a$	$11a$	$13a$	$15a$	$17a$

(i) Determine the value of a.

(ii) Find $P(X < 3), P(X \geq 3), P(0 < X < 5)$.

(iii) What is the smallest value of x for which $P(X \leq x) > 0.5$? and

(iv) Find out the distribution function of X ?

Ans. *(i)* $a = 1/81$, *(ii)* 9/81, 72/81, 24/81, *(iii)* 6

(iv)

x	0	1	2	3	4	5	6	7	8
$F(x)$	a	$4a$	$9a$	$16a$	$25a$	$36a$	$49a$	$64a$	$81a$

6. *(a)* Let $p(x)$ be the probability function of a discrete random variable X which assumes the values x_1, x_2, x_3, x_4, such that $2p(x_1) = 3p(x_2) = p(x_3) = 5p(x_4)$. Find probability distribution and cumulative probability distribution of X. **(Sardar Patel Univ. B.Sc. 1987)**

Ans.

x	x_1	x_2	x_3	x_4
$p(x)$	15/16	10/16	30/16	6/16

(b) The following is the distribution function of a discrete random variable X :

x	-3	-1	0	1	2	3	5	8
$f(x)$	0·10	0·30	0·45	0·5	0·75	0·90	0·95	1·00

(i) Find the probability distribution of X.

(ii) Find $P(X \text{ is even})$ and $P(1 \leq X \leq 8)$.

(iii) Find $P(X = -3 \mid X < 0)$ and $P(X \geq 3 \mid X > 0)$.

[**Ans.** *(ii)* 0·30, 0·55, *(iii)* 1/3, 5/11].

7. If $p(x) = \frac{x}{15}$; $x = 1, 2, 3, 4, 5$

$= 0$, elsewhere

Find *(i)* $P\{X = 1 \text{ or } 2\}$, and *(ii)* $P\left\{\frac{1}{2} < X < \frac{5}{2} \mid X > 1\right\}$

[Allahabad Univ. B.Sc., April 1992]

Hint. *(i)* $P\{X = 1 \text{ or } 2\} = P(X = 1) + P(X = 2) = \frac{1}{15} + \frac{2}{15} = \frac{1}{5}$

$$(ii)\ P\left\{\frac{1}{2} < X < \frac{5}{2} \mid X > 1\right\} = \frac{P\left\{\left(\frac{1}{2} < X < \frac{5}{2}\right) \cap X > 1\right\}}{P(X > 1)}$$

$$= \frac{P\{(X = 1 \text{ or } 2) \cap X > 1\}}{P(X > 1)} = \frac{P(X = 2)}{1 - P(X = 1)} = \frac{2/15}{1 - (1/15)} = \frac{1}{7}$$

8. The probability mass function of a random variable X is zero except at the points $x = 0, 1, 2$. At these points it has the values $p(0) = 3c^3$, $p(1) = 4c - 10c^2$ and $p(2) = 5c - 1$ for some $c > 0$.

(i) Determine the value of c.

(ii) Compute the following probabilities, $P(X < 2)$ and $P(1 < X \leq 2)$.

(iii) Describe the distribution function and draw its graph.

(iv) Find the largest x such that $F(x) < 1/2$.

(v) Find the smallest x such that $F(x) \geq 1/3$. **[Poona Univ. B.Sc., 1987]**

Ans. *(i)* $\frac{1}{3}$, *(ii)* $\frac{1}{3}, \frac{2}{3}$, *(iv)* 1, *(v)* 1.

9. *(a)* Suppose that the random variable X assumes three values 0,1 and 2 with probabilities $\frac{1}{3}$, $\frac{1}{6}$ and $\frac{1}{2}$ respectively. Obtain the distribution function of X.. **[Gujarat Univ. B.Sc., 1992]**

(b) Given that $f(x) = k(1/2)^x$ is a probability distribution for a random variable which can take on the values $x = 0, 1, 2, 3, 4, 5, 6$, find k and find an expression for the corresponding cumulative probabilities $F(x)$.

[Nagpur Univ. B.Sc., 1987]

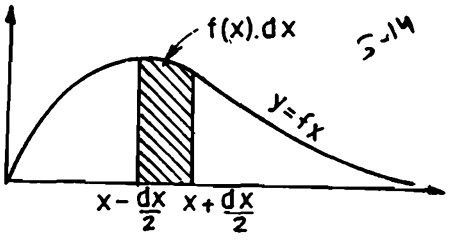
5·4. Continuous Random Variable. A random variable X is said to be continuous if it can take all possible values between certain limits. *In other words, a random variable is said to be continuous when its different values cannot be put in 1-1 correspondence with a set of positive integers.*

A continuous random variable is a random variable that (at least conceptually) can be measured to any desired degree of accuracy. Examples of continuous random variables are age, height, weight etc.

5·4·1. Probability Density Function *(Concept and Definition).* Consider the small interval $(x, x + dx)$ of length dx round the point x. Let $f(x)$ be any continuous

function of x so that $f(x)\,dx$ represents the probability that X falls in the infinitesimal interval $(x, x+dx)$. Symbolically

$$P(x \le X \le x+dx) = f_X(x)\,dx \qquad \ldots (5\cdot5)$$



In the figure, $f(x)\,dx$ represents the area bounded by the curve $y = f(x)$, x-axis and the ordinates at the points x and $x+dx$. The function $f_X(x)$ so defined is known as *probability density function or simply density function of random variable X and is usually abbreviated as p.d.f.* The expression, $f(x)\,dx$, usually written as $dF(x)$, is known as the *probability differential* and the curve $y = f(x)$ is known as the *probability density curve* or simply *probability curve.*

Definition. p.d.f. $f_X(x)$ of the *r.v.* X is defined as :

$$f_X(x) = \lim_{\delta x \to 0} \frac{P(x \le X \le x+\delta x)}{\delta x} \qquad \ldots (5\cdot5\,a)$$

The probability for a variate value to lie in the interval dx is $f(x)\,dx$ and hence the probability for a variate value to fall in the finite interval $[\alpha, \beta]$ is :

$$P(\alpha \le X \le \beta) = \int_{\alpha}^{\beta} f(x)\,dx \qquad \ldots (5\cdot5\,b)$$

which represents the area between the curve $y = f(x)$, x-axis and the ordinates at $x = \alpha$ and $x = \beta$. Further since total probability is unity, we have $\int_a^b f(x)\,dx = 1$, where $[a, b]$ is the range of the random variable X. The range of the variable may be finite or infinite.

The probability density function (*p.d.f.*) of a random variable (*r.v.*) X usually denoted by $f_X(x)$ or simply by $f(x)$ has the following obvious properties

(i) $f(x) \ge 0, \quad -\infty < x < \infty$ $\qquad \ldots (5\cdot5\,c)$

(ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$ $\qquad \ldots (5\cdot5\,d)$

(iii) The probability $P(E)$ given by

$$P(E) = \int_E f(x)\,dx \qquad \ldots (5\cdot5\,e)$$

is well defined for any event E.

Important Remark. In case of discrete random variable, the probability at a point, *i.e.*, $P(x = c)$ is not zero for some fixed c. However, in case of continuous random variables the probability at a point is always zero, *i.e.*, $P(x = c) = 0$ for all possible values of c. This follows directly from (5·5 *b*) by taking $\alpha = \beta = c$.

This also agrees with our discussion earlier that $P(E) = 0$ does not imply that the event E is null or impossible event. This property of continuous r.v., *viz.*,

$$P(X = c) = 0, \quad \forall\ c \qquad \ldots (5{\cdot}5f)$$

leads us to the following important result :

$$P(\alpha \leq X \leq \beta) = P(\alpha \leq X < \beta) = P(\alpha < X \leq \beta) = P(\alpha < X < \beta) \quad \ldots (5{\cdot}5g)$$

i.e., in case of continuous r.v., it does matter whether we include the end points of the interval from α to β.

However, this result is in general not true for discrete random variables.

5·4·2. Various Measures of Central Tendency, Dispersion, Skewness, and Kurtosis for Continuous Probability Distribution. The formulae for these measures in case of discrete frequency distribution can be easily extended to the case of continuous probability distribution by simply replacing $p_i = f_i/N$ by $f(x)\,dx$, x_i by x and the summation over 'i' by integration over the specified range of the variable X.

Let $f_X(x)$ *or* $f(x)$ be the *p.d.f.* of a random variable X where X is defined from a to b. Then

(i) $$\textit{Arithmetic mean} = \int_a^b x\, f(x)\, dx \qquad \ldots (5{\cdot}6)$$

(ii) Harmonic mean. Harmonic mean H is given by

$$\frac{1}{H} = \int_a^b \left(\frac{1}{x}\right) f(x)\, dx \qquad \ldots (5{\cdot}6\,a)$$

(iii) Geometric mean. Geometric mean G is given by

$$\log G = \int_a^b \log x f(x)\, dx \qquad \ldots (5{\cdot}6\,b)$$

(iv) $$\mu_r' \text{ (about origin)} = \int_a^b x^r f(x)\, dx \qquad \ldots (5{\cdot}7)$$

$$\mu_r' \text{ (about the point } x = A) = \int_a^b (x - A)^r f(x)\, dx \qquad \ldots (5{\cdot}7\,a)$$

$$\text{and } \mu_r \text{ (about mean)} = \int_a^b (x - \textit{mean})^r f(x)\, dx \qquad \ldots (5{\cdot}7\,b)$$

In particular, from (5·7), we have

$$\mu_1' \text{ (about origin)} = \textit{Mean} = \int_a^b x\, f(x)\, dx$$

and $$\mu_2' = \int_a^b x^2 f(x)\, dx$$

Hence $$\mu_2 = \mu_2' - \mu_1'^2 = \int_a^b x^2 f(x)\, dx - \left(\int_a^b x f(x)\, dx\right)^2 \qquad \ldots (5{\cdot}7\,c)$$

From (5·7), on putting $r=3$ and 4 respectively, we get the values of μ_3' and μ_4' and consequently the moments about mean can be obtained by using the relations :

and
$$\left.\begin{aligned}\mu_3 &= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 \\ \mu_4 &= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4\end{aligned}\right\} \quad \ldots (5{\cdot}7\,d)$$

and hence β_1 and β_2 can be computed.

(v) Median. Median is the point which divides the entire distribution in two equal parts. In case of continuous distribution, median is the point which divides the total area into two equal parts. Thus if M is the median, then

$$\int_a^M f(x)\,dx = \int_M^b f(x)\,dx = \frac{1}{2} \quad \ldots (5{\cdot}8)$$

Thus solving

$$\int_a^M f(x)\,dx = \frac{1}{2} \quad or \quad \int_M^b f(x)\,dx = \frac{1}{2} \quad \ldots (5{\cdot}8\,a)$$

for M, we get the value of median.

(vi) Mean Deviation. Mean deviation about the mean μ_1' is given by

$$M.D. = \int_a^b |\, x - mean \,|\, f(x)\,dx \quad \ldots (5{\cdot}9)$$

(vii) Quartiles and Deciles. Q_1 and Q_3 are given by the equations

$$\int_a^{Q_1} f(x)\,dx = \frac{1}{4} \quad \text{and} \quad \int_a^{Q_3} f(x)\,dx = \frac{3}{4} \quad \ldots (5{\cdot}10)$$

D_i, i th decile is given by

$$\int_a^{D_i} f(x)\,dx = \frac{i}{10} \quad \ldots (5{\cdot}10\,a)$$

(viii) Mode. Mode is the value of x for which $f(x)$ is maximum. Mode is thus the solution of

$$f'(x) = 0 \text{ and } f''(x) < 0 \quad \ldots (5{\cdot}11)$$

provided it lies in $[a,b]$.

Example 5·3. *The diameter of an electric cable, say X, is assumed to be a continuous random variable with p.d.f. $f(x) = 6x(1-x)$, $0 \le x \le 1$.*

(i) Check that above is p.d.f.,

(ii) Determine a number b such that $P(X < b) = P(X > b)$

[Aligarh Univ. B.Sc. (Hons). 1990]

Solution. Obviously, for $0 \le x \le 1$, $f(x) \ge 0$

Now
$$\int_0^1 f(x)\,dx = 6\int_0^1 x(1-x)\,dx$$

$$= 6\int_0^1 (x - x^2)\,dx = 6\left|\frac{x^2}{2} - \frac{x^3}{3}\right|_0^1 = 1$$

Hence $f(x)$ is the *p.d.f.* of *r.v.* X

(ii) $$P(X < b) = P(X > b) \quad \ldots (*)$$

$$\Rightarrow \quad \int_0^b f(x)\,dx = \int_b^1 f(x)\,dx$$

$$\Rightarrow \quad 6\int_0^b x(1-x)\,dx = 6\int_b^1 x(1-x)\,dx$$

$$\Rightarrow \quad \left|\frac{x^2}{2}-\frac{x^3}{3}\right|_0^b = \left|\frac{x^2}{2}-\frac{x^3}{3}\right|_b^1$$

$$\Rightarrow \quad \left(\frac{b^2}{2}-\frac{b^3}{3}\right) = \left[\left(\frac{1}{2}-\frac{1}{3}\right)-\left(\frac{b^2}{2}-\frac{b^3}{3}\right)\right]$$

$$\Rightarrow \quad 3b^2-2b^3 = [\,1-3b^2+2b^3\,]$$

$$\Rightarrow \quad 4b^3-6b^2+1=0$$

$$(2b-1)(2b^2-2b-1)=0$$

$$\Rightarrow \quad 2b-1=0 \quad or \quad 2b^2-2b-1=0$$

Hence $b = 1/2$ is the only real value lying between 0 and 1 and satisfying (*).

Example 5·4. *A continuous random variable X has a p.d.f.*

$$f(x) = 3x^2,\ 0 \le x \le 1.$$ *Find a and b such that*

(i) $P\{X \le a\} = P\{X > a\}$, and

(ii) $P\{X > b\} = 0{\cdot}05$. **[Calicut Univ. B.Sc., Sept. 1988]**

Solution. *(i)* Since $P(X \le a) = P(X > a)$, each must be equal to 1/2, because total probability is always one.

$$\therefore \quad P(X \le a) = \frac{1}{2} \quad \Rightarrow \quad \int_0^a f(x)\,dx = \frac{1}{2}$$

$$\Rightarrow \quad 3\int_0^a x^2\,dx = \frac{1}{2} \quad \Rightarrow \quad 3\left|\frac{x^3}{3}\right|_0^a = \frac{1}{2}$$

$$\Rightarrow \quad a^3 = \frac{1}{2} \quad \Rightarrow \quad a = \left(\frac{1}{2}\right)^{\frac{1}{3}}$$

(ii) $$P(X > b) = 0{\cdot}05 \quad \Rightarrow \quad \int_b^1 f(x)\,dx = 0{\cdot}05$$

$$\Rightarrow \quad 3\left|\frac{x^3}{3}\right|_b^1 = \frac{1}{20} \quad \Rightarrow \quad 1-b^3 = \frac{1}{20}$$

$$\Rightarrow \quad b^3 = \frac{19}{20} \quad \Rightarrow \quad b = \left(\frac{19}{20}\right)^{\frac{1}{3}}.$$

Example 5·5. *Let X be a continuous random variate with p.d.f.*

$$f(x) = ax,\ 0 \le x \le 1$$
$$= a,\ 1 \le x \le 2$$
$$= -ax + 3a,\ 2 \le x \le 3$$
$$= 0,\ \textit{elsewhere}$$

(i) Determine the constant a.

(ii) Compute $P(X \leq 1{\cdot}5)$. **[Sardar Patel Univ. B.Sc., Nov.1988]**

Solution. *(i)* Constant 'a' is determined from the consideration that total probability is unity, *i.e.*,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\Rightarrow \int_{-\infty}^{0} f(x)\,dx + \int_0^1 f(x)\,dx + \int_1^2 f(x)\,dx + \int_2^3 f(x)\,dx + \int_3^{\infty} f(x)\,dx = 1$$

$$\Rightarrow \int_0^1 ax\,dx + \int_1^2 a.dx + \int_2^3 (-ax + 3a)\,dx = 1$$

$$\Rightarrow a\left|\frac{x^2}{2}\right|_0^1 + a\,|\,x\,|_1^2 + a\left|-\frac{x^2}{2} + 3x\right|_2^3 = 1$$

$$\Rightarrow \frac{a}{2} + a + a\left[\left(-\frac{9}{2} + 9\right) - (-2 + 6)\right] = 1$$

$$\Rightarrow \frac{a}{2} + a + \frac{a}{2} = 1 \quad \Rightarrow \quad 2a = 1 \quad \Rightarrow \quad a = \tfrac{1}{2}$$

(ii) $P(X \leq 1{\cdot}5) = \int_{-\infty}^{1{\cdot}5} f(x)\,dx = \int_{-\infty}^{0} f(x)\,dx + \int_0^1 f(x)\,dx + \int_1^{1{\cdot}5} f(x)\,dx$

$$= a\int_0^1 x\,dx + \int_1^{1{\cdot}5} a.dx$$

$$= a\left|\frac{x^2}{2}\right|_0^1 + a\,|\,x\,|_1^{1{\cdot}5} = \frac{a}{2} + 0{\cdot}5\,a$$

$$= a = \tfrac{1}{2} \qquad [\because a = \tfrac{1}{2},\ \textit{Part (i)}]$$

Example 5·6. *A probability curve* $y = f(x)$ *has a range from 0 to* ∞. *If* $f(x) = e^{-x}$, *find the mean and variance and the third moment about mean.*

[Andhra Univ. B.Sc. 1988; Delhi Univ. B.Sc. Sept. 1987]

Solution.

μ_r' (rth moment about origin) $= \int_0^{\infty} x^r f(x)\,dx$

$$= \int_0^{\infty} x^r e^{-x}\,dx = \Gamma(r+1) = r!$$

(Using Gamma Integral)

Substituting $r = 1, 2$ and 3 successively, we get

Mean $= \mu_1' = 1! = 1,\ \mu_2' = 2! = 2,\ \mu_3' = 3! = 6$

Hence variance $= \mu_2 = \mu_2' - \mu_1'^2 = 2 - 1 = 1$

and $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 6 - 3 \times 2 + 2 = 2$

Example 5·7. *In a continuous distribution whose relative frequency density is given by*

$$f(x) = y_o \, . \, x\,(2-x)\,,\ 0 \le x \le 2,$$

find mean, variance, β_1, and β_2 and hence show that the distribution is symmetrical. Also (i) find mean deviation about mean and (ii) show that for this distribution $\mu_{2n+1} = 0$, (iii) find the mode, harmonic mean and median.

[Delhi Univ. B.Sc.(Stat. Hons.), 1992; B.Sc., Oct. 1992]

Solution. Since total probability is unity, we have

$$\int_0^2 f(x)\,dx = 1$$

$$\Rightarrow \qquad y_o \int_0^2 x\,(2-x)\,dx = 1 \quad \Rightarrow \quad y_o = 3/4$$

$$\therefore \qquad f(x) = \frac{3}{4}\,x\,(2-x)$$

$$\mu_r' = \int_0^2 x^r f(x)\,dx = \frac{3}{4}\int_0^2 x^{r+1}\,(2-x)\,dx = \frac{3\cdot 2^{r+1}}{(r+2)\,(r+3)}$$

In particular

$$\text{Mean} = \mu_1' = \frac{3\cdot 2^2}{3\cdot 4} = 1, \quad \mu_2' = \frac{3\cdot 2^3}{4\cdot 5} = \frac{6}{5},$$

$$\mu_3' = \frac{3\cdot 2^4}{5\cdot 6} = \frac{8}{5}, \quad \text{and} \quad \mu_4' = \frac{3\cdot 2^5}{6\cdot 7} = \frac{16}{7}$$

Hence varience $= \mu_2 = \mu_2' - \mu_1'^2 = \frac{6}{5} - 1 = \frac{1}{5}$

$$\mu_3 = \mu_3' - 3\,\mu_2'\,\mu_1' + 2\mu_1'^3 = \frac{8}{5} - 3\,.\,\frac{6}{5}\,.\,1 + 2 = 0$$

$$\mu_4 = \mu_4' - 4\,\mu_3'\,\mu_1' + 6\,\mu_2'\,\mu_1'^2 - 3\,\mu_1'^4 = \frac{16}{7} - 4\,.\,\frac{8}{5}\,.\,1 + 6\,.\,\frac{6}{5}\,.\,1 - 3\,.\,1 = \frac{3}{35}$$

$$\therefore \qquad \beta_1 = \frac{\mu_3^{\,2}}{\mu_2^{\,3}} = 0 \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^{\,2}} = \frac{3/35}{(1/5)^2} = \frac{15}{7}$$

Since $\beta_1 = 0$, the distributior is *symmetrical.*

Mean deviation about mean

$$= \int_0^2 |\,x-1\,|\,f(x)\,dx$$

$$= \int_0^1 |\,x-1\,|\,f(x)\,dx + \int_1^2 |\,x-1\,|\,f(x)\,dx$$

$$= \frac{3}{4}\left[\int_0^1 (1-x)\,x\,(2-x)\,dx + \int_1^2 (x-1)\,x\,(2-x)\,dx\right]$$

$$= \frac{3}{4}\left[\int_0^1 (2x - 3x^2 + x^3)\,dx + \int_1^2 (3x^2 - x^3 - 2x)\,dx\right]$$

$$= \frac{3}{4}\left[\left| x^2 - \frac{3.x^3}{3} + \frac{x^4}{4}\right|_0^1 + \left| 3.\frac{x^3}{3} - \frac{x^4}{4} - \frac{2x^2}{2}\right|_1^2\right] = \frac{3}{8}$$

$$\mu_{2n+1} = \int_0^2 (x - mean)^{2n+1} f(x)\,dx$$

$$= \frac{3}{4}\int_0^2 (x-1)^{2n+1} x(2-x)\,dx$$

$$= \frac{3}{4}\int_{-1}^1 t^{2n+1}(t+1)(1-t)\,dt \qquad (x - 1 = t)$$

$$= \frac{3}{4}\int_{-1}^1 t^{2n+1}(1-t^2)\,dt$$

Since t^{2n+1} 1 is an odd function of t and $(1 - t^2)$ is an even function of t, the integrand $t^{2n+1}(1 - t^2)$ is an odd function of t.

Hence $\mu_{2n+1} = 0.$

Now $f'(x) = \frac{3}{4}(2 - 2x) = 0 \quad \Rightarrow \quad x = 1$

and $f''(x) = \frac{3}{4}(-2) = -\frac{3}{2} < 0$

Hence mode = 1

Harmonic mean H is given by

$$\frac{1}{H} = \int_0^2 \frac{1}{x} f(x)\,dx$$

$$= \frac{3}{4}\int_0^2 (2 - x)\,dx = \frac{3}{2}$$

$$\Rightarrow \qquad H = \frac{2}{3}$$

If M is the median, then

$$\int_0^M f(x)\,dx = \frac{1}{2}$$

$$\Rightarrow \qquad \frac{3}{4}\int_0^M x(2-x)\,dx = \frac{1}{2}$$

$$\Rightarrow \qquad \left| x^2 - \frac{x^3}{3}\right|_0^M = \frac{2}{3}$$

$$\Rightarrow \qquad 3M^2 - M^3 = 2$$

$$\Rightarrow \qquad M^3 - 3M^2 + 2 = 0$$

$$\Rightarrow \qquad (M-1)(M^2 - 2M - 2) = 0$$

The only value of M lying in $[0, 2]$ is $M = 1$. Hence median is 1.

Aliter. Since we have proved that distribution is symmetrical,

Mode = Median = Mean = 1

Example 5·8. *The elementary probability law of a continuous random variable X is*

$$f(x) = y_o e^{-b(x-a)}, \quad a \le x < \infty, \quad b > 0$$

where a, b and y_o are constants.

Show that $y_o = b = 1/\sigma$ and $a = m - \sigma$, where m and σ are respectively the mean and standard deviation of the distribution. Show also that $\beta_1 = 4$ and $\beta_2 = 9$. **[Gauhati Univ. B.Sc., 1992]**

Solution. Since total probability is unity,

$$\int_a^\infty f(x)\,dx = 1 \quad \Rightarrow \quad y_o \int_a^\infty e^{-b(x-a)}\,dx = 1$$

$$\Rightarrow \quad y_o \left| \frac{e^{-b(x-a)}}{-b} \right|_a^\infty = 1 \quad \Rightarrow \quad y_o \frac{1}{b} = 1, \ (b > 0)$$

$$\Rightarrow \quad y_o = b$$

μ_r' (rth moment about the point '$x = a$')

$$= \int_a^\infty (x-a)^r f(x)\,dx = b \int_a^\infty (x-a)^r e^{-b(x-a)}\,dx$$

$$= b \int_0^\infty t^r e^{-bt}\,dt \qquad [\text{On putting } x - a = t]$$

$$= b \frac{\Gamma(r+1)}{b^{r+1}} = \frac{r!}{b^r} \qquad [\text{Using Gamma Integral}]$$

In particular

$$\mu_1' = 1/b, \ \mu_2' = 2/b^2, \ \mu_3' = 6/b^3, \ \mu_4' = 24/b^4$$

$$\therefore \quad m = \text{Mean} = a + \mu_1' = a + (1/b)$$

and $\quad \sigma^2 = \mu_2 = \mu_2' - \mu_1'^2 = 1/b^2$

$$\Rightarrow \quad \sigma = \frac{1}{b} \text{ and } m = a + \frac{1}{b} = a + \sigma$$

Hence $\quad y_o = b = \frac{1}{\sigma}$ and $a = m - \sigma$

Also $\quad \mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = \frac{1}{b^3}(6 - 3 \cdot 2 + 2) = \frac{2}{b^3} = 2\sigma^3$

and $\quad \mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$

$$= \frac{1}{b^4}(24 - 4.6.1 + 6.2.1 - 3) = \frac{9}{b^4} = 9\sigma^4$$

Hence $\beta_1 = \mu_3^2/\mu_2^3 = 4\sigma^6/\sigma^6 = 4$ and $\beta_2 = \mu_4/\mu_2^2 = 9\sigma^4/\sigma^4 = 9$

Example 5·9. *For the following probability distribution*

$$dF = y_o \cdot e^{-|x|}\, dx, \quad -\infty < x < \infty$$

show that $y_o = \frac{1}{2}$, $\mu_1' = 0$, $\sigma = \sqrt{2}$ *and mean deviation about mean = 1.*

Solution. We have $\int_{-\infty}^{\infty} f(x)\, dx = 1$

$$\Rightarrow y_o \int_{-\infty}^{\infty} e^{-|x|}\, dx = 1 \quad \Rightarrow \quad 2y_o \int_0^{\infty} e^{-|x|}\, dx = 1,$$

(since $e^{-|x|}$ is an even function of x)

$$\Rightarrow 2y_o \int_0^{\infty} e^{-x}\, dx = 1, \qquad (\text{since in } 0 \le x < \infty,\ |x| = x)$$

$$\Rightarrow 2y_o \left| \frac{e^{-x}}{-1} \right|_0^{\infty} = 1 \quad \Rightarrow \quad 2y_o = 1, \ i.e., \ y_o = \frac{1}{2}$$

$$\mu_1' \text{ (about origin)} = \int_{-\infty}^{\infty} x\, f(x)\, dx = \frac{1}{2} \int_{-\infty}^{\infty} x\, e^{-|x|}\, dx$$

$$= 0,$$

(since the integrand $x \cdot e^{-|x|}$ is an odd function of x)

$$\mu_2' = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \frac{1}{2} \int_{-\infty}^{\infty} x^2 e^{-|x|}\, dx$$

$$= \frac{1}{2}\, 2 \int_0^{\infty} x^2 e^{-|x|}\, dx$$

[since the integrand $x^2 e^{-|x|}$ is an even function of x]

$$\therefore \qquad \mu_2' = \int_0^{\infty} x^2 e^{-x}\, dx = \Gamma(3) \qquad (\text{on using Gamma Integral})$$

$$\Rightarrow \qquad \mu_2' = 2! = 2$$

$$\text{Now} \qquad \sigma^2 = \mu_2 = \mu_2' - \mu_1'^2 = 2$$

$$\text{M.D..about mean} = \int_{-\infty}^{\infty} |x - mean|\, f(x)\, dx$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} |x|\, e^{-|x|}\, dx \qquad (\because \text{Mean} = \mu_1' = 0)$$

$$= \frac{1}{2} \cdot 2 \int_0^{\infty} |x|\, e^{-|x|}\, dx$$

$$= \int_0^{\infty} x e^{-x}\, dx = \Gamma(2) = 1$$

Example 5·10. *A random variable X has the probability law :*

$$dF(x) = \frac{x}{b^2} . e^{-x^2/2b^2}\, dx, \quad 0 \le x < \infty$$

Find the distance between the quartiles and show that the ratio of this distance to the standard devation of X is independent of the patameter 'b'.

Solution. If Q_1 and Q_3 are the first and third quartiles respectively, we have

$$\int_0^{Q_1} f(x)\, dx = \frac{1}{4} \quad \Rightarrow \quad \frac{1}{b^2} \int_0^{Q_1} x\, e^{-x^2/2b^2}\, dx = \frac{1}{4}$$

Put $\quad y = \dfrac{x^2}{2b^2} \quad$ *then* $\quad dy = \dfrac{x}{b^2}\, dx$

$$\therefore \quad \int_0^{Q_1^2/2b^2} e^{-y}\, dy = \frac{1}{4} \quad \Rightarrow \quad \left| \frac{e^{-y}}{-1} \right|_0^{Q_1^2/2b^2} = \frac{1}{4}$$

$$\Rightarrow \quad 1 - e^{-Q_1^2/2b^2} = \frac{1}{4} \quad \Rightarrow \quad e^{-Q_1^2/2b^2} = \frac{3}{4}$$

$$\Rightarrow \quad Q_1 = \sqrt{2b}\, \sqrt{\log (4/3)}$$

Again we have $\int_0^{Q_3} f(x)\, dx = \dfrac{3}{4}$ which, on proceeding similarly, will give

$$1 - e^{-Q_3^2/2b^2} = 3/4 \quad \Rightarrow \quad e^{-Q_3^2/2b^2} = 1/4$$

$$\Rightarrow \quad Q_3 = \sqrt{2b}\, \sqrt{\log (4)}$$

The distance between the quartiles is given by

$$Q_3 - Q_1 = \sqrt{2b}\, [\sqrt{\log 4} - \sqrt{\log (4/3)}]$$

$$\mu_1' = \int_0^\infty x\, f(x)\, dx = \int_0^\infty x\, \frac{x}{b^2}\, e^{-x^2/2b^2}\, dx$$

$$= \int_0^\infty \sqrt{2b}\, y^{1/2}\, e^{-y}\, dy \qquad \left(y = \frac{x^2}{2b^2} \right)$$

$$= \sqrt{2b} \int_0^\infty e^{-y}\, y^{(3/2)-1}\, dy$$

$$= \sqrt{2}\, b\, \Gamma\left(\frac{3}{2}\right) = \sqrt{2}\, .b .\, \frac{1}{2}\, \Gamma\left(\frac{1}{2}\right) = \sqrt{2b}\, \frac{\sqrt{\pi}}{2} = b\sqrt{(\pi/2)}$$

$$\mu_2' = \int_0^\infty x^2\, f(x)\, dx = \int_0^\infty x^2\, \frac{x}{b^2}\, e^{-x^2/2b^2}\, dx$$

$$= 2b^2 \int_0^\infty y\, e^{-y}\, dy \qquad \left(y = \frac{x^2}{2b^2} \right)$$

$$= 2b^2\, \Gamma(2) = 2b^2 . 1 ! = 2b^2$$

$$\therefore \quad \sigma^2 = \mu_2 = \mu_2' - \mu_1'^2 = 2b^2 - b^2 \cdot \frac{\pi}{2} = b^2\left(2 - \frac{\pi}{2}\right)$$

$$\Rightarrow \quad \sigma = b\sqrt{2 - (\pi/2)}$$

Hence $$\frac{Q_3 - Q_1}{\sigma} = \frac{\sqrt{2}\,[\sqrt{\log 4} - \sqrt{\log (4/3)}]}{\sqrt{2 - (\pi/2)}},$$

which is independent of the parameter 'b'.

Example 5·11. *Prove that the geometric mean G of the distribution*

$$dF = 6(2 - x)(x - 1)\,dx, \quad 1 \le x \le 2$$

is given by $6 \log (16G) = 19$. **[Kanpur Univ. B.Sc., Oct. 1992]**

Solution. By definition, we have

$$\log G = \int_1^2 \log x\, f(x)\,dx = 6\int_1^2 \log x\,(2 - x)(x - 1)\,dx$$

$$= -6\int_1^2 (x^2 - 3x + 2)\log x\,dx$$

Integrating by parts, we get

$$\log G = -6\left[\left(\frac{x^3}{3} - \frac{3x^2}{2} + 2x\right)\log x\Big|_1^2 - \int_1^2 \left(\frac{x^3}{3} - \frac{3x^2}{2} + 2x\right)\frac{1}{x}\,dx\right]$$

$$= -4\log 2 + 6 \times \frac{19}{36} \qquad \text{(on simplification)}$$

$$\therefore \quad \log G + 4\log 2 = \frac{19}{6} \quad \Rightarrow \quad \log G + \log 2^4 = \frac{19}{6}$$

$$\Rightarrow \quad \log G + \log 16 = \frac{19}{6} \quad \Rightarrow \quad \log (16\,G) = \frac{19}{6}$$

$$\Rightarrow \quad 6\log (16\,G) = 19$$

Example 5·12. *The time one has to wait for a bus at a downtown bus stop is observed to be random phenomenon (X) with the following probability density function :*

$$f_X(x) = 0, \qquad \text{for} \quad x < 0$$
$$= \tfrac{1}{9}(x + 1), \quad \text{for} \quad 0 \le x < 1$$
$$= \tfrac{4}{9}(x - \tfrac{1}{2}), \quad \text{for} \quad 1 \le x < \tfrac{3}{2}$$
$$= \tfrac{4}{9}(\tfrac{5}{2} - x), \quad \text{for} \quad \tfrac{3}{2} \le x < 2$$
$$= \tfrac{1}{9}(4 - x), \quad \text{for} \quad 2 \le x < 3$$
$$= \tfrac{1}{9}, \qquad \text{for} \quad 3 \le x < 6$$

$= 0, \qquad for \quad 6 \le x,$

Let the events A and B be defined as follows :

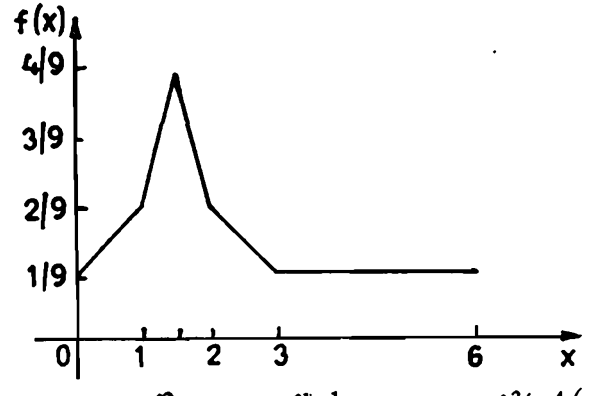
A : One waits between 0 to 2 minutes inclusive.

B : One waits between 0 to 3 minutes inclusive.

(i) Draw the graph of probability density function.

(ii) Show that (a) $P(B \mid A) = \frac{2}{3}$, (b) $P(\bar{A} \cap \bar{B}) = \frac{1}{3}$

Solution. *(i)* The graph of p.d.f. is given below.



(ii) (a) $P(A) = \int_0^2 f(x)\,dx = \int_0^1 \frac{1}{9}(x+1)\,dx + \int_1^{3/2} \frac{4}{9}\left(x - \frac{1}{2}\right)dx + \int_{3/2}^2 \frac{4}{9}\left(\frac{5}{2} - x\right)dx$

$= \frac{1}{2}$ (on simplification)

$P(A \cap B) = P(1 \le X \le 2) = \int_1^2 f(x)\,dx$

$= \int_1^{3/2} \frac{4}{9}\left(x - \frac{1}{2}\right)dx + \int_{3/2}^2 \frac{4}{9}\left(\frac{5}{2} - x\right)dx$

$= \frac{4}{9}\left[\frac{x^2}{2} - \frac{x}{2}\right]_1^{3/2} + \frac{4}{9}\left[\frac{5}{2}x - \frac{x^2}{2}\right]_{3/2}^2 = \frac{1}{3}$

(on simplification)

$\therefore \quad P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{1/3}{1/2} = \frac{2}{3}$

(b) $\bar{A} \cap \bar{B}$ means that waiting time is more than 3 minutes.

$\therefore \quad P(\bar{A} \cap \bar{B}) = P(X > 3) = \int_3^\infty f(x)\,dx = \int_3^6 f(x)\,dx + \int_6^\infty f(x)\,dx$

$= \int_3^6 \frac{1}{9}\,dx = \frac{1}{9} \mid x \mid_3^6 = \frac{1}{3}$

Example 5·13. *The amount of bread (in hundreds of pounds) X that a certain bakery is able to sell in a day is found to be a numerical valued random phenomenon, with a probability function specified by the probability density function $f(x)$, given by*

$$f(x) = A \cdot x, \quad \text{for } 0 \le x < 5$$
$$= A(10 - x), \quad \text{for } 5 \le x < 10$$
$$= 0, \quad \text{otherwise}$$

(a) Find the value of A such that $f(x)$ is a probability density function.

(b) What is the probability that the number of pounds of bread that will be sold tomorrow is

(i) more than 500 pounds,

(ii) less than 500 pounds,

(iii) between 250 and 750 pounds? **[Agra Univ. B.Sc., 1989]**

(c) Denoting by A, B, C the events that the pounds of bread sold are as in b (i), b (ii) and b (iii) respectively, find $P(A \mid B), P(A \mid C)$. Are (i) A and B independent events? (ii) Are A and C independent events ?

Solution. *(a)* In order that $f(x)$ should be a probability density function

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

i.e.,
$$\int_0^5 A x\,dx + \int_5^{10} A(10 - x)\,dx = 1$$

$$\Rightarrow \qquad A = \frac{1}{25} \qquad \text{(On simplification)}$$

(b) (i) The probability that the number of pounds of bread that will be sold tomorrow is more than 500 pounds, *i.e.,*

$$P(5 \le X \le 10) = \int_5^{10} \frac{1}{25}(10 - x)\,dx = \frac{1}{25}\left| 10x - \frac{x^2}{2} \right|_5^{10}$$

$$= \frac{1}{25}\left(\frac{25}{2}\right) = \frac{1}{2} = 0{\cdot}5$$

(ii) The probability that the number of pounds of bread that will be sold tomorrow is less than 500 pounds, *i.e.,*

$$P(0 \le X \le 5) = \int_0^5 \frac{1}{25} \cdot x\,dx = \frac{1}{25}\left| \frac{x^2}{2} \right|_0^5 = \frac{1}{2} = 0{\cdot}5$$

(iii) The required probability is given by

$$P(2{\cdot}5 \le X \le 7{\cdot}5) = \int_{2{\cdot}5}^5 \frac{1}{25}\,x\,dx + \int_5^{7{\cdot}5} \frac{1}{25}(10 - x)\,dx = \frac{3}{4}$$

(c) The events A, B and C are given by

$$A : 5 < X \le 10; \quad B : 0 \le X < 5; \quad C : 2{\cdot}5 < X < 7{\cdot}5$$

Then from parts *b (i), (ii)* and *(iii)*, we have

$$P(A) = 0{\cdot}5\,,\quad P(B) = 0{\cdot}5\,,\quad P(C) = \frac{3}{4}$$

The events $A \cap B$ and $A \cap C$ are given by

$$A \cap B = \phi \text{ and } A \cap C : 5 < X < 7{\cdot}5$$

$$\therefore \qquad P(A \cap B) = P(\phi) = 0$$

and
$$P(A \cap C) = \int_5^{7{\cdot}5} f(x)\,dx = \frac{1}{25}\int_5^{7{\cdot}5} (10 - x)\,dx$$

$$= \frac{1}{25} \times \frac{75}{8} = \frac{3}{8}$$

$$P(A)\,.\,P(C) = \frac{1}{2} \times \frac{3}{4} = \frac{3}{8} = P(A \cap C)$$

$\Rightarrow$ A and C are independent.

Again $P(A)\,.\,P(B) = \frac{1}{4} \neq P(A \cap B)$

$\Rightarrow$ A and B are not independent.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = 0$$

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)} = \frac{3/8}{3/4} = \frac{1}{2}$$

Example 5·14. *The mileage C in thousands of miles which car owners get with a certain kind of tyre is a random variable having probability density function*

$$f(x) = \frac{1}{20}\, e^{-x/20}, \text{ for } x > 0$$

$$= 0, \text{ for } x \leq 0$$

Find the probabilities that one of these tyres will last

(i) at most 10,000 miles,

(ii) anywhere from 16,000 to 24,000 miles.

(iii) at least 30,000 miles. **(Bombay Univ. B.Sc. 1989)**

Solution. Let *r.v.* X denote the mileage (in '000 miles) with a certain kind of tyre. Then required probability is given by:

(i)
$$P(X \leq 10) = \int_0^{10} f(x)\,dx = \frac{1}{20}\int_0^{10} e^{-x/20}\,dx$$

$$= \frac{1}{20}\left|\frac{e^{-x/20}}{-1/20}\right|_0^{10} = 1 - e^{-1/2}$$

$$= 1 - 0{\cdot}6065 = 0{\cdot}3935$$

(ii) $P(16 \le X \le 24) = \frac{1}{20}\int_{16}^{24} \exp\left(-\frac{x}{20}\right)dx = \left| -e^{-x/20} \right|_{16}^{24}$

$= e^{-16/20} - e^{-24/20} = e^{-4/5} - e^{-6/5}$

$= 0{\cdot}4493 - 0{\cdot}3012 = 0{\cdot}1481$

(iii) $P(X \ge 30) = \int_{30}^{\infty} f(x)\,dx = \frac{1}{20}\left| \frac{e^{-x/20}}{-1/20} \right|_{30}^{\infty}$

$= e^{-1\cdot5} = 0{\cdot}2231$

EXERCISE 5 (c)

1. *(a)* A continuous random variable X follows the probability law

$$f(x) = Ax^2,\ 0 \le x \le 1$$

Determine A and find the probability that *(i)* X lies between 0·2 and 0·5, *(ii)* X is less than 0·3, *(iii)* $1/4 < X < 1/2$ and *(iv)* $X > 3/4$ given $X > 1/2$.

Ans. $A = 0{\cdot}3$, *(i)* 0·117, *(ii)* 0·027, *(iii)* 15/256 and *(iv)* 27/56.

(b) If a random variable X has the density function

$$f(x) = \begin{cases} 1/4, & -2 < x < 2 \\ 0, & \textit{elsewhere}. \end{cases}$$

Obtain *(i)* $P(X < 1)$, *(ii)* $P(|X| > 1)$ *(iii)* $P(2X + 3 > 5)$

(Kerala Univ. B.Sc., Sept.1992)

Hint. *(ii)* $P(|X| > 1) = P(X > 1 \text{ or } X < -1) = \int_{-2}^{-1} f(x)\,dx + \int_{1}^{2} f(x)\,dx$

or $P(|X| > 1) = 1 - P(|X| \le 1) = 1 - P(-1 \le X \le 1)$

Ans. *(i)* 3/4, *(ii)* 1/2 *(iii)* 1/4.

2. Are any of the following probability mass or density functions? Prove your answer in each case.

(a) $f(x) = x;\ x = \frac{1}{16}, \frac{3}{16}, \frac{1}{4}, \frac{1}{2}$

(b) $f(x) = \lambda e^{-\lambda x};\ x \ge 0;\ \lambda > 0$

(c) $f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 4 - 2x, & 1 < x < 2 \\ 0, & \textit{elsewhere}, \end{cases}$

(Calicut Univ. B. Sc., Oct. 1988)

Ans. *(a)* and *(b)* are p.m.f./p.d.f.'s, *(c)* is not.

3. If f_1 and f_2 are p.d.f.'s and $\theta_1 + \theta_2 = 1$, check if

$$g(x) = \theta_1 f_1(x) + \theta_2 f_2(x), \quad \text{is a p.d.f.}$$

Ans. $g(x)$ is a p.d.f. if $0 \le (\theta_1, \theta_2) \le 1$ *a*·· $\theta_1 + \theta_2 = 1$.

4. A continuous random variable X has the probability density function :

$$f(x) = A + Bx, \quad 0 \le x \le 1.$$

If the mean of the distribution is $\frac{1}{2}$, find A and B.

Hint : Solve $\int_0^1 f(x)\,dx = 1$ and $\int_0^1 x\,f(x)\,dx = \frac{1}{2}$. Find A and B.

5. For the following density function

$$f(x) = c\,x^2(1-x), \quad 0 < x < 1,$$

find *(i)* the constant c, and *(ii)* mean.

[Calicut Univ. B.Sc.(subs.), 1991]

Ans. *(i)* $c = 12$; *(ii)* mean = 3/5 .

6. A continuous distribution of a variable X in the range (–3, 3) is defined by

$$f(x) = \tfrac{1}{16}(3+x)^2, \quad -3 \le x \le -1$$
$$= \tfrac{1}{16}(6-2x^2), \quad -1 \le x \le 1$$
$$= \tfrac{1}{16}(3-x)^2, \quad 1 \le x \le 3$$

(i) Verify that the area under the curve is unity.

(ii) Find the mean and variance of the above distribution.

(Madras Univ. B.Sc., Oct. 1992; Gujarat Univ. B.Sc., Oct. 1986)

Hint: $\int_{-3}^{3} f(x)\,dx = \int_{-3}^{-1} f(x)\,dx + \int_{-1}^{1} f(x)\,dx + \int_{1}^{3} f(x)\,dx$

Ans. Mean=0, Variance=1

7. If the random variable X has the p.d.f.,

$$f(x) = \tfrac{1}{2}(x+1), \quad -1 < x < 1$$
$$= 0, \text{ elsewhere,}$$

find the coefficient of skewness and kurtosis.

8. *(a)* A random variable X has the probability density function given by

$$f(x) = 6x(1-x), \quad 0 \le x \le 1$$

Find the mean μ, mode and S.D. σ, Compute $P(\mu - 2\sigma < X < \mu + 2\sigma)$. Find also the mean deviation about the median.

(Lucknow Univ. B.Sc., 1988)

(b) For the continuous distribution

$$dF = y_o(x - x^2)\,dx \; ; \quad 0 \le x \le 1, \; y_o \text{ being a constant.}$$

Find *(i)* arithmetic mean, *(ii)* harmonic mean, *(iii)* Median, *(iv)* Mode and *(v)* rth moment about mean. Hence find β_1 and β_2 and show that the distribution is symmetrical. **(Delhi Univ. B.Sc., 1992; Karnatak Univ. B.Sc., 1991)**

Ans. Mean = Median = Mode = $\frac{1}{2}$

(c) Find the mean, mode and median for the distribution,

$$dF(x) = \sin x\, dx, \quad 0 \le x \le \pi/2$$

Ans. 1, $\pi/2$, $\pi/3$

9. If the function $f(x)$ is defined by

$$f(x) = c\, e^{-\alpha x}, \quad 0 \le x < \infty, \quad \alpha > 0$$

(i) Find the value of constant c.

(ii) Evaluate the first four moments about mean.

[Gauhati Univ. B.Sc. 1987]

Ans. *(i)* $c = \alpha$, *(ii)* 0, $1/\alpha^2$, $2/\alpha^3$, $9/\alpha^4$.

10. *(a)* Show that for the exponential distribution

$$dP = y_o . e^{-x/\sigma}\, dx, \quad 0 \le x < \infty, \quad \sigma > 0$$

the mean and S.D. are both equal to σ and that the interquartile range is $\sigma \log_e 3$. Also find μ_r' and show that $\beta_1 = 4$, $\beta_2 = 9$.

[Agra Univ. B.Sc., 1986 ; Madras Univ. B.Sc., 1987]

(b) Define the harmonic mean (H.M.) of variable X as the reciprocal of the expected value of $1/X$, show that the H.M. of variable which ranges from 0 to ∞ with probability density $\frac{1}{6} x^3 e^{-x}$ is 3.

11. *(a)* Find the mean, variance and the co-efficients β_1, β_2 of the distribution,

$$dF = k\, x^2\, e^{-x}\, dx, \quad 0 < x < \infty.$$

Ans. $k = 1/2$; 3, 3, 4/3 and 5.

(b) Calculate β_1 for the distribution,

$$dF = k\, x\, e^{-x}\, dx, \quad 0 < x < \infty$$

Ans. 2 **[Delhi Univ. B.Sc. (Hons. Subs.), 1988]**

12. A continuous random variable X has a p.d.f. given by

$$f(x) = k\, x\, e^{\lambda x}, \quad x \ge 0, \quad \lambda > 0$$
$$= 0, \text{ otherwise}$$

Determine the constant k, obtain the mean and variance of X.

[Nagpur Univ. B.Sc. 1990]

13. For the probability density function,

$$f(x) = \frac{2(b+x)}{b(a+b)}, \quad -b \le x < 0$$
$$= \frac{2(a-x)}{a(a+b)}, \quad 0 \le x \le a$$

Find mean, median and variance. **[Calcutta Univ. B.Sc. 1984]**

Ans. Mean $= (a-b)/3$, Variance $= (a^2 + b^2 + ab)/18$,
Median $= a - \sqrt{a(a+b)/2}$

(ii) Show that, if terms of order $(a-b)^2/a^2$ are neglected, then

$$\text{mean} - \text{median} = (\text{mean} - \text{mode})/4$$

14. A variable X can assume values only between 0 and 5 and the equation of its frequency curve is

$$y = A \sin \tfrac{1}{5}\pi x, \quad 0 \le x \le 5$$

where A is a constant such that the area under the curve is unity. Determine the value of A and obtain the median and quartiles of the distribution.

Show also that the variance of the distribution is $50\left\{\frac{1}{8} - \frac{1}{\pi^2}\right\}$.

Ans. 1/10; 2·5, 4/3, 10/3

15. A continuous variable X is distributed over the interval [0, 1] with p.d.f. $ax^2 + bx$, where a, b are constants. If the arithmetic mean of X is 0·5, find the values of a and b.

Ans. $-6, 6$

16. A man leaves his house at the same time every morning and the time taken to journey to work has the following probability density function : less than 30 minutes, zero, between 30 minutes and 60 minutes, uniform with density k ; between 60 minutes and 70 minutes, uniform with density $2k$; and more than 70 minutes, zero. What is the probability that on one particular day he arrives at work later than on the previous day but not more than 5 minutes later.

17. The density function of sheer strength of spot welds is given by

$$f(x) = x/160{,}000 \quad \text{for} \quad 0 \le x \le 400$$
$$= (800 - x)/160{,}000 \quad \text{for} \quad 400 \le x \le 800$$

Find the number a such that

Prob. $(X < a) = 0{\cdot}50$ and the number b such that

Prob. $(X < b) = 0{\cdot}90$. Find the mean, median and variance of X. .

[Delhi Univ. B.E., 1987]

18. A batch of small calibre ammunition is accepted as satisfactory if none of a sample of five shot falls more than 2 feet from the centre of the target at a given range. If X, the distance from the centre of the target to a given impact point, actually has the density

$$f(x) = k.2xe^{-x^2}, \quad 0 < x < 3$$

where k is a number which makes it probability density function, what is the value of k and what is the probability that the batch will be accepted?

[Nagpur Univ. B.E., 1987]

Hint. $\int_0^3 f(x)\,dx = 1 \quad \Rightarrow \quad k = 1/(1 - e^{-9})$

Reqd. Prob. = P [Each of a sample of 5 shots falls within a distance of 2 ft. from the centre]

$$= [P(0<X<2)]^5 = \left[\int_0^2 f(x)\,dx\right]^5 = \left[\frac{1-e^{-4}}{1-e^{-9}}\right]^5$$

19. A random variable X has the p.d.f. :

$$f(x) = \begin{cases} 2x, & 0<x<1 \\ 0, & \textit{otherwise} \end{cases}$$

Find *(i)* $P\left(X<\frac{1}{2}\right)$, *(ii)* $P\left(\frac{1}{4}<X<\frac{1}{2}\right)$, *(iii)* $P\left(X>\frac{3}{4} \mid X>\frac{1}{2}\right)$, and *(iv)* $P\left(X<\frac{3}{4} \mid X>\frac{1}{2}\right)$. **(Gorakhpur Univ. B.Sc., 1988)**

Ans. *(i)* 1/4 , *(ii)* 3/16 , *(iii)* $\frac{P(X>3/4)}{P(X>1/2)} = \frac{7/16}{3/4} = \frac{7}{12}$; *(iv)* $\frac{P(1/2<X<3/4)}{P(X>1/2)}$

5·4·3. Continuous Distribution Function. If X is a continuous random variable with the p.d.f. $f(x)$, then the function

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt\,, \quad -\infty < x < \infty\,. \qquad ...(5\cdot12)$$

is called the *distribution function* (d.f.) or sometimes the *cumulative distribution function* (c.d.f.) of the random variable X.

Remarks 1. $0 \le F(x) \le 1, \quad -\infty < x < \infty$.

2. From analysis (Riemann integral), we know that

$$F'(x) = \frac{d}{dx}F(x) = f(x) \ge 0 \qquad [\because f(x) \text{ is p.d.f.}]$$

$\Rightarrow$ $F(x)$ is non-decreasing function of x.

3. $$F(-\infty) = \lim_{x\to-\infty} F(x) = \lim_{x\to-\infty} \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{-\infty} f(x)\,dx = 0$$

and $$F(+\infty) = \lim_{x\to\infty} F(x) = \lim_{x\to\infty} \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{\infty} f(x)\,dx = 1$$

4. $F(x)$ is a continuous function of x on the right.
5. The discontinuities of $F(x)$ are at the most countable.
6. It may be noted that

$$P(a \le X \le b) = \int_a^b f(x)\,dx = \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)\,dx$$

$$= P(X \le b) - P(X \le a) = F(b) - F(a)$$

Similarly

$$P(a<X<b) = P(a<X\le b) = P(a\le X<b) = \int_a^b f(t)\,dt$$

7. Since $F'(x) = f(x)$, we have

$$\frac{d}{dx} F(x) = f(x) \quad \Rightarrow \quad dF(x) = f(x)\, dx$$

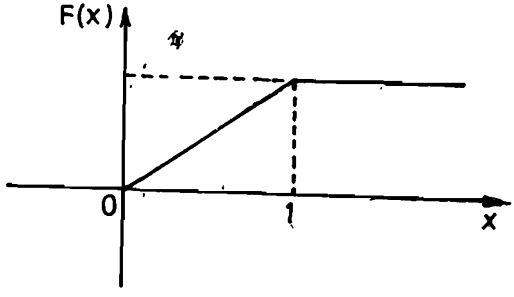
This is known as probability differential of X.

Remarks. 1. It may be pointed out that the properties (2), (3) and (4) above uniquely characterise the distribution functions. This means that any function $F(x)$ satisfying (2) to (4) is the distribution function of some random variable, and any function $F(x)$ violating any one or more of these three properties cannot be the distribution function of any random variable.

2. Often, one can obtain a p.d.f. from a distribution function $F(x)$ by differentiating $F(x)$, provided the derivative exists. For example, consider

$$F_X(x) = \begin{cases} 0, & \text{for } x < 0 \\ x, & \text{for } 0 \le x \le \\ 1, & \text{for } x > 1 \end{cases}$$

The graph of $F(x)$ is given by bold lines. Obviously we see that $F(x)$ is continuous from right as stipulated in (4) and we also see that $F(x)$ is not continuous at $x = 0$ and $x = 1$ and hence is not derivable at $x = 0$ and $x = 1$.



Differentiating $F(x)$ w.r.t. x, we get

$$\frac{d}{dx} F(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \textit{otherwise} \end{cases}$$

[Note the strict inequality in $0 < x < 1$, since $F(x)$ is not derivable at $x = 0$ and $x = 1$]

Let us define

$$f(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \textit{otherwise} \end{cases}$$

Then $f(x)$ is a p.d.f. for F.

Example 5·15. *Verify that the following is a distribution function:*

$$F(x) = \begin{cases} 0, & x < -a \\ \frac{1}{2}\left(\frac{x}{a} + 1\right), & -a \le x \le a \\ 1, & x > a \end{cases}$$

(Madras Univ. B.Sc., 1992)

Solution. Obviously the properties *(i), (ii), (iii)* and *(iv)* are satisfied. Also we observe that $F(x)$ is continuous at $x = a$ and $x = -a$, as well.

Now

$$\frac{d}{dx} F(x) = \begin{cases} \frac{1}{2a}, & -a \le x \le a \\ 0, & \text{otherwise} \end{cases}$$

$$= f(x), \text{ say}$$

In order that $F(x)$ is a distribution function, $f(x)$ must be a p.d.f. Thus we have to show that

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

Now $$\int_{-\infty}^{\infty} f(x)\, dx = \int_{-a}^{a} f(x)\, dx = \frac{1}{2a}\int_{-a}^{a} 1 \,.\, dx = 1$$

Hence $F(x)$ is a d.f.

Example 5·16. *Suppose the life in hours of a certain kind of radio tube has the probability density function :*

$$f(x) = \frac{100}{x^2}, \quad \text{when } x \ge 100$$

$$= 0, \quad \text{when } x < 100$$

Find the distribution function of the distribution. What is the probabilty that none of three such tubes in a given radio set will have to be replaced during the first 150 hours of operation ? What is the probability that all three of the original tubes will have been replaced during the first 150 hours ? **(Delhi Univ. B.Sc, Oct. 1988)**

Solution. Probability that a tube will last for first 150 hours is given by

$$P(X \le 150) = P(0 < X < 100) + P(100 \le X \le 150)$$

$$= \int_{100}^{150} f(x)\, dx = \int_{100}^{150} \frac{100}{x^2}\, dx = \frac{1}{3}$$

Hence the probability that none of the three tubes will have to be replaced during the first 150 hours is $(1/3)^3 = 1/27$.

The probability that a tube will not last for the first 150 hours is $1 - \frac{1}{3} = \frac{2}{3}$.

Hence the probability that all three of the original tubes will have to be replaced during the first 150 hours is $(2/3)^3 = 8/27$.

Example 5·17. *Suppose that the time in minutes that a person has to wait at a certain station for a train is found to be a random phenomenon, a probability function specified by the distribution function,*

$$\begin{aligned} F(x) &= 0, && \text{for } x \le 0 \\ &= \frac{x}{2}, && \text{for } 0 \le x < 1 \\ &= \frac{1}{2}, && \text{for } 1 \le x < 2 \\ &= \frac{x}{4}, && \text{for } 2 \le x < 4 \\ &= 1, && \text{for } x \ge 4 \end{aligned}$$

(a) Is the Distribution Function continuous ? If so, give the formula for its probability density function ?

(b) What is the probability that a person will have to wait (i) more than 3 minutes, (ii) less than 3 minutes, and (iii) between 1 and 3 minutes ?

(c) What is the conditional probability that the person will have to wait for a train for (i) more than 3 minutes, given that it is more than 1 minute, (ii) less than 3 minutes given that it is more than 1 minute ? **(Calicut Univ. B.Sc., 1985)**

Solution. *(a)* Since the value of the distribution function is the same at the points $x = 0, x = 1, x = 2$, and $x = 4$ given by the two forms of $F(x)$ for $x < 0$ and $0 \le x < 1$, $0 \le x < 1$ and $1 \le x < 2$, $1 \le x < 2$ and $2 \le x < 4$, $2 \le x < 4$ and $x \ge 4$, the distribution function is continuous.

Probability density function $= f(x) = \dfrac{d}{dx} F(x)$

$$\therefore \quad \begin{aligned} f(x) &= 0, && \text{for } x < 0 \\ &= \frac{1}{2}, && \text{for } 0 \le x < 1 \\ &= 0, && \text{for } 1 \le x < 2 \\ &= \frac{1}{4}, && \text{for } 2 \le x < 4 \\ &= 0, && \text{for } x \ge 4 \end{aligned}$$

(b) Let the random variable X represent the waiting time in minutes.

Then

(i) Required probability $= P(X > 3) = 1 - P(X \le 3) = 1 - F(3)$

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

(ii) Required probability $= P(X < 3) = P(X \le 3) - P(X = 3)$

$$= F(3) = \frac{3}{4}$$

(Since, the probability that a continuous variable takes a fixed value is zero)

(iii) Required Probability $= P(1 < X < 3) = P(1 < X \leq 3)$

$= F(3) - F(1) = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}$

(c) Let A denote the event that a person has to wait for more than 3 minutes and B the event that he has to wait for more than 1 minute. Then

$$P(A) = P(X > 3) = \frac{1}{4} \qquad [\textit{c.f. (b), (i)}]$$

$$P(B) = P(X > 1) = 1 - P(X \leq 1) = 1 - F(1) = 1 - \frac{1}{2} = \frac{1}{2}$$

$$P(A \cap B) = P(X > 3 \cap X > 1) = P(X > 3) = \frac{1}{4}$$

(i) Required probability is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{1/2} = \frac{1}{2}$$

(ii) Required probability $= P(\bar{A} \mid B) = \dfrac{P(\bar{A} \cap B)}{P(B)}$

Now $P(\bar{A} \cap B) = P(X \leq 3 \cap X > 1) = P(1 < X \leq 3) = F(3) - F(1) = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}$

$$\therefore \qquad P(\bar{A} \mid B) = \frac{1/4}{1/2} = \frac{1}{2}$$

Example 5·18. *A petrol pump is supplied with petrol once a day. If its daily volume X of sales in thousands of litres is distributed by*

$$f(x) = 5(1-x)^4, \quad 0 \leq x \leq 1,$$

what must be the capacity of its tank in order that the probability that its supply will be exhausted in a given day shall be 0·01 ? **(Madras Univ. B.E., 1986)**

Solution. Let the capacity of the tank (in '000 of litres) be 'a' such that

$$P(X \geq a) = 0{\cdot}01 \quad \Rightarrow \quad \int_a^1 f(x)\,dx = 0{\cdot}01$$

$$\Rightarrow \quad \int_a^1 5(1-x)^4\,dx = 0{\cdot}01 \quad \textit{or} \quad \left[5 \cdot \frac{(1-x)^5}{(-5)}\right]_a^1 = 0{\cdot}01$$

$$\Rightarrow \quad (1-a)^5 = 1/100 \quad \textit{or} \quad 1 - a = (1/100)^{1/5}$$

$$\therefore \quad a = 1 - (1/100)^{1/5} = 1 - 0{\cdot}3981 = 0{\cdot}6019$$

Hence the capacity of the tank $= 0{\cdot}6019 \times 1000$ *litres* $= 601{\cdot}9$ *litres.*

Example 5·19. *Prove that mean deviation is least when measured from the median.* **[Delhi Univ. B.Sc. (Maths. Hons.), 1989]**

Solution. If $f(x)$ is the probability function of a random variable X, $a \leq X \leq b$, then mean deviation $M(A)$, say, about the point $x = A$ is given by

$$M(A) = \int_a^b |x - A| f(x)\,dx$$

$$= \int_a^A |x - A| f(x)\,dx + \int_A^b |x - A| f(x)\,dx$$

$$= \int_a^A (A - x) f(x)\,dx + \int_A^b (x - A) f(x)\,dx \quad \ldots (1)$$

We want to find the value of 'A' so that $M(A)$ is minimum. From the principle of maximum and minimum in differential calculus, $M(A)$ will be minimum for variations in A if

$$\frac{\partial M(A)}{\partial A} = 0 \quad \text{and} \quad \frac{\partial^2 M(A)}{\partial A^2} > 0 \quad \ldots (2)$$

Differentiating (1) w.r.t. 'A' under the integral sign, since the functions $(A - x) f(x)$ and $(x - A) f(x)$ vanish at the point $x = A^*$, we get

$$\frac{\partial M(A)}{\partial A} = \int_a^A f(x)\,dx - \int_A^b f(x)\,dx \quad \ldots (3)$$

Also $$\frac{\partial M(A)}{\partial A} = \int_a^A f(x)\,dx - \left[1 - \int_a^A f(x)\,dx\right],$$

$$\left[\because \int_a^b f(x)\,dx = 1\right]$$

$$= 2\int_a^A f(x)\,dx - 1 = 2F(A) - 1,$$

where $F(\cdot)$ is the distribution function of X. Differentiating again w.r.t. A, we get

$$\frac{\partial^2}{\partial A^2} M(A) = 2f(A) \quad \ldots (4)$$

Now $\dfrac{\partial M(A)}{\partial A} = 0$, on using (3) gives

$$\int_a^A f(x)\,dx = \int_A^b f(x)\,dx$$

i.e., A is the median value.

Also from (4), we see that

$$\frac{\partial^2 M(A)}{\partial A^2} > 0,$$

assuming that $f(x)$ does not vanish at the median value. Thus mean deviation is least when taken from median.

---

*If $f(x, \theta)$ is a continuous function of both variables x and θ, possessing continuous partial derivatives $\dfrac{\partial^2 f}{\partial x \partial \theta}$, $\dfrac{\partial^2 f}{\partial \theta \partial x}$ and a and b are differentiable functions of θ, then

$$\frac{\partial}{\partial \theta}\left[\int_a^b f(x, \theta)\,dx\right] = \int_a^b \frac{\partial f}{\partial \theta}\,dx + f(b, \theta)\frac{db}{d\theta} - f(a, \theta)\frac{da}{d\theta}$$

EXERCISE 5 (d)

1. *(a)* Explain the terms *(i)* probability differential, *(ii)* probability density function, and *(iii)* distribution function.

(b) Explain what is meant by a random variable. Distinguish between a discrete and a continuous random variable. Define distribution function of a random variable and show that it is monotonic non-decreasing everywhere and continuous on the right at every point.

[Madras Univ. B.Sc. (Stat Main), 1987]

(c) Show that the distribution function $F(x)$ of a random variable X is a non-decreasing function of x. Determine the jump of $F(x)$ at a point x_0 of its discontinuity in terms of the probability that the random variable has the value x_0.

[Calcutta Univ. B.Sc. (Hons.), 1984]

2. The length (in hours) X of a certain type of light bulb may be supposed to be a continuous random variable with probability density function :

$$f(x) = \frac{a}{x^3}, \quad 1500 < x < 2500$$
$$= 0, \quad \text{elsewhere}$$

Determine the constant a, the distribution function of X, and compute the probability of the event $1,700 \leq X \leq 1,900$.

Ans. $a = 70,31,250$; $F(x) = \frac{a}{2}\left(\frac{1}{22,50,000} - \frac{1}{x^2}\right)$ and

$$P(1,700 < X < 1,900) = F(1,900) - F(1,700) = \frac{a}{2}\left(\frac{1}{28,90,000} - \frac{1}{36,10,000}\right)$$

3. Define the "distribution function" (or cumulative distribution function) of a random variable and state its essential properties.

Show that, whatever the distribution function $F(x)$ of a random variable X, $P[a \leq F(x) \leq b] = b - a$, $0 \leq a, b \leq 1$.

4. *(a)* The distribution function of a random variable X is given by

$$F(x) = \begin{cases} 1 - (1+x)e^{-x}, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0 \end{cases}$$

Find the corresponding density function of random variable X.

(b) Consider the distribution for X defined by

$$F(x) = \begin{cases} 0, & \text{for } x < 0 \\ 1 - \frac{1}{4}e^{-x} & \text{for } x \geq 0 \end{cases}$$

Determine $P(x = 0)$ and $P(x > 0)$.

[Allahabad Univ. B.Sc., 1992]

5. *(a)* Let X be a continuous random variable with probability density function given by

$$f(x) = \begin{cases} ax, & 0 \le x \le 1 \\ a, & 1 \le x \le 2 \\ -ax + 3a, & 2 \le x \le 3 \\ 0, & \textit{elsewhere} \end{cases}$$

(i) Determine the constant a.

(ii) Determine $F(x)$, and sketch its graph.

(iii) If three independent observations are made, what is the probability that exactly one of these three numbers is larger than 1·5 ?

[Rajasthan Univ. M.Sc., 1987]

Ans. *(i)* 1/2, *(iii)* 3/8.

(b) For the density $f_X(x) = k\,e^{-ax}(1 - e^{-ax})\ I_{0,\infty}(x)$, find the normalising constant k, $f_X(x)$ and evaluate $P(X > 1)$.

[Delhi Univ. B.Sc. (Maths Hons.), 1989]

Ans. $k = 2a$; $F(x) = 1 - 2e^{-ax} + e^{-2ax}$; $P(X > 1) = 2e^{-a} - e^{-2a}$

6. A random variable X has the density function :

$$f(x) = K \cdot \frac{1}{1 + x^2}, \quad \text{if } -\infty < x < \infty$$

$$= 0, \quad \textit{otherwise}$$

Determine K and the distribution function.

Evaluate the probability $P(X \ge 0)$. Find also the mean and variance of X.

[Karnatak Univ. B.Sc. 1985]

Ans. $K = 1$, $F(x) = \frac{1}{\pi}\left\{\tan^{-1} x + \frac{\pi}{2}\right\}$, $P(x \ge 0) = 1/2$, Mean = 0, Variance does not exist.

7. A continuous random variable X has the distribution function

$$F(x) = \begin{cases} 0, & \text{if } x \le 1 \\ k(x-1)^4, & \text{if } 1 < x \le 3 \\ 1, & \text{if } x > 3 \end{cases}$$

Find *(i)* k, *(ii)* the probability density function $f(x)$, and *(iii)* the mean and the median of X.

Ans. *(i)* $k = \frac{1}{16}$, *(ii)* $f(x) = \frac{1}{4}(x-1)^3$, $1 \le x \le 3$

8. Given $f(x) = \begin{cases} kx(1-x), & \text{for } 0 < x < 1 \\ 0, & \textit{elsewhere} \end{cases}$

Show that

(i) $k = 1/5$, *(ii)* $F(x) = 0$ for $x \le 0$ and $F(x) = 1 - e^{-x/5}$, for $x > 0$

Using $F(x)$, show that

(iii) $P(3 < X < 5) = 0{\cdot}1809$, *(iv)* $P(X < 4) = 0{\cdot}5507$, *(v)* $P(X > 6) = 0{\cdot}3012$

9. A bombing plane carrying three bombs flies directly above a railroad track. If a bomb falls within 40 feet of track, the track will be sufficiently damaged to disrupt the traffic. Within a certain bomb site the points of impact of a bomb have the probability density function :

$$f(x) = (100 + x)/10{,}000, \quad \textit{when } -100 \le x \le 0$$
$$= (100 - x)/10{,}000, \quad \textit{when } 0 \le x \le 100$$
$$= 0, \quad \textit{elsewhere}$$

where x represents the vertical deviation (in feet) from the aiming point, which is the track in this case. Find the distribution function. If all the bombs are used, what is the probability that track will be damaged ?

Hint. Probability that track will be damaged by the bomb is given by

$$P(|X| < 40) = P(-40 < X < 40)$$
$$= \int_{-40}^{0} f(x)\,dx + \int_{0}^{40} f(x)\,dx$$
$$= \int_{-40}^{0} \frac{100+x}{10{,}000}\,dx + \int_{0}^{40} \frac{100-x}{10{,}000}\,dx = \frac{16}{25}$$

$\therefore$ Probability that a bomb will not damage the track $= 1 - \frac{16}{25} = \frac{9}{25}$

Probability that none of the three bombs damages the track $= \left(\frac{9}{25}\right)^3 = 0{\cdot}046656$

Required probability that the track will be damaged $= 1 - 0{\cdot}046656 = 0{\cdot}953344$.

10. The length of time (in minutes) that a certain lady speaks on the telephone is found to be random phenomenon, with a probability function specified by the probability density function $f(x)$ as

$$f(x) = A\,e^{-x/5}, \quad \text{for } x \ge 0$$
$$= 0, \quad \textit{otherwise}$$

(a) Find the value of A that makes $f(x)$ a p.d.f.

Ans. $A = 1/5$

(b) What is the probability that the number of minutes that she will talk over the phone is

(i) More than 10 minutes, *(ii)* less than 5 minutes, and *(iii)* between 5 and 10 minutes ? **[Shivaji Univ. B.Sc., 1990]**

Ans. *(i)* $\frac{1}{e^2}$, *(ii)* $\frac{e-1}{e}$, *(iii)* $\frac{e-1}{e^2}$.

11. The probability that a person will die in the time interval (t_1, t_2) is given by

$$A \int_{t_1}^{t_2} f(t)\, dt\,;$$

where A is a constant and the function $f(t)$ determined from long records, is

$$f(t) = \begin{cases} t^2 (100 - t)^2, & 0 \le t \le 100 \\ 0, & \textit{elsewhere} \end{cases}$$

Find the probability that a person will die between the ages 60 and 70 assuming that his age is ≥ 50. **[Calcutta Univ. B.A. (Hons.), 1987]**

5·5. Joint Probability Law. Two random variables X and Y are said to be jointly distributed if they are defined on the same probability space. The sample points consist of 2–tuples. If the joint probability function is denoted by $P_{XY}(x, y)$ then the probability of a certain event E is given by

$$P_{XY}(x, y) = P[(X, Y) \in E] \qquad \ldots (5\cdot13)$$

(X, Y) is said to belong to E, if in the 2 dimensional space the 2–tuples lie in the Borel set B, representing the event E.

5·5·1. Joint Probability Mass Function and Marginal and Conditional Probability Functions. Let X and Y be random variables on a sample space S with respective image sets $X(S) = \{x_1, x_2, \ldots, x_n\}$ and $Y(S) = \{y_1, y_2, \ldots, y_m\}$. We make the product set

$$X(S) \times Y(S) = \{x_1, x_2, \ldots, x_n\} \times \{y_1, y_2, \ldots, y_m\}$$

into a probability space by defining the probability of the ordered pair (x_i, y_j) to be $P(X = x_i, Y = y_j)$ which we write $p(x_i, y_j)$. The function p on $X(S) \times Y(S)$ defined by

$$p_{ij} = P(X = x_i \cap Y = y_j) = p(x_i, y_j) \qquad \ldots (5\cdot14)$$

is called the *joint probability function* of X and Y and is usually represented in the form of the following table :

X \ Y	y_1	y_2	y_3	...	y_j	...	y_m	Total
x_1	p_{11}	p_{12}	p_{13}	...	p_{1j}	...	p_{1m}	$p_{1.}$
x_2	p_{21}	p_{22}	p_{23}	...	p_{2j}	...	p_{2m}	$p_{2.}$
x_3	p_{31}	p_{32}	p_{33}	...	p_{3j}	...	p_{3m}	$p_{3.}$
⋮	⋮	⋮						⋮
x_i	p_{i1}	p_{i2}	p_{i3}	...	p_{ij}	...	p_{im}	$p_{i.}$
⋮	⋮	⋮						⋮
x_n	p_{n1}	p_{n2}	p_{n3}	...	p_{nj}	...	p_{nm}	$p_{n.}$
Total	$p_{.1}$	$p_{.2}$	$p_{.3}$	...	$p_{.j}$	...	$p_{.m}$	1

$$\therefore \quad \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) = 1$$

Suppose the joint distribution of two random variables X and Y is given, then the probability distribution of X is determined as follows :

$$p_X(x_i) = P(X = x_i) = P[X = x_i \cap Y = y_1] + P[X = x_i \cap Y = y_2] + \ldots + P[X = x_i \cap Y = y_j] + \ldots + P[X = x_i \cap \ldots y_m]$$

$$= p_{i1} + p_{i2} + \ldots + p_{ij} + \ldots + p_{im}$$

$$= \sum_{j=1}^{m} p_{ij} = \sum_{j=1}^{m} p(x_i, y_j) = p_i . \qquad \ldots (5\cdot14\,a)$$

and is known as *marginal probability function of X.*

Also $\sum_{i=1}^{n} p_{i.} = p_{1.} + p_{2.} + \ldots + p_{n.} = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} = 1$

Similarly, we can prove that

$$p_Y(y_j) = P(Y = y_j) = \sum_{i=1}^{n} p_{ij} = \sum_{i=1}^{n} p(x_i, y_j) = p_{.j} \qquad \ldots (5\cdot14\,b)$$

which is the *marginal probability function of Y.*

Also

$$P[X = x_i \mid Y = y_j] = \frac{P[X = x_i \cap Y = y_j]}{P[Y = y_j]} = \frac{p(x_i, y_j)}{p(y_j)} = \frac{p_{ij}}{p_{.j}}$$

This is known as *conditional probability function of X given* $Y = y_j$

Similarly

$$P[Y = y_j \mid X = x_i] = \frac{p(x_i, y_j)}{p(x_i)} = \frac{p_{ij}}{p_{i.}} \qquad \ldots (5\cdot14\,c)$$

is the *conditional probability function of Y given* $X = x_i$

Also $\sum_{i=1}^{n} \frac{p_{ij}}{p_{.j}} = \frac{p_{1j} + p_{2j} + \ldots + p_{ij} + \ldots + p_{nj}}{p_{.j}} = \frac{p_{.j}}{p_{.j}} = 1$

Similarly

$$\sum_{j=1}^{n} \frac{p_{ij}}{p_{i.}} = 1$$

Two random variables X and Y are said to be *independent* if

$$P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j), \qquad \ldots (5\cdot14\,d)$$

otherwise they are said to be dependent.

5·5·2. Joint Probability Distribution Function. Let (X, Y) be a two-dimensional random variable then their joint distribution function is denoted by $F_{XY}(x, y)$ and it represents the probability that simultaneously the observation

(X, Y) will have the property $(X \le x$ and $Y \le y)$, *i.e.*,

$$F_{XY}(x, y) = P(-\infty < X \le x, -\infty < Y \le y)$$

$$= \int_{-\infty}^{x} \left[\int_{-\infty}^{y} f_{XY}(x, y)\, dx\, dy \right] \quad \ldots (5\cdot15)$$

(For continuous variables)

where $f_{XY}(x, y) \ge 0$

And $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)\ dx\, dy = 1$ *or* $\sum_{x} \sum_{y} f(x, y) = 1$

Properties of Joint Distribution Function

1. (*i*) For the real numbers a_1, b_1, a_2 and b_2

$$P(a_1 < X \le b_1, a_2 < Y \le b_2) = F_{XY}(b_1, b_2) + F_{XY}(a_1, a_2)$$

$$- F_{XY}(a_1, b_2) - F_{XY}(b_1, a_2)$$

[For proof, See Example 5·29]

(*ii*) Let $a_1 < a_2, b_1 < b_2$. We have

$$(X \le a_1, Y \le a_2) + (a_1 < X \le b_1, Y \le a_2) = (X \le b_1, Y \le a_2)$$

and the events on the L.H.S. are mutually exclusive. Taking probabilities on both-sides, we get :

$$F(a_1, a_2) + P(a_1 < X \le b_1, Y \le a_2) = F(b_1, a_2)$$

$$\Rightarrow \quad F(b_1, a_2) - F(a_1, a_2) = P(a_1 < X \le b_1, Y \le a_2)$$

$$\therefore \quad F(b_1, a_2) \ge F(a_1, a_2) \qquad [\text{Since } P(a_1 < X \le b_1, Y \le a_2) \ge 0]$$

Similarly it follows that

$$F(a_1, b_2) - F(a_1, a_2) = P(X \le a_1, a_2 < Y \le b_2) \ge 0$$

$$\Rightarrow \quad F(a_1, b_2) \ge F(a_1, a_2),$$

which shows that $F(x, y)$ is monotonic non-decreasing function.

2. $F(-\infty, y) = 0 = F(x, -\infty)$, $F(+\infty, +\infty) = 1$

3. If the density function $f(x, y)$ is continuous at (x, y) then

$$\frac{\partial^2 F}{\partial x\, \partial y} = f(x, y)$$

5·5·3. Marginal Distribution Functions. From the knowledge of joint distribution function $F_{XY}(x, y)$, it is possible to obtain the individual distribution functions, $F_X(x)$ and $F_Y(y)$ which are termed as marginal distribution function of X and Y respectively with respect to the joint distribution function $F_{XY}(x, y)$.

$$F_X(x) = P(X \le x) = P(X \le x, Y < \infty) = \lim_{y \to \infty} F_{XY}(x, y).$$

$$= F_{XY}(x, \infty) \quad \ldots (5\cdot16)$$

Similarly, $F_Y(y) = P(Y \le y) = P(X < \infty, Y \le y)$

$$= \lim_{x \to \infty} F_{XY}(x, y) = F_{XY}(\infty, y)$$

$F_X(x)$ is termed as the marginal distribution function of X corresponding to the joint distribution function $F_{XY}(x, y)$ and similarly $F_Y(y)$ is called marginal distribution function of the random variable Y corresponding to the joint distribution function $F_{XY}(x, y)$.

In the case of jointly discrete random variables, the marginal distribution functions are given as

$$F_X(x) = \sum_y P(X \le x, Y = y),$$

$$F_Y(y) = \sum_x P(X = x, Y \le y)$$

Similarly in the case of jointly continuous random variable, the marginal distribution functions are given as

$$F_X(x) = \int_{-\infty}^{x} \left\{ \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy \right\} dx$$

$$F_Y(y) = \int_{-\infty}^{y} \left\{ \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx \right\} dy$$

5·5·4 Joint Density Function, Marginal Density Functions. From the joint distribution function $F_{XY}(x, y)$ of two dimensional continuous random variable we get the joint probabilty density function by differentiation as follows :

$$f_{XY}(x, y) = \partial^2 F(x, y)/\partial x \partial y$$

$$= \lim_{\delta x \to 0, \delta y \to 0} \frac{P(x \le X \le x + \delta x, y \le Y \le y + \delta y)}{\delta x\, \delta y}$$

Or it may be expressed in the following way also :

"The probability that the point (x, y) will lie in the infinitesimal rectangular region, of area $dx\, dy$ is given by

$$P\left\{x - \tfrac{1}{2}dx \le X \le x + \tfrac{1}{2}dx, y - \tfrac{1}{2}dy \le Y \le y + \tfrac{1}{2}dy\right\} = dF_{XY}(x, y)$$

and is denoted by $f_{XY}(x, y)\, dx\, dy$, where the function $f_{XY}(x, y)$ is called the joint probability density function of X and Y.

The marginal probability function of Y and X are given respectively

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx \qquad \text{(for continuous variables)}$$

$$= \sum_x p_{XY}(x, y) \qquad \text{(for discrete variables)}$$

...(5·17)

and $$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy \qquad \text{(for continuous variables)}$$

$$= \sum_{y} p_{XY}(x, y) \qquad \text{(for discrete variables)}$$

(5·17*a*)

The marginal density functions of X and Y can be obtained in the following manner also.

$$f_X(x) = \frac{dF_X(x)}{dx} = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy$$

and $$f_Y(y) = \frac{dF_Y(y)}{dy} = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx \qquad \text{... (5·17 } b\text{)}$$

Important Remark. If we know the joint p.d.f. (p.m.f.) $f_{XY}(x, y)$ of two random variables X and Y, we can obtain the individual distributions of X and Y in the form of their marginal p.d.f.'s (p.m.f's) $f_X(x)$ and $f_Y(y)$ by using (5·17) and (5·17a). However, the converse is not true *i.e., from the marginal distributions of two jointly distributed random variables, we cannot determine the joint distributions of these two random variables.*

To verify this, it will suffice to show that two different joint p.m.f's (p.d.f.'s) have the same marginal distribution for X and the same marginal distribution for Y. We give below two joint discrete probability distributions which have the same marginal distributions.

JOINT DISTRIBUTIONS HAVING SAME MARGINALS

Probability Distribution I

X / Y	0	1	$f_Y(y)$
0	0·28	0·37	0·65
1	0·22	0·13	0·35
$f_X(x)$	0·50	0·50	1·00

Probability Distribution II

X / Y	0	1	$f_Y(y)$
0	0·35	0·30	0·65
1	0·15	0·20	0·35
$f_X(x)$	0·50	0·50	1·00

As an illustration for continuous random variables, let (X, Y) be continuous r.v. with joint p.d.f.

$$f_{XY}(x, y) = x + y \;;\; 0 \le (x, y) \le 1 \qquad \text{...(5·17 } c\text{)}$$

The marginal p.d.f.'s of X and Y are given by :

$$f_X(x) = \int_0^1 f(x, y)\, dy = \int_0^1 (x + y)\, dy = \left| xy + \frac{y^2}{2} \right|_0^1$$

$$\left.\begin{aligned} \Rightarrow \quad & f_X(x) = x + \tfrac{1}{2} \; ; && 0 \le x \le 1 \\ \text{Similarly} \quad & f_Y(y) = \int_0^1 f(x, y)\, dx = y + \tfrac{1}{2} \; ; && 0 \le y \le 1 \end{aligned}\right\} \quad \ldots (5{\cdot}17\,d)$$

Consider another continuous joint p.d.f.

$$g(x, y) = \left(x + \tfrac{1}{2}\right)\left(y + \tfrac{1}{2}\right) \; ; \quad 0 \le (x, y) \le 1 \qquad \ldots (5{\cdot}17\,e)$$

Then marginal p.d.f.'s of X and Y are given by :

$$g_1(x) = \int_0^1 g(x, y)\, dy = \left(x + \tfrac{1}{2}\right) \int_0^1 \left(y + \tfrac{1}{2}\right) dy$$

$$= \left(x + \tfrac{1}{2}\right) \left| \frac{y^2}{2} + \frac{1}{2} y \right|_0^1$$

$$\left.\begin{aligned} \Rightarrow \quad & g_1(x) = x + \tfrac{1}{2} \; ; && 0 \le x \le 1 \\ \text{Similarly} \quad & g_2(y) = y + \tfrac{1}{2} \; ; && 0 \le y \le 1 \end{aligned}\right\} \quad \ldots (5{\cdot}17\,f)$$

(5·17 *d*) and (5·17 *f*) imply that the two joint p.d.f.'s in (5·17 *c*) and (5·17 *e*) have the same marginal p.d.f.'s (5·17 *d*) or (5·17 *f*).

Another illustration of continuous r.v.'s is given in Remark to Bivariate Normal Distribution, § 10·10·2.

5·5·5. The Conditional Distribution Function and Conditinal Probability Density Function. For two diamensional random variable (X, Y), the joint distribution function $F_{XY}(x, y)$, for any real numbers x and y is given by

$$F_{XY}(x, y) = P(X \le x, Y \le y)$$

Now let A be the event $(Y \le y)$ such that the event A is said to occur when Y assumes values up to and inclusive of y.

Using conditional probabilities we may now write

$$F_{XY}(x, y) = \int_{-\infty}^{x} P[A \mid X = x]\, dF_X(x) \qquad \ldots (5{\cdot}18)$$

The *conditional distribution function* $F_{Y|X}(y \mid x)$ denotes the distribution function of Y when X has already assumed the particular value x. Hence

$$F_{Y|X}(y \mid x) = P[Y \le y \mid X = x) = P[A \mid X = x]$$

Using this expression, the joint distribution function $F_{XY}(x, y)$ may be expressed in terms of the conditional distribution function as follows :

$$F_{XY}(x, y) = \int_{-\infty}^{x} F_{Y|X}(y \mid x)\, dF_X(x) \qquad \ldots (5{\cdot}18\,a)$$

Similarly

$$F_{XY}(x, y) = \int_{-\infty}^{y} F_{X|Y}(x \mid y)\, dF_Y(y) \qquad \ldots (5{\cdot}18\,b)$$

The *conditional probability density function* of Y given X for two random variables X and Y which are jointly continuously distributed is defined as follows, for two real numbers x and y :

$$f_{Y|X}(y|x) = \frac{\partial}{\partial y} F_{Y|X}(y|x) \qquad \ldots (5\cdot19)$$

Remarks : 1. $f_X(x) > 0$, then

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

Proof. We have

$$F_{XY}(x, y) = \int_{-\infty}^{x} F_{Y|X}(y|x)\, dF_X(x)$$

$$= \int_{-\infty}^{x} F_{Y|X}(y|x)\, f_X(x)\, dx$$

Differentiating w.r.t. x, we get

$$\frac{\partial}{\partial x} F_{XY}(x, y) = F_{Y|X}(y|x)\, f_X(x)$$

Differentiating w.r.t. y, we get

$$\frac{\partial}{\partial y}\left[\frac{\partial}{\partial x} F_{XY}(x, y)\right] = f_{Y|X}(y|x)\, f_X(x)$$

$$\Rightarrow \quad f_{XY}(x, y) = f_{Y|X}(y|x)\, f_X(x)$$

$$\Rightarrow \quad f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

2. If $f_Y(y) > 0$, then

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

3. In terms of the differentials, we have

$$P(x < X \le x + dx \,|\, y < Y \le y + dy)$$

$$= \frac{P(x < X \le x + dx,\ y < Y \le y + dy)}{P(y < Y \le y + dy)}$$

$$= \frac{f_{XY}(x, y)\, dx\, dy}{f_Y(y)\, dy} = f_{X|Y}(x|y)\, dx$$

Whence $f_{X|Y}(x|y)$ may be interpreted as the conditional density function of X on the assumption $Y = y$.

5·5·6. Stochastic Independence. Let us consider two random variables X and Y (of discrete or continuous type) with joint p.d.f. $f_{XY}(x, y)$ and marginal p.d.f.'s $f_X(x)$ and $g_Y(y)$ respectively. Then by the compound probability theorem

$$f_{XY}(x, y) = f_X(x)\, g_Y(y|x)$$

where $g_Y(y \mid x)$ is the conditional p.d.f. of Y for given value of $X = x$.

If we assume that $g(y \mid x)$ does not depend on x, then by the definition of marginal p.d.f.'s, we get for continuous r.v.'s

$$g(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

$$= \int_{-\infty}^{\infty} f_X(x)\, g(y \mid x)\, dx$$

$$= g(y \mid x) \int_{-\infty}^{\infty} f_X(x)\, dx$$

[since $g(y \mid x)$ does not depend on x]

$$= g(y \mid x) \qquad [\because f(.) \text{ is p.d.f. of } X]$$

Hence

$$g(y) = g(y \mid x)$$

and $$f_{XY}(x, y) = f_X(x)\, g_Y(y) \qquad \ldots (*)$$

provided $g(y \mid x)$ does not depend on x. This motivates the following definition of independent random variables.

Independent Random variables. *Two r.v.'s X and Y with joint p.d.f. $f_{XY}(x, y)$ and marginal p.d.f.'s $f_X(x)$ and $g_Y(y)$ respectively are said to be stochastically independent if and only if*

$$f_{X,Y}(x, y) = f_X(x)\, g_Y(y) \qquad \ldots (5{\cdot}20)$$

Remarks. 1. In terms of the distribution function, we have the following definition :

Two jointly distributed random variables X and Y are stochastically independent if and only if their joint distribution function $F_{X,Y}(.,.)$ is the product of their marginal distribution functions $F_X(\cdot)$ and $G_Y(\cdot)$, i.e., if for real (x, y)

$$F_{X,Y}(x, y) = F_X(x)\, G_Y(y) \qquad \ldots (5{\cdot}20\,a)$$

2. The variables which are not stochastically independent are said to be stochastically dependent.

Theorem 5·8. *Two random variables X and Y with joint p.d.f. $f(x, y)$ are stochastically independent if and only if $f_{X,Y}(x, y)$ can be expressed as the product of a non-negative function of x alone and a non-negative function of y alone, i.e., if*

$$f_{X,Y}(x, y) = h_X(x)\, k_Y(y) \qquad \ldots (5{\cdot}21)$$

where $h(\cdot) \geq 0$ and $k(\cdot) \geq 0$.

Proof. If X and Y are independent then by definition, we have

$$f_{X,Y}(x, y) = f_X(x) \cdot g_Y(y)$$

where $f(x)$ and $g(y)$ are marginal p.d.f. of X and Y respectively. Thus condition (5·21) is satisfied.

Conversely if (5·21) holds, then we have to prove that X and Y are independent. For continuous random variables X and Y, the marginal p.d.f.'s are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{-\infty}^{\infty} h(x)\, k(y)\, dy$$

$$= h(x) \int_{-\infty}^{\infty} k(y)\, dy = c_1\, h(x), \text{ say} \quad \ldots (*)$$

and $$g_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_{-\infty}^{\infty} h(x)\, k(y)\, dx$$

$$= k(y) \int_{-\infty}^{\infty} h(x)\, dx = c_2\, k(y), \text{ say} \quad \ldots (**)$$

where c_1 and c_2 are constants independent of x and y. Moreover

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

$$\Rightarrow \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)\, k(y)\, dx\, dy = 1$$

$$\Rightarrow \quad \left(\int_{-\infty}^{\infty} h(x)\, dx\right)\left(\int_{-\infty}^{\infty} k(y)\, dy\right) = 1$$

$$\Rightarrow \quad c_2\, c_1 = 1 \quad \ldots (***)$$

Finally, we get

$$f_{X,Y}(x, y) = h_X(x)\, k_Y(y) = c_1 c_2\, h_X(x)\, k_Y(y) \quad [\text{using } (***)]$$

$$= (c_1 h_X(x))\ (c_2 k_Y(y))$$

$$= f_X(x)\ g_Y(y) \quad [\text{from } (*) \text{ and } (**)]$$

$\Rightarrow$ X and Y are stochastically independent.

Theorem 5·9. *If the random variables X and Y are stochastically independent, then for all possible selections of the corresponding pairs of real numbers (a_1, b_1), (a_2, b_2) where $a_i \le b_i$ for all $i = 1, 2$ and where the values $\pm \infty$ are allowed, the events $(a_1 < X \le b_1)$ and $(a_2 < Y \le b_2)$ are independent, i.e.,*

$$P[(a_1 < X \le b_1) \cap (a_2 < Y \le b_2)] = P(a_1 < X \le b_1)\, P(a_2 < Y \le b_2)$$

Proof. Since X and Y are stochastically independent, we have in the usual notations

$$f_{X,Y}(x, y) = f_X(x)\ g_Y(y) \quad \ldots (*)$$

In case of continuous r.v.'s, we have

$$P[(a_1 < X \le b_1) \cap (a_2 < Y \le b_2)] = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y)\, dx\, dy$$

$$= \left(\int_{a_1}^{b_1} f_X(x)\, dx \right) \left(\int_{a_2}^{b_2} g_Y(y)\, dy \right) \qquad [\text{from } (*)]$$

$$= P(a_1 < X \leq b_1)\, P(a_2 < Y \leq b_2)$$

as desired.

Remark. In case of discrete r.v.'s theorems 5·8 and 5·9 can be proved on replacing integration by summation over the given range of the variables.

Example 5.20. *For the following bivariate probability distribution of X and Y, find*

(i) $P(X \leq 1, Y = 2)$, *(ii)* $P(X \leq 1)$, *(iii)* $P(Y = 3)$, *(iv)* $P(Y \leq 3)$ *and (v)* $P(X < 3, Y \leq 4)$

X \ Y	1	2	3	4	5	6
0	0	0	$\frac{1}{32}$	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{3}{32}$
1	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$
2	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{64}$	0	$\frac{2}{64}$

Solution. The marginal distributions are given below :

X \ Y	1	2	3	4	5	6	$p_X(x)$
0	0	0	$\frac{1}{32}$	$\frac{2}{32}$	$\frac{2}{32}$	$\frac{3}{32}$	$\frac{8}{32}$
1	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{10}{16}$
2	$\frac{1}{32}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{64}$	0	$\frac{2}{64}$	$\frac{8}{64}$
$p_Y(y)$	$\frac{3}{32}$	$\frac{3}{32}$	$\frac{11}{64}$	$\frac{13}{64}$	$\frac{6}{32}$	$\frac{16}{64}$	$\Sigma p(x) = 1$ $\Sigma p(y) = 1$

(i) $P(X \leq 1, Y = 2) = P(X = 0, Y = 2) + P(X = 1, Y = 2)$

$$= 0 + \frac{1}{16} = \frac{1}{16}$$

(ii) $P(X \leq 1) = P(X = 0) + P(X = 1)$

$$= \frac{8}{32} + \frac{10}{16} = \frac{7}{8} \qquad (\text{From above table})$$

(iii) $P(Y = 3) = \frac{11}{64}$ (From above table)

(iv) $P(Y \leq 3) = P(Y = 1) + P(Y = 2) + P(Y = 3)$

$$= \frac{3}{32} + \frac{3}{32} + \frac{11}{64} = \frac{23}{64}$$

(v) $$P(X<3,\ Y\le 4)=P(X=0,\ Y\le 4)+P(X=1,\ Y\le 4)+P(X=2,\ Y\le 4)$$
$$=\left(\frac{1}{32}+\frac{2}{32}\right)+\left(\frac{1}{16}+\frac{1}{16}+\frac{1}{8}+\frac{1}{8}\right)+\left(\frac{1}{32}+\frac{1}{32}+\frac{1}{64}+\frac{1}{64}\right)=\frac{9}{16}$$

Example 5·21. *The joint probability distribution of two random variables X and Y is given by :*

$$p(x,y)=\frac{2}{n(n+1)},\quad x=1,2,...,n;\quad y=1,2,...,x$$

Examine whether X and Y are independent. **(Calicut Univ. B.Sc., 1991)**

Solution. The joint probability distribution table along with the marginal distributions of X and Y is given below.

Y \ X	1	2	3	.....	n	$p_Y(y)$
1	$\frac{2}{n(n+1)}$	$\frac{2}{n(n+1)}$	$\frac{2}{n(n+1)}$	.....	$\frac{2}{n(n+1)}$	$\frac{2n}{n(n+1)}$
2	–	$\frac{2}{n(n+1)}$	$\frac{2}{n(n+1)}$	.....	$\frac{2}{n(n+1)}$	$\frac{2(n-1)}{n(n+1)}$
3	–	–	$\frac{2}{n(n+1)}$	.....	$\frac{2}{n(n+1)}$	$\frac{2(n-2)}{n(n+1)}$
⋮	⋮					
$n-1$	–	–	–	$\frac{2}{n(n+1)}$	$\frac{2}{n(n+1)}$	$\frac{2\times 2}{n(n+1)}$
n	–	–	–	–	$\frac{2}{n(n+1)}$	$\frac{2}{n(n+1)}$
$p_X(x)$	$\frac{2}{n(n+1)}$	$\frac{2\times 2}{n(n+1)}$	$\frac{2\times 3}{n(n+1)}$	.....	$\frac{2\times n}{n(n+1)}$	

Note that $y=1,2,...,x$.

When $x=1$, $y=1$; when $x=2$, $y=1,2$; when $x=3$, $y=1,2,3$ and so on.

From the above table, we see that

$$p_{XY}(x,y)\neq p_X(x)\,p_Y(y)\ ;\quad \forall\ x,y$$

$\Rightarrow$ X and Y are not independent.

Example 5·22. *Given the following bivariate probability distribution, obtain (i) marginal distributions of X and Y, (ii) the conditional distribution of X given Y = 2.*

Y \ X	−1	0	1
0	1/15	2/15	1/15
1	3/15	2/15	1/15
2	2/15	1/15	2/15

(Mysore Univ. B.Sc., Oct. 1987)

Solution.

Y \ X	−1	0	1	$\sum_x p(x,y)$
0	1/15	2/15	1/15	4/15
1	3/15	2/15	1/15	6/15
2	2/15	1/15	2/15	5/15
$\sum_y p(x,y)$	6/15	5/15	4/15	1

(i) Marginal distribution of X. From the above table, we get

$$P(X=-1)=\frac{6}{15}=\frac{2}{5}\ ;\ P(X=0)=\frac{5}{15}=\frac{1}{3}\ ;\ P(X=1)=\frac{4}{15}$$

Marginal distribution of Y :

$$P(Y=0)=\frac{4}{15}\ ;\ P(Y=1)=\frac{6}{15}=\frac{2}{5}\ ;\ P(Y=2)=\frac{5}{15}=\frac{1}{3}$$

(ii) Conditional distribution of X given Y = 2. We have

$$P(X=x\cap Y=2)=P(Y=2).P(X=x\mid Y=2)$$

$$\Rightarrow\quad P(X=x\mid Y=2)=\frac{P(X=x\cap Y=2)}{P(Y=2)}$$

$$\therefore\quad P(X=-1\mid Y=2)=\frac{P(X=-1\cap Y=2)}{P(Y=2)}=\frac{2/15}{1/3}=\frac{2}{5}$$

Example 5·23. *X and Y are two random variables having the joint density function,* $f(x,y)=\frac{1}{27}(2x+y)$, *where x and y can assume only the integer values 0, 1 and 2. Find the conditional distribution of Y for X = x.*

[South Gujarat Univ. B.Sc., 1988]

Solution. The joint probability function

$$f(x, y) = \frac{1}{27}(2x + y); \quad x = 0, 1, 2; \quad y = 0, 1, 2$$

gives the following table of joint probability distribution of X and Y.

JOINT PROBABILITY DISTRIBUTION $f(x, y)$ OF X AND Y

$X \downarrow$ \ $Y \rightarrow$	0	1	2	$f_X(x)$
0	0	1/27	2/27	3/27
1	2/27	3/27	4/27	9/27
2	4/27	5/27	6/27	15/27

For example $f(0, 0) = \frac{1}{27}(0 + 2 \times 0) = 0$

$f(1, 0) = \frac{1}{27}(0 + 2 \times 1) = \frac{2}{27}$; $f(2, 0) = \frac{1}{27}(0 + 2 \times 2) = \frac{4}{27}$ and so on.

The marginal probability distribution of X is given by

$$f_X(x) = \sum_y f(x, y),$$

and is tabulated in last column of above table.

The conditional distribution of Y for $X = x$ is given by

$$f_{Y|X}(Y = y \mid X = x) = \frac{f(x, y)}{f_X(x)}$$

and is obtained in the following table.

CONDITIONAL DISTRIBUTION OF Y FOR $X = x$

X \ Y	0	1	2
0	0	1/3	2/3
1	2/9	3/9	4/9
2	4/15	5/15	6/15

Example 5.24. *Two discrete random variables X and Y have the joint probability density function :*

$$p(x, y) = \frac{\lambda^x e^{-\lambda} p^y (1-p)^{x-y}}{y!(x-y)!}, \quad y = 0, 1, 2, ..., x \; ; \quad x = 0, 1, 2, ...$$

where λ, p *are constants with* $\lambda > 0$ *and* $0 < p < 1$.

Find (i) The marginal probability density functions of X and Y.

(ii) The conditional distribution of Y for a given X and of X for a given Y.

(Poona Univ. B.Sc., 1986 ; Nagpur Univ. M.Sc., 1989)

Solution. *(i)*

$$p_X(x) = \sum_{y=0}^{x} p(x,y) = \sum_{y=0}^{x} \frac{\lambda^x e^{-\lambda} p^y (1-p)^{x-y}}{y!(x-y)!}$$

$$= \frac{\lambda^x e^{-\lambda}}{x!} \sum_{y=0}^{x} \frac{x! p^x (1-p)^{x-y}}{y!(x-y)!} = \frac{\lambda^x e^{-\lambda}}{x!} \sum_{y=0}^{x} {}^xC_y p^y (1-p)^{x-y}$$

$$= \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$

which is the probability function of a Poisson distribution with parameter λ.

$$p_Y(y) = \sum_{x=0}^{\infty} p(x,y) = \sum_{x=y}^{\infty} \frac{\lambda^x e^{-\lambda} p^y (1-p)^{x-y}}{y!(x-y)!}$$

$$= \frac{(\lambda p)^y e^{-\lambda}}{y!} \sum_{x=y}^{\infty} \frac{[\lambda(1-p)]^{x-y}}{(x-y)!} = \frac{(\lambda p)^y e^{-\lambda}}{y!} e^{\lambda(1-p)}$$

$$= \frac{e^{-\lambda p} (\lambda p)^y}{y!}, \quad y = 0, 1, 2, \ldots,$$

which is the probability function of a Poisson distribution with parameter λp.

(ii) The conditional distribution of Y for given X is

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)} = \frac{\lambda^x e^{-\lambda} p^y (1-p)^{x-y} x!}{y!(x-y)! \lambda^x e^{-\lambda}}$$

$$= \frac{x!}{y!(x-y)!} p^y (1-p)^{x-y} = {}^xC_y p^y (1-p)^{x-y}, \quad x > y$$

The conditional probability distribution of X for given Y is

$$p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

$$= \frac{\lambda^x e^{-\lambda} p^y (1-p)^{x-y}}{y!(x-y)!} \cdot \frac{y!}{e^{-\lambda p}(\lambda p)^y} \qquad [\textit{c.f.} \text{ Part (i)}]$$

$$= \frac{e^{-\lambda q}(\lambda q)^{x-y}}{(x-y)!}; \quad q = 1-p, \; x > y$$

Example 5·25. *The joint p.d.f. of two random variables X and Y is given by* :

$$f(x,y) = \frac{9(1+x+y)}{2(1+x)^4 (1+y)^4}; \quad \begin{pmatrix} 0 \le x < \infty \\ 0 < y < \infty \end{pmatrix}$$

Find the marginal distributions of X and Y, and the conditional distribution of Y for X = x.

Solution. Marginal p.d.f. of X is given by

$$f_X(x) = \int_0^\infty f(x,y)\, dy$$

$$= \frac{9}{2(1+x)^4} \int_0^\infty \frac{(1+y)+x}{(1+y)^4}\, dy$$

$$= \frac{9}{2(1+x)^4} \cdot \int_0^\infty \left[(1+y)^{-3} + x\,(1+y)^{-4}\right] dy$$

$$= \frac{9}{2(1+x)^4}\left[\left|\frac{-1}{2(1+y)^2}\right|_0^\infty + x\left|\frac{-1}{3(1+y)^3}\right|_0^\infty\right]$$

$$= \frac{9}{2(1+x)^4} \cdot \left[\frac{1}{2} + \frac{x}{3}\right]$$

$$= \frac{3}{4} \cdot \frac{3+2x}{(1+x)^4}\ ;\ 0 < x < \infty$$

Since $f(x,y)$ is symmetric in x and y, the marginal p.d.f. of Y is given by

$$f_Y(y) = \int_0^\infty f(x,y)\, dx$$

$$= \frac{3}{4} \cdot \frac{3+2y}{(1+y)^4}\ ;\ 0 < y < \infty$$

The conditional distribution of Y for $X = x$ is given by

$$f_{XY}(Y = y \mid X = x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

$$= \frac{9(1+x+y)}{2(1+x)^4(1+y)^4} \cdot \frac{4(1+x)^4}{3(3+2x)}$$

$$= \frac{6(1+x+y)}{(1+y)^4(3+2x)}\ ;\ 0 < y < \infty$$

Example 5·26. *The joint probability density function of a two-dimensional random variable (X,Y) is given by*

$$f(x,y) = 2\ ;\ 0 < x < 1,\ 0 < y < x$$

$$= 0,\ \textit{elsewhere}$$

(i) Find the marginal density functions of X and Y,

(ii) find the conditional density function of Y given X = x and conditional density function of X given Y = y, and

(iii) check for independence of X and Y.

[M.S.Baroda Univ. B.Sc., 1987; Karnataka Univ. B.Sc., Oct. 1988]

Solution. Evidently $f(x, y) \geq 0$ and

$$\int_0^1 \int_0^x 2\, dx\, dy = 2 \int_0^1 x\, dx = 1$$

(i) The marginal p.d.f.'s of X and Y are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy = \int_0^x 2\, dy = 2x, \quad 0 < x < 1$$

$$= 0, \text{ elsewhere}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx = \int_y^1 2dx = 2(1 - y), \quad 0 < y < 1$$

$$= 0, \text{ elsewhere}$$

(ii) The conditional density function of Y given X is

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{2}{2x} = \frac{1}{x}, \quad 0 < x < 1$$

The conditional density function of X given Y is

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)} = \frac{2}{2(1-y)} = \frac{1}{(1-y)}, \quad 0 < y < 1$$

(iii) Since $f_X(x)\, f_Y(y) = 2(2x)(1-y) \neq f_{XY}(x, y)$, X and Y are not independent.

Example 5·27. *A gun is aimed at a certain point (origin of the coordinate system). Because of the random factors, the actual hit point can be any point (X,Y) in a circle of radius R about the origin. Assume that the joint density of X and Y is constant in this circle given by :*

$$f_{XY}(x, y) = k, \text{ for } x^2 + y^2 \leq R^2$$

$$= 0, \text{ otherwise}$$

(i) Compute k, (ii) show that

$$f_X(x) = \frac{2}{\pi R} \left\{ 1 - \left(\frac{x}{R} \right)^2 \right\}^{1/2}, \quad \text{for } -R \leq x \leq R$$

$$= 0, \text{ otherwise}$$

[Calcutta Univ. B.Sc.(Stat. Hons.),1987]

Solution. *(i)* The constant k is computed from the consideration that the total probability is 1, *i.e.*,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1 \quad \Rightarrow \quad \iint_{x^2 + y^2 \leq R^2} k\, dx\, dy = 1$$

$$\Rightarrow \qquad 4 \iint_I k\,dx\,dy = 1$$

where region I is the first quadrant of the circle $x^2 + y^2 = R^2$.

$$\Rightarrow \qquad 4k \int_0^R \left(\int_0^{\sqrt{R^2 - x^2}} 1 \,.\, dy \right) dx = 1$$

$$\Rightarrow \qquad 4k \int_0^R \sqrt{R^2 - x^2}\, dx = 1$$

$$\Rightarrow \qquad 4k \left| x\sqrt{R^2 - x^2} + \frac{R^2}{2} \sin^{-1}\left(\frac{x}{R}\right) \right|_0^R = 1$$

$$\Rightarrow \qquad 4k \cdot \left(\frac{R^2}{2} \cdot \frac{\pi}{2} \right) = 1 \quad \Rightarrow \quad k = \frac{1}{\pi R^2}$$

$$\therefore \qquad f_{XY}(x, y) = 1/(\pi R^2) \;;\quad x^2 + y^2 \le R^2$$
$$= 0, \quad \textit{otherwise}$$

(ii) $$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \frac{1}{\pi R^2} \int_{-\sqrt{R^2 - x^2}}^{\sqrt{R^2 - x^2}} 1 \cdot dy$$

$$\left[\textit{because } x^2 + y^2 \le R^2 \;\Rightarrow\; -\left(R^2 - x^2\right)^{1/2} \le y \le \left(R^2 - x^2\right)^{1/2} \right]$$

$$= \frac{2}{\pi R^2} \int_0^{\sqrt{R^2 - x^2}} 1 \cdot dy = \frac{2}{\pi R^2} \left(R^2 - x^2\right)^{1/2}$$

$$= \frac{2}{\pi R} \left[1 - \left(\frac{x}{R}\right)^2 \right]^{1/2}$$

Example 5·28. *Given:*

$$f(x, y) = e^{-(x+y)}\, I_{(0,\infty)}(x) \,.\, I_{(0,\infty)}(y)\,,$$

find *(i)* $P(X > 1)$, *(ii)* $P(X < Y \mid X < 2Y)$, *(iii)* $P(1 < X + Y < 2)$

[Delhi Univ. B.Sc. (Maths Hons.), 1987]

Solution. We are given :

$$f(x, y) = e^{-(x+y)} \;;\quad 0 \le x < \infty,\ 0 \le y < \infty \qquad \ldots (1)$$
$$= \left(e^{-x}\right)\left(e^{-y}\right)$$
$$= f_X(x) \,.\, f_Y(y) \;;\quad 0 \le x < \infty,\ 0 \le y < \infty$$

$\Rightarrow$ X and Y are independent and

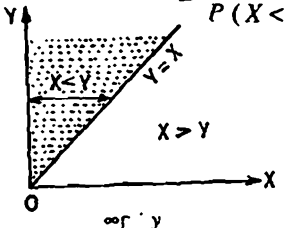
$$f_X(x) = e^{-x} \;;\; x \ge 0 \quad \text{and} \quad f_Y(y) = e^{-y} \;;\; y \ge 0 \qquad \ldots (2)$$

(i) $$P(X > 1) = \int_1^\infty f_X(x)\,dx = \int_1^\infty e^{-x}\,dx$$

$$= \left| \frac{e^{-x}}{-1} \right|_1^\infty = \frac{1}{e}$$

(ii) $$P(X < Y \mid X < 2Y) = \frac{P(X < Y \cap X < 2Y)}{P(X < 2Y)}$$

$$= \frac{P(X < Y)}{P(X < 2Y)} \qquad \ldots (3)$$



$$P(X < Y) = \int_0^\infty \left[\int_0^y f(x,y)\,dx \right] dy$$

$$= \int_0^\infty \left[e^{-y} \left| \frac{e^{-x}}{-1} \right|_0^y \right] dy = -\int_0^\infty e^{-y}\left(e^{-y} - 1\right) dy$$

$$= -\left| \frac{e^{-2y}}{-2} + e^{-y} \right|_0^\infty = 1 - \frac{1}{2} = \frac{1}{2}$$

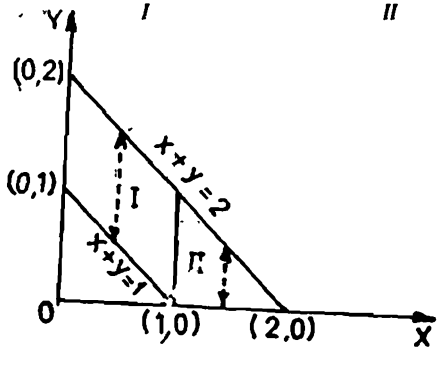
$$P(X < 2Y) = \int_0^\infty \left[\int_0^{2y} f(x,y)\,dx \right] dy = -\int_0^\infty e^{-y}(e^{-2y} - 1)\,dy$$

$$= -\left| \frac{e^{-3y}}{-3} + e^{-y} \right|_0^\infty = 1 - \frac{1}{3} = \frac{2}{3}$$

Substituting in (3),

$$P(X < Y \mid X < 2Y) = \frac{1/2}{2/3} = \frac{3}{4}$$

(iii) $$P(1 < X + Y < 2) = \iint_I f(x,y)\,dx\,dy = \iint_{II} f(x,y)\,dx\,dy$$



$$= \int_0^1 \left(\int_{1-x}^{2-x} f(x,y)\,dy \right) dx + \int_1^2 \left(\int_0^{2-x} f(x,y)\,dy \right) dx$$

$$= \int_0^1 \left(e^{-x} \int_{1-x}^{2-x} e^{-y}\,dy \right) dx + \int_1^2 \left(e^{-x} \int_0^{2-x} e^{-y}\,dy \right) dx$$

$$= \int_0^1 \frac{e^{-x}}{-1}\left(e^{x-2} - e^{x-1}\right) dx + \int_1^2 \frac{e^{-x}}{-1}\left(e^{x-2} - 1\right) dx$$

$$= -\left(e^{-2} - e^{-1}\right) \int_0^1 1 \cdot dx - \int_1^2 \left(e^{-2} - e^{-x}\right) dx$$

$$= -\left(e^{-2} - e^{-1}\right) \Big| x \Big|_0^1 - \Big| e^{-2} \cdot x + e^{-x} \Big|_1^2$$

$$= 2/e - 3/e^2$$

Example 5·29. *(i) Let $F(x,y)$ be the d.f. of X and Y. Show that $P(a < X \le b,\ c < Y \le d) = F(b,d) - F(b,c) - F(a,d) + F(a,c)$ where a, b, c, d are real constants $a < b$; $c < d$.*

Deduce that if: $F(x,y) = 1$, for $x + 2y \ge 1$

$F(x,y) = 0$, for $x + 2y < 1$,

then $F(x,y)$ cannot be joint distribution function of variables X and Y.

(ii) Show that, with usual notation : for all x, y,

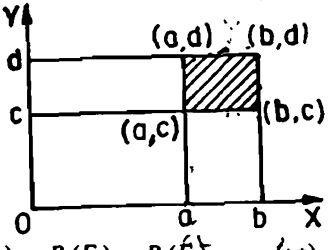
$$F_X(x) + F_Y(y) - 1 \le F_{XY}(x,y) \le \sqrt{F_X(x)\,F_Y(y)}$$

[Delhi Univ. B.Sc. (Maths Hons.), 1985]

Solution. *(i)* Let us define the events :

$$A : \{X \le a\} \ ; \ B : \{X \le b\} \ ; \ C = \{Y \le c\} \ ; \ D = \{Y \le d\} \ ;$$

for $a < b$; $c < d$.



$$P(a < X \le b \cap c < Y \le d)$$

$$= P[(B - A) \cap (D - C)]$$

$$= P[B \cap (D - C) - A \cap (D - C)] \quad \ldots(*)$$

(By distributive property of sets)

We know that if $E \subset F \Rightarrow E \cap F = E$, then

$$P(F - E) = P(\bar{E} \cap F) = P(F) - P(E \cap F) = P(F) - P(E) \quad \ldots(**)$$

Obviously $A \subset B \ \Rightarrow \ [A \cap (D - C)] \subset [B \cap (D - C)]$

Hence using (**), we get from (*)

$$P(a < X \le b \cap c < Y \le d) = P[B \cap (D - C)] - P[A \cap (D - C)]$$

$$= P[(B \cap D) - (B \cap C)] - P[(A \cap D) - (A \cap C)]$$

$$= P(B \cap D) - P(B \cap C) - P(A \cap D) + P(A \cap C) \quad \ldots (***)$$

[On using (**), since $C \subset D \Rightarrow (B \cap C) \subset (B \cap D)$ and $(A \cap C) \subset (A \cap D)$]

We have :

$P(B \cap D) = P[X \leq b \cap Y \leq d] = F(b, d)$.

Similarly

$P(B \cap C) = F(b, c)$; $P(A \cap D) = F(a, d)$ and $P(A \cap C) = F(a, c)$

Substituting in (***), we get :

$$P\{a < X \leq b \cap c < Y \leq d) = F(b, d) - F(b, c) - F(a, d) + F(a, c) \quad \ldots (1)$$

We are given $\left.\begin{array}{ll} F(x, y) = 1, & \text{for } x + 2y \geq 1 \\ = 0, & \text{for } x + 2y < 1 \end{array}\right\} \quad \ldots (2)$

In (1) let us take : $a = 0$, $b = 1/2$, ; $c = 1/4$, $d = 3/4$ s.t. $a < b$ and $c < d$. Then using (2) we get :

$F(b, d) = 1$; $F(b, c) = 1$; $F(a, d) = 1$; $F(a, c) = 0$.

Substituting in (1) we get :

$P(a < X \leq b \cap c < Y \leq d) = 1 - 1 - 1 + 0 = -1$;

which is not possible since $P(.) \geq 0$.

Hence $F(x, y)$ defined in (2) cannot be the distribution function of variates X and Y.

(ii) Let us define the events : $A = \{X \leq x\}$; $B = \{Y \leq y\}$

$$\left.\begin{array}{l} \textit{Then } P(A) = P(X \leq x) = F_X(x) ; \; P(B) = P(Y \leq y) = F_Y(y) \\ \text{and } P(A \cap B) = P(X \leq x \cap Y \leq y) = F_{XY}(x, y) \end{array}\right\} \quad \ldots (3)$$

$(A \cap B) \subset A \Rightarrow P(A \cap B) \leq P(A) \Rightarrow F_{XY}(x, y) \leq F_X(x)$

$(A \cap B) \subset B \Rightarrow P(A \cap B) \leq P(B) \Rightarrow F_{XY}(x, y) \leq F_Y(y)$

Multiplying these inequalities we get :

$$F^2_{X,Y}(x, y) \leq F_X(x) F_Y(y) \Rightarrow F_{XY}(x, y) \leq \sqrt{F_X(x) F_Y(y)} \quad \ldots (4)$$

Also $P(A \cup B) \leq 1 \Rightarrow P(A) + P(B) - P(A \cap B) \leq 1$

$\Rightarrow P(A) + P(B) - 1 \leq P(A \cap B)$

$$\Rightarrow F_X(x) + F_Y(y) - 1 \leq F_{XY}(x, y) \quad \ldots (5)$$

From (4) and (5) we get :

$F_X(x) + F_Y(y) - 1 \leq F_{XY}(x, y) \leq \sqrt{F_X(x) F_Y(y)}$, as required.

Example 5·30. *If X and Y are two random variables having joint density function*

$$f(x, y) = \frac{1}{8}(6 - x - y) \; ; \; 0 < x < 2, \; 2 < y < 4$$

$$= 0, \quad \textit{otherwise}$$

Find (i) $P(X < 1 \cap Y < 3)$, *(ii)* $P(X + Y < 3)$ *and (iii)* $P(X < 1 | Y < 3)$

(Madras Univ. B.Sc., Nov. 1986)

Solution. We have

(i) $$P(X<1 \cap Y<3) = \int_{-\infty}^{1}\int_{-\infty}^{3} f(x,y)\,dx\,dy$$

$$= \int_0^1 \int_2^3 \frac{1}{8}(6-x-y)\,dx\,dy = \frac{3}{8}$$

(ii) The probability that $X+Y$ will be less than 3 is

$$P(X+Y<3) = \int_0^1 \int_2^{3-x} \frac{1}{8}(6-x-y)\,dx\,dy = \frac{5}{24}$$

(iii) The probability that $X<1$ when it is known that $Y<3$ is

$$P(X<1 \mid Y<3) = \frac{P(X<1 \cap Y<3)}{P(Y<3)} = \frac{3/8}{5/8} = \frac{3}{5}$$

$$\left[P(Y<3) = \int_0^2 \int_2^3 \frac{1}{8}(6-x-y)\,dx\,dy = \frac{5}{8}\right]$$

Example 5·31. *If the joint distribution function of X and Y is given by :*

$$F(x,y) = 1 - e^{-x} - e^{-y} + e^{-(x+y)};\quad x>0,\ y>0$$

$$= 0; \qquad \textit{elsewhere}$$

(a) Find the marginal densities of X and Y.

(b) Are X and Y independent ?

(c) Find $P(X \le 1 \cap Y \le 1)$ *and* $P(X+Y \le 1)$. **(I.C.S., 1989)**

Solution. *(a) & (b)* The joint p.d.f. of the r.v.'s (X, Y) is given by:

$$f_{XY}(x,y) = \frac{\partial^2 F(x,y)}{\partial x\,\partial y} = \frac{\partial}{\partial x}\left[e^{-y} - e^{-(x+y)}\right]$$

$$= e^{-(x+y)};\quad x \ge 0,\ y \ge 0$$

$$= 0; \qquad \text{otherwise} \qquad \ldots (i)$$

We have

$$f_{XY}(x,y) = e^{-x}.e^{-y} = f_X(x) f_Y(y) \qquad \ldots (ii)$$

where $$f_X(x) = e^{-x};\ x \ge 0; \qquad f_Y(y) = e^{-y};\ y \ge 0 \qquad \ldots (iii)$$

(ii) $\Rightarrow$ X and Y are independent,

and *(iii)* gives the marginal p.d.f.'s of X and Y.

(c) $$P(X \le 1 \cap Y \le 1) = \int_0^1 \int_0^1 f(x,y)\,dx\,dy$$

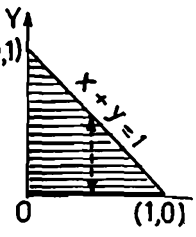
$$= \left(\int_0^1 e^{-x}\,dx\right)\left(\int_0^1 e^{-y}\,dy\right)$$

$$= \left(1 - e^{-1}\right)^2$$

$$P(X+Y\le 1)=\iint_{x+y\le 1} f(x,y)=\int_0^1\left(\int_0^{1-x} f(x,y)\,dy\right)dx$$

$$=\int_0^1\left[e^{-x}\int_0^{1-x}e^{-y}\,dy\right]dx$$

$$=\int_0^1 e^{-x}\left(1-e^{-(1-x)}\right)dx=1-2e^{-1}$$



Example 5·32. *Joint distribution of X and Y is given by*

$$f(x,y)=4xy\,e^{-(x^2+y^2)};\quad x\ge 0,\ y\ge 0.$$

Test whether X and Y are independent.

For the above joint distribution, find the conditional density of X given $Y=y$. **(Calicut Univ. B.Sc., 1986)**

Solution. Joint p.d.f. of X and Y is

$$f(x,y)=4xy\,e^{-(x^2+y^2)};\quad x\ge 0,\ y\ge 0.$$

Marginal density of X is given by

$$f_1(x)=\int_0^\infty f(x,y)\,dy=\int_0^\infty 4xy\,e^{-(x^2+y^2)}\,dy$$

$$=4x\,e^{-x^2}\int_0^\infty y\,e^{-y^2}\,dy$$

$$=4x\,e^{-x^2}\cdot\int_0^\infty e^{-t}\cdot\frac{dt}{2}\qquad\left(Put\ y^2=t\right)$$

$$=2x\cdot e^{-x^2}\left|-e^{-t}\right|_0^\infty$$

$$\Rightarrow\quad f_1(x)=2x\,e^{-x^2};\quad x\ge 0$$

Similarly, the marginal p.d.f. of Y is given by

$$f_2(y)=\int_0^\infty f(x,y)\,dx=2y\,e^{-y^2};\quad y\ge 0$$

Since $f(x,y)=f_1(x)\cdot f_2(y)$, X and Y are independently distributed. The conditional distribution of X for given Y is given by :

$$f(X = x \mid Y = y) = \frac{f(x, y)}{f_2(y)}$$
$$= 2x\, e^{-x^2}; \quad x \geq 0.$$

EXERCISE 5(e)

1. (a) Two fair dice are tossed simultaneously. Let X denote the number on the first die and Y denote the number on the second die.

(i) Write down the sample space of this experiment.

(ii) Find the following probabilities :

(1) $P(X + Y = 8)$, *(2)* $P(X + Y \geq 8)$, *(3)* $P(X = Y)$, *(4)* $P(X + Y = 6 \mid Y = 4)$, *(5)* $P(X - Y = 2)$.

(Sardar Patel Univ, B.Sc., 1991)

2. *(a)* Explain the concepts *(i)* conditional probability, *(ii)* random variable, *(iii)* independence of random variables, and *(iv)* marginal and conditional probability distributions.

(b) Explain the notion of the joint distribution of two random variables. If $F(x, y)$ be the joint distribution function of X and Y, what will be the distribution functions for *the marginal distribution of X and Y ?*

What is meant by the *conditional distribution* of Y under the condition that $X = x$? Consider separately the cases where *(i)* X and Y are both discrete and *(ii)* X and Y are both continuous.

3. The joint probability distribution of a pair of random variables is given by the following table :-

Y \ X	1	2	3
1	0.1	0.1	0.2
2	0.2	0.3	0.1

Find :

(i) The marginal distributions.

(ii) The conditional distribution of X given $Y = 1$.

(iii) $P\{(X + Y) < 4\}$.

4. *(a)* What do you mean by marginal and conditional distributions ? The following table represents the joint probability distribution of the discrete random variable (X, Y)

Y \ X	1	2	3
1	1/12	1/6	0
2	0	1/9	1/5
3	1/18	1/4	2/15

(i) Evaluate marginal distribution of X.

(ii) Evaluate the conditional distribution of Y given $X = 2$.

(Aligarh Univ. B.Sc., 1992)

(b) Two discrete random variables X and Y have

$$P(X=0, Y=0)=\frac{2}{9}\ ;\ P(X=0, Y=1)=\frac{1}{9}$$

$$P(X=1, Y=0)=\frac{1}{9}\ ;\ P(X=1, Y=1)=\frac{5}{9}$$

Examine whether X and Y are independent.

(Kerala Univ. B.Sc., Oct. 1987)

5. *(a)* Let the joint p.m.f. of X_1 and X_2 be

$$p(x_1, x_2)=\frac{x_1+x_2}{21}\ ;\ x_1=1,2,3\ ;\ x_2=1,2$$

$$=0, \text{otherwise}$$

Show that marginal p.m.f.'s of X_1 and X_2 are

$$p_1(x_1)=\frac{2x_1+3}{21};\ x_1=1,2,3\ ;\quad p_2(x_2)=\frac{6+3x_2}{21}\ ;\ x_2=1,2$$

(b) Let

$$f(x_1, x_2)=C(x_1x_2+e^{x_1})\ ;\ 0<(x_1, x_2)<1$$

$$=0, \text{ elsewhere}$$

(i) Determine C.

(ii) Examine whether X_1 and X_2 are stochastically independent.

Ans. *(i)* $C=\dfrac{4}{4e-3}$, *(ii)* $g(x_1)=C\left(\frac{1}{2}x_1+e^{x_1}\right)$, $g(x_2)=C\left(\frac{1}{2}x_2+e-1\right)$

Since $g(x_1).g(x_2)\neq f(x_1, x_2)$, X_1 and X_2 are not stochastically independent.

6. Find k so that $f(x, y)=kxy$, $1\leq x\leq y\leq 2$ will be a probability density function. **(Mysore Univ. B.Sc., 1986)**

Hint. $\iint f(x, y)\,dx\,dy=1 \Rightarrow k\int_1^2 x\left(\int_x^2 y\,dy\right)dx=1 \Rightarrow k=8/9$

7. *(a)* If $f(x, y)=e^{-(x+y)}\ ;\ x\geq 0,\ y\geq 0$

$$=0, \textit{elsewhere}$$

is the joint probability density function of random variables X and Y, find

(i) $P(X<1)$, *(ii)* $P(X>Y)$, and *(iii)* $P(X+Y<1)$.

Ans. *(i)* $1-\dfrac{1}{e}$, *(ii)* $\dfrac{1}{2}$ and *(iii)* $1-\dfrac{2}{e}$

(b) The joint frequency function of (X, Y) is given to be

$$f(x,y) = A e^{-x-y} \; ; \; 0 \le x \le y, \; 0 \le y < +\infty$$
$$= 0 \quad ; \text{ otherwise}$$

(i) Determine A.

(ii) Find the marginal density function of X.

(iii) Find the marginal density function of Y.

(iv) Examine if X and Y are independent.

(v) Find the conditional density function of Y given $X = 2$.

[Madras Univ. B.Sc. (Main Stat.), 1992]

(c) Suppose that the random variables X and Y have the joint p.d.f.

$$f(x,y) = \begin{cases} kx(x-y), & 0 < x < 2, \; -x < y < x \\ 0, & \textit{elsewhere}. \end{cases}$$

(i) Evaluate the constant k.

(ii) Find the marginal probability density functions of the random variables.

(South Gujarat Univ. B.Sc., 1988)

8. *(a)* Two-dimensional random variable (X, Y) have the joint density

$$f(x,y) = 8xy, \; 0 < x < y < 1$$
$$= 0, \text{ otherwise}$$

(i) Find. $P(X < 1/2 \cap Y < 1/4)$.

(ii) Find the marginal and conditional distributions.

(iii) Are X and Y independent? Give reasons for your answer.

(South Gujarat Univ. B.Sc., 1992)

Ans. $f_1(x) = 4x(1-x^2), 0 < x < 1$ | $f_1(x \mid y) = 2x/y^2 \quad ; 0 < x < y, 0 < y < 1$

$= 0$, *otherwise*

$f_2(y) = 4y^3, 0 < y < 1$ | $f_2(y \mid x) = 2y/(1-x^2)$; $x < y < 1, 0 < x < 1$

9. *(a)* The random variables X and Y have the joint density function :

$$f(x,y) = 2, \text{ if } x + y \le 1, \; x \ge 0 \text{ and } y \ge 0$$
$$= 0, \; \textit{otherwise}$$

Find the conditional distribution of Y, given $X = x$.

(Calcutta Univ. B.Sc. (Hons.), 1984)

(b) The random variables X and Y have the joint distribution given by the probability density function :

$$f(x,y) = \begin{cases} 6(1-x-y), \text{ for } x > 0, \; y > 0, \; x + y < 1 \\ 0, \; \textit{elsewhere} \end{cases}$$

Find the marginal distributions of X and Y. Hence examine if X and Y are independent. **[Calcutta Univ. B.Sc. (Hons.), 1986]**

10. If the joint distribution function of X and Y is given by

$$F(x,y) = (1 - e^{-x})(1 - e^{-y}) \quad \text{for } x > 0, \; y > 0$$
$$= 0, \text{ elsewhere}$$

Find $P(1 < X < 3,\ 1 < Y < 2)$. **[Delhi Univ. M.A.(Econ.), 1988]**

Hint. Reqd. Prob. $= \left(\int_1^3 e^{-x}\,dx\right)\left(\int_1^2 e^{-y}\,dy\right) = \left(1 - e^{-3}\right)\left(1 - e^{-2}\right)$

11. Let X and Y be two random variables with the joint probability density function

$$f(x,y) = \begin{cases} 8xy, & 0 < x \le y < 1 \\ 0, & \textit{otherwise} \end{cases}$$

Obtain :

(i) the joint distribution function of X and Y.

(ii) the marginal probability density function of Y; and

(iii) $P\left(X \le \frac{1}{4} \mid \frac{1}{2} < Y \le 1\right)$.

12. Let X and Y be jointly distributed with p.d.f.

$$f(x,y) = \begin{cases} \frac{1}{4}(1+xy), & |x| < 1,\ |y| < 1 \\ 0, & \textit{otherwise} \end{cases}$$

Show that X and Y are not independent but X^2 and Y^2 are independent.

Hint. $f_1(x) = \int_{-1}^{1} f(x,y)\,dy = \frac{1}{2},\ -1 < x < 1;$

$$f_2(y) = \int_{-1}^{1} f(x,y)\,dx = \frac{1}{2},\ -1 < y < 1$$

Since $f(x,y) \neq f_1(x) f_2(y)$, X and Y are not independent. However,

$$P(X^2 \le x) = P(|X| \le \sqrt{x}) = \int_{-\sqrt{x}}^{\sqrt{x}} f_1(x)\,dx = \sqrt{x}$$

$$P(X^2 \le x \cap Y^2 \le y) = P(|X| \le \sqrt{x} \cap |Y| \le \sqrt{y})$$

$$= \int_{-\sqrt{x}}^{\sqrt{x}} \left[\int_{-\sqrt{y}}^{\sqrt{y}} f(u,v)\,dv \right] du$$

$$= \sqrt{x}\cdot\sqrt{y}$$

$$= P(X^2 \le x) \cdot P(Y^2 \le y)$$

$\Rightarrow$ X^2 and Y^2 are independent.

13. *(a)* The joint probability density function of the two dimensional random variable (X, Y) is given by :

$$f(x,y)=\begin{cases} x^3y^3/16\ , & 0\le x\le 2\ ,\ 0\le y\le 2 \\ 0\ , & \textit{elsewhere}\end{cases}$$

Find the marginal densities of X and Y. Also find the cumulative distribution functions for X and Y. **(Annamalai Univ. B.E., 1986)**

Ans. $f_X(x)=\frac{x^3}{4}\ ;\ 0\le x\le 2\ ;\quad f_Y(y)=\frac{y^3}{4}\ ;\ 0\le y\le 2$

$$F_X(x)=\begin{cases}0 & ;\ x<o\\ x^4/16 & ;\ 0\le x\le 2\\ 1 & ;\ x>2\end{cases}\qquad F_Y(y)=\begin{cases}0 & ;\ y<0\\ y^4/16 & ;\ 0\le y\le 2\\ 1 & ;\ y>2\end{cases}$$

(b) The joint probability density function of the two dimensional random variable (X, Y) is given by :

$$f(x,y)=\begin{cases}\frac{8}{9}xy\ , & 1\le x\le y\le 2\\ 0\ , & \textit{elsewhere}\end{cases}$$

(i) Find the marginal density functions of X and Y,

(ii) Find the conditional density function of Y given $X = x$, and conditional density funciton of X given $Y = y$.

[Madras Univ. B.Sc. (Stat. Main), 1987]

Ans. *(i)* $f_X(x)=\int_x^2 f(x,y)\,dy=\frac{4}{9}x\left(4-x^2\right)\ ;\ 1\le x\le 2$

$$=0\qquad ;\ \text{otherwise}$$

$$f_Y(y)=\int_1^y f(x,y)\,dx=\frac{4}{9}y\left(y^2-1\right)\ ;\ 1\le y\le 2$$

$$f_{X|Y}(x\,|\,y)=\frac{2x}{y^2-1}\qquad ;\ 1\le x\le y$$

$$f_{Y|X}(y\,|\,x)=\frac{f(x,y)}{f_X(y)}=\frac{2y}{4-x^2}\ ;\ x\le y\le 2$$

14. The two random variables X and Y have, for $X = x$ and $Y = y$, the joint probability density function :

$$f(x,y)=\frac{1}{2x^2y}\ ,\ \text{for } 1\le x<\infty \text{ and } \frac{1}{x}<y<x$$

Derive the marginal distributions of X and Y. Further obtain the conditional distribution of Y for $X = x$ and also that of X given $Y = y$.

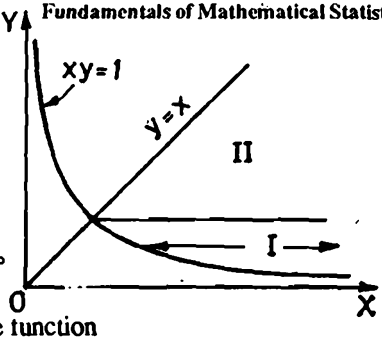
(Civil Services Main, 1986)

Hint. $f_X(x)=\int_y f(x,y)\,dy=\int_{1/x}^{x} f(x,y)\,dy$

$$f_Y(y) = \int_x f(x,y)\,dx$$

$$= \int_{1/y}^{\infty} f(x,y)\,dx; \quad 0 \le y \le 1$$

$$= \int_{y}^{\infty} f(x,y)\,dx; \quad 1 \le y < \infty$$



15. Show that the conditions for the function

$$f(x,y) = k \exp\left[A x^2 + 2Hxy + B y^2\right], \quad -\infty < (x,y) < \infty$$

to be a bivariate p.d.f. are

(i) $A \le 0$, *(ii)* $B \le 0$ *(iii)* $AB - H^2 \ge 0$.

Further show that under these conditions

$$k = \frac{1}{\pi}\left(AB - H^2\right)^{1/2}$$

Hint. $f(x,y)$ will represent the p.d.f. of a bivariate distriibution if and only if

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$$

$$\Rightarrow \quad k\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[A x^2 + 2Hxy + B y^2\right]dx\,dy = 1 \qquad \ldots (*)$$

We have

$$A x^2 + 2Hxy + B y^2 = A\left[x^2 + \frac{2H}{A}xy + \frac{B}{A}y^2\right]$$

$$= A\left[\left(x + \frac{H}{A}y\right)^2 + \frac{AB - H^2}{A^2}\cdot y^2\right] \qquad \ldots (**)$$

Similarly, we can write

$$A x^2 + 2Hxy + B y^2 = B\left[\left(y + \frac{H}{B}x\right)^2 + \frac{AB - H^2}{B^2}x^2\right] \qquad \ldots (***)$$

Substituting from (**) and (***) in (*) we observe that the double integral on the left hand side will converge if and only if

$$A \le 0,\ B \le 0 \text{ and } AB - H^2 \ge 0,$$

as desired.

Let us take $A = -a$; $B = -b$; $H = h$ so that $AB - H^2 = ab - h^2$, where $a > 0, b > 0$.

Substituting in (*), we get

$$k\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left[-\frac{ab-h^2}{a}y^2-\frac{1}{a}(-ax+hy)^2\right]dx\,dy=1$$

$$\Rightarrow \quad k\int_{-\infty}^{\infty}\left[\exp\left(-\frac{ab-h^2}{a}y^2\right)\cdot\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{a}(ax-hy)^2\right\}dx\right]dy$$

$$= 1 \quad \ldots (****)$$

(By Fubini's theorem)

Now $$\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{a}(ax-hy)^2\right\}dx=\int_{-\infty}^{\infty}\exp\left(-\frac{u^2}{a}\right)\frac{du}{a}$$

$(ax-hy=u)$

$$=\frac{1}{a}\sqrt{\pi}\sqrt{a}=\sqrt{\frac{\pi}{a}}$$

$$\left(\because \int_{-\infty}^{\infty} e^{-c^2u^2}\,du=\frac{\sqrt{\pi}}{c}\right)$$

Hence from (****), we get

$$k\sqrt{\frac{\pi}{a}}\int_{-\infty}^{\infty}\exp\left\{-\frac{ab-h^2}{a}y^2\right\}dy=1$$

$$\Rightarrow \quad k\sqrt{\frac{\pi}{a}}\cdot\sqrt{\frac{\pi a}{ab-h^2}}=1$$

$$\Rightarrow \quad k=\frac{1}{\pi}\sqrt{ab-h^2}=\frac{1}{\pi}\sqrt{AB-H^2}\,.$$

OBJECTIVE TYPE QUESTIONS

I. Which of the following statements are TRUE or FALSE.

(i) Given a continuous random variable X with probability density function $f(x)$, then $f(x)$ cannot exceed unity.

(ii) A random variable X has the following probability density function :

$$f(x)=x,\ 0<x<1$$
$$=0, \text{ elsewhere}$$

(iii) The function defined as

$$f(x)=|x|,\ -1<x<1$$
$$=0, \text{ elsewhere}$$

is a possible probability density function.

(iv) The following represents joint probability distribution.

Y \ X	1	2	3
−1	1/9	1/18	1/18
0	1/18	2/9	3/9
1	1/8	1/18	1/18

II. Fill in the blanks :

(i) If $p_1(x)$ and $p_2(y)$ be the marginal probability functions of two independent discrete random variablies X and Y, then their joint probability function

$$p(x, y) = \ldots$$

(ii) The function $f(x)$ defined as

$$f(x) = |x|, \quad -1 < x < 1$$
$$= 0, \text{ elsewhere}$$

is a possible .....

5·6. Transformation of One-dimensional Random Variable. Let X be a random variable defined on the event space S and let $g(\cdot)$ be a function such that $Y = g(X)$ is also a r.v. defined on S. In this section we shall deal with the following problem :

"Given the probability density of a r.v. X, to determine the density of a new r.v. $Y = g(X)$."

It can be proved in general that, if $g(\cdot)$ is any continuous function, then the distribution of $Y = g(X)$ is uniquely determined by that of X. The proof of this result is rather difficult and beyond the scope of this book. Here we shall consider the following, relatively simple theorem.

Theorem 5·9. *Let X be a continuous r.v. with p.d.f. $f_X(x)$. Let $y = g(x)$ be strictly monotonic (increasing or decreasing) function of x. Assume that $g(x)$ is differentiable (and hence continuous) for all x. Then the p.d.f. of the r.v. Y is given by*

$$h_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|,$$

where x is expressed in terms of y.

Proof. **Case (i).** *$y = g(x)$ is strictly increasing function of x (i.e., $dy/dx > 0$).* The *d.f.* of Y is given by

$$H_Y(y) = P(Y \leq y) = P[g(X) \leq y] = P(X \leq g^{-1}(y)],$$

the inverse exists and is unique, since $g(\cdot)$ is strictly increasing.

$$\therefore \quad H_Y(y) = F_X\left[g^{-1}(y)\right], \text{ where } F \text{ is the d.f. of } X$$
$$= F_X(x) \qquad \left[\because y = g(x) \Rightarrow g^{-1}(y) = x\right]$$

Differentiating w.r.t. y, we get

$$h_Y(y) = \frac{d}{dy}[F_X(x)] = \frac{d}{dx}\left(F_X(x)\right)\frac{dx}{dy}$$
$$= f_X(x)\frac{dx}{dy} \qquad \ldots (*)$$

Case (ii). $y = g(x)$ *is strictly monotonic decreasing.*

$$H_Y(y) = P(Y \le y) = P[g(X) \le y] = P[X \ge g^{-1}(y)]$$
$$= 1 - P\left[X \le g^{-1}(y)\right] = 1 - F_X\left[g^{-1}(y)\right] = 1 - F_X(x),$$

where $x = g^{-1}(y)$, the inverse exists and is unique. Differentiating w.r.t. y, we get

$$h_Y(y) = \frac{d}{dx}\left[1 - F_X(x)\right]\frac{dx}{dy} = -f_X(x)\cdot\frac{dx}{dy}$$

$$= f_X(x)\cdot\frac{-dx}{dy} \qquad \ldots(**)$$

Note that the algebraic sign (–ive) obtained in (**) is correct, since y is a decreasing function of $x \Rightarrow x$ is a decreasing function of $y \Rightarrow dx/dy < 0$.

The results in (*) and (**) can be combined to give

$$h_Y(y) = f_X(x)\left|\frac{dx}{dy}\right|$$

Example 5·33. *If the cumulative distribution function of X is $F(x)$, find the cumulative distribution function of*

(i) $Y = X + a$, *(ii)* $Y = X - b$, *(iii)* $Y = aX$,

(iv) $Y = X^3$, and *(v)* $Y = X^2$

What are the corresponding probability density functions ?

Solution. Let $G(\cdot)$ be the c.d.f. of Y. Then

(i) $G(x) = P(Y \le x) = P[X + a \le x] = P[X \le x - a] = F(x - a)$

(ii) $G(x) = P(Y \le x) = P[X - b \le x] = P[X \le x + b] = F(x + b)$

(iii) $G(x) = P[aX \le x] = P\left[X \le \frac{x}{a}\right]$, $a > 0$

$$= F\left(\frac{x}{a}\right), \text{ if } a > 0$$

and $$G(x) = P\left[X \ge \frac{x}{a}\right] = 1 - P\left[X < \frac{x}{a}\right]$$

$$= 1 - F\left(\frac{x}{a}\right), \text{ if } a < 0$$

(iv) $G(x) = P[Y \le x] = P[X^3 \le x] = P\left[X \le x^{1/3}\right] = F\left(x^{1/3}\right)$

(v) $G(x) = P\left[X^2 \le x\right] = \left[-x^{1/2} \le X \le x^{1/2}\right]$

$$= P\left[X \le x^{1/2}\right] - P\left[X \le -x^{1/2}\right]$$

$= 0$, if $x < 0$

$= F(\sqrt{x}) - F(-\sqrt{x} - 0)$, if $x > 0$

Variable	*d.f.*	*p.d.f.*
X	$F(x)$	$f(x)$
$X - a$	$F(x+a)$	$f(x+a)$
aX	$F(x/a)$ $a > 0$ $1 - F(x/a)$, $a < 0$	$(1/a)\, f(x/a)$, $a > 0$ $(-1/a)\, f(x/a)$, $a < 0$
X^2	$F(\sqrt{x}) - F(-\sqrt{x} - 0)$ for $x > 0$ 0, *otherwise*	$\frac{1}{2(\sqrt{x})}[f\sqrt{x} + f(-\sqrt{x})]$ for $x > 0$ $= 0$ for $x \le 0$
X^3	$F(x^{1/3})$	$\frac{1}{3} f(x^{1/3}) \cdot \frac{1}{x^{2/3}}$

EXERCISE 5(f)

1. *(a)* A random variable X has $F(x)$ as its distribution function [$f(x)$ is the density function]. Find the distribution and the density functions of the random variable :

(i) $Y = a + bX$, a and b are real numbers, *(ii)* $Y = X^{-1}$, [$P(X = 0) = 0$], *(iii)* $Y = \tan X$, and *(iv)* $Y = \cos X$.

(b) Let $f(x) = \begin{cases} \frac{1}{2}, & -1 < x < 1 \\ 0, & \textit{elsewhere} \end{cases}$

be the p.d.f. of the r.v. X. Find the distribution function and the p.d.f. of $Y = X^2$.

[Delhi Univ. B.Sc. (Maths Hons.), 1988]

Hint. $F(x) = P(X \le x) = \int_{-1}^{x} f(x)\, dx = \frac{1}{2}(x + 1)$ ... (*)

Distribution function $G(\cdot)$ of $Y = X^2$ is given by :

$G_Y(x) = F(\sqrt{x}) - F(-\sqrt{x})$; $x > 0$ [c.f. Example 5·33 *(v)*]

$= \frac{1}{2}(\sqrt{x} + 1) - \frac{1}{2}(-\sqrt{x} + 1)$ [From (*)]

$= \sqrt{x}$; $0 < x < 1$

(As $-1 < x < 1$, $Y = X^2$ lies between 0 and 1)

p.d.f. of $Y = X^2$ is $g(x) = G'(x) = \frac{1}{2\sqrt{x}}$; $0 < x < 1$

2. Let X be a continuous random variable with p.d.f. $f(x)$. Let $Y = X^2$. Show that the random variable Y has p.d.f. given by

$$g(y) = \begin{cases} \frac{1}{2\sqrt{y}} [f(\sqrt{y}) + f(-\sqrt{y})], & y > 0 \\ 0, & y \leq 0 \end{cases}$$

3. Find the distribution and denisitiy functions for *(i)* $Y = aX + b$, $a \neq 0$, b real, *(ii)* $Y = e^X$, assuming that $F(x)$ and $f(x)$, the distribution and the density of X are known.

Ans. *(i)* $\left.\begin{array}{ll} G(y) = F[(y-b)/a], & \text{if } a > 0 \\ G(y) = 1 - F[(y-b)/a], & \text{if } a < 0 \end{array}\right\}$ $g_1(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right)$

(ii) $\left.\begin{array}{ll} G(y) = F(\log y), & y > 0 \\ = 0, & y \leq 0 \end{array}\right\}$ $\left.\begin{array}{ll} g(y) = \frac{1}{y} f(\log y), & y > 0 \\ = 0, & y \leq 0 \end{array}\right.$

4. *(a)* The random variable X has an exponential distribution

$$f(x) = e^{-x}, \quad 0 < x \leq \infty$$

Find the density function of the variable *(i)* $Y = 3X + 5$, *(ii)* $Y = X^3$.

(b) Suppose that X has p.d.f.,

$$f(x) = 2x, \quad 0 < x < 1$$
$$= 0, \quad \textit{elsewhere}$$

Find the p.d.f. of $Y = 3X + 1$.

Ans. $g(y) = \frac{2}{9}(y-1), \quad 1 < y < 4$

5. Let X be a random variable with p.d.f.

$$f(x) = \frac{2}{9}(x+1) \qquad -1 < x < 2$$
$$= 0, \quad \textit{elsewhere}$$

Find the p.d.f. of $U = X^2$. **[Poona Univ. B.E., 1992]**

6. Let the p.d.f. of X be

$$f(x) = \frac{1}{6}, \quad -3 \leq x \leq 3$$
$$= 0, \quad \textit{elsewhere}$$

Find the p.d.f. of $Y = 2X^2 - 3$.

7. Let X be a random variable with the distribution function :

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

Determine the distribution function $F_Y(y)$ of the random variable $Y = \sqrt{X}$ and hence compute mean of Y. **[Calcutta Univ. B.A.(Hons.), 1986]**

5·7. Transformation of Two-dimensional Random Variable. In this section we shall consider the problem of change of variables in the two-dimensional

case. Let the *r.v.'s* U and V by the transformation $u = u(x, y)$, $v = v(x, y)$, where u and v are continuously differentiable functions for which Jacobian of transformation

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix}$$

is either > 0 or < 0 throughout the (x, y) plane so that the inverse transformation is uniquiely given by $x = x(u, v)$, $y = y(u, v)$.

Theorem 5·10. *The joint p.d.f.* $g_{UV}(u, v)$ *of the transformed variables U and V is given by*

$$g_{UV}(u, v) = f_{XY}(x, y) \cdot |J|$$

where $|J|$ *is the modulus value of the Jacobian of transformation and* $f(x, y)$ *is expressed in terms of u and v.*

Proof. $P(x < X \leq x + dx, y < Y \leq y + dy)$

$$= P(u < U \leq u + du, v < V \leq v + dv)$$

$$\Rightarrow \quad f_{XY}(x, y)\, dx\, dy = g_{UV}(u, v)\, du\, dv$$

$$\Rightarrow \quad g_{UV}(u, v)\, du\, dv = f_{XY}(x, y) \left| \frac{\partial(x, y)}{\partial(u, v)} \right| du\, dv$$

$$\Rightarrow \quad g_{UV}(u, v) = f_{XY}(x, y) \left| \frac{\partial(x, y)}{\partial(u, v)} \right| = f_{XY}(x, y) |J|$$

Theorem 5·11. *If X and Y are independent continuous r.v.'s, then the p.d.f. of* $U = X + Y$ *is given by*

$$h(u) = \int_{-\infty}^{\infty} f_X(v) f_Y(u - v)\, dv$$

Proof. Let $f_{XY}(x, y)$ be the joint *p.d.f.* of independent continuous *r.v.'s* X and Y and let us make the transformation :

$$u = x + y, \quad v = x \quad \Rightarrow \quad x = v, \quad y = u - v$$

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1$$

Thus the joint p.d.f. of *r.v.'s* U and V is given by

$$g_{UV}(u, v) = f_{XY}(x, y) |J|$$

$$= f_X(x) \cdot f_Y(y) |J|$$

(Since X and Y are independent)

$$= f_X(v) \cdot f_Y(u - v)$$

The marginal density of U is given by

$$h(u) = \int_{-\infty}^{\infty} g_{UV}(u, v)\, dv$$

$$= \int_{-\infty}^{\infty} f_X(v)\, f_Y(u - v)\, dv$$

Remark. The function $h(\cdot)$ is given a special name and is said to be the *convolution* of $f_X(\cdot)$ and $f_Y(\cdot)$ and we write

$$h(\cdot) = f_X(\cdot) * f_Y(\cdot)$$

Example 5·34. *Let (X, Y) be a two-dimensional non-negative continuous r.v. having the joint density :*

$$f(x, y) = \begin{cases} 4xy\, e^{-(x^2+y^2)} ; & x \geq 0,\ y \geq 0 \\ 0 , & \textit{elsewhere} \end{cases}$$

Prove that the densitiy function of $U = \sqrt{X^2 + Y^2}$ is

$$h(u) = \begin{cases} 2u^3\, e^{-u^2} , & 0 \leq u < \infty \\ 0 , & \textit{elsewhere} \end{cases}$$

[Meerut Univ. M.Sc., 1986]

Solution. Let us make the transformation :

$$u = \sqrt{x^2 + y^2} \quad \text{and} \quad v = x$$

$$\Rightarrow \qquad v \geq 0,\ u \geq 0 \text{ and } u \geq v \qquad \Rightarrow \qquad u \geq 0 \text{ and } 0 \leq v \leq u$$

The Jacobian of transformation J is given by

$$\frac{1}{J} = \frac{\partial(u, v)}{\partial(x, y)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix} = -\frac{y}{\sqrt{x^2 + y^2}}$$

The joint p.d.f. of U and V is given by

$$g(u, v) = f(x, y)\, |J|$$

$$= 4xy\, e^{-(x^2+y^2)} \left| -\frac{\sqrt{x^2 + y^2}}{y} \right|$$

$$= 4x\sqrt{x^2 + y^2}\, e^{-(x^2+y^2)}$$

$$= \begin{cases} 4vu \,.\, e^{-u^2} ; & u \geq 0,\ 0 \leq v \leq u \\ 0 , & \textit{otherwise} \end{cases}$$

Hence the density function of $U = \sqrt{X^2 + Y^2}$ is

$$h(u) = \int_0^u g(u, v)\, dv = 4u e^{-u^2} \int_0^u v\, dv$$

$$= \begin{cases} 2u^3 e^{-u^2}, & u \geq 0 \\ 0, & \text{elsewhere} \end{cases}$$

Example 5·35. *Let the probability density function of the random variable* (X, Y) *be*

$$f(x, y) = \begin{cases} \alpha^{-2} e^{-(x+y)/\alpha} & ; \quad x, y > 0,\ \alpha > 0 \\ 0 & , \quad \text{elsewhere} \end{cases}$$

Find the distribution of $\frac{1}{2}(X - Y)$. **[Nagpur Univ. B.E., 1988]**

Solution. Let us make the transformation :

$$u = \frac{1}{2}(x - y) \quad \text{and} \quad v = y$$

$$\Rightarrow \qquad x = 2u + v \quad \text{and} \quad y = v$$

The Jacobian of the transformation is :

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 2 & 1 \\ 0 & 1 \end{vmatrix} = 2$$

Thus, the joint p.d.f. of the random variables (U, V) is given by :

$$g(u, v) = \begin{cases} \frac{2}{\alpha^2} e^{-(2/\alpha)(u+v)}, & -\infty < u < \infty,\ v > -2u, \text{ if } u < 0 \\ & v > 0 \text{ if } u \geq 0 \text{ and } \alpha > 0 \\ 0, & \text{elsewhere} \end{cases}$$

The marginal p.d.f. of U is given by

$$g_U(u) = \begin{cases} \int_{-2u}^{\infty} \frac{2}{\alpha^2} \exp\{-(2/\alpha)(u+v)\}\, dv \\ \quad = \frac{1}{\alpha} e^{-2u/\alpha} \quad , u < 0 \\ \int_0^{\infty} \frac{2}{\alpha} \exp\{-(2/\alpha)(u+v)\}\, dv \\ \quad = \frac{1}{\alpha} e^{-2u/\alpha} \quad , u \geq 0 \end{cases}$$

Hence

$$g_U(u) = \frac{1}{\alpha} e^{-(2/\alpha)|u|} \quad ; \quad -\infty < u < \infty$$

Example 5·36. *Given the joint density function of X and Y as*

$$f(x,y) = \frac{1}{2} x\, e^{-y} \; ; \; 0 < x < 2, \; y > 0$$

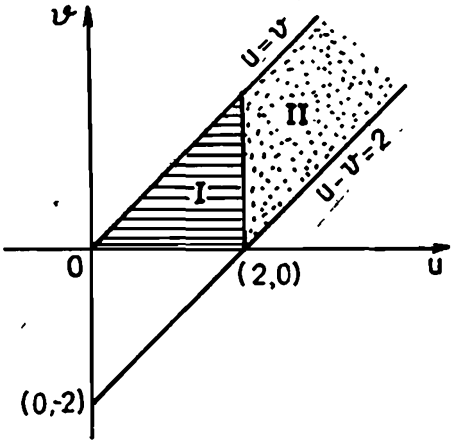
$$= 0, \quad \textit{elsewhere}$$

Find the distribution of $X + Y$.

Solution. Let us make the transformation :

$$u = x + y \text{ and } v = y \quad \Rightarrow \quad y = v, \; x = u - v$$

The Jacobian of transformation $J = \dfrac{\partial (x,y)}{\partial (u,v)} = 1$ and the region $0 < x < 2$ and $y > 0$ transforms to $0 < u - v < 2$ and $v > 0$ as shown in the following figure.



The joint density function of U and V is given by

$$g(u,v) = \frac{1}{2}(u-v)\, e^{-v} \; ; \; 0 < v < u, \; u > 0$$

To find the density of $U = X + Y$, we split the range of U into two parts (i) $0 < u \leq 2$ (region I) (ii) $u > 2$ (region II) (which is suggested by the diagram).

For $0 < u \leq 2$, (Region I) :

$$h(u) = \int_0^u g(u,v)\, dv = \frac{1}{2} \int_0^u (u-v)\, e^{-v}\, dv$$

$$= \frac{1}{2} \left| -e^{-v}(u-v) + e^{-v} \right|_{v=0}^{v=u} \quad \text{(Integration by parts)}$$

$$= \frac{1}{2}(e^{-u} + u - 1)$$

For $2 < u < \infty$, (Region II) :

$$h(u) = \frac{1}{2}\int_{u-2}^{u} (u - v)\, e^{-v}\, dv$$

$$= \frac{1}{2}\left| e^{-v}(1 + v - u)\right|_{v=u-2}^{v=u}$$

$$= \frac{1}{2}\, e^{-u}\left(1 + e^{2}\right) \qquad \text{(on simplification)}$$

Hence

$$g(u) = \begin{cases} \frac{1}{2}\left(e^{-u} + u - 1\right), & 0 < u \le 2 \\ \frac{1}{2}\, e^{-u}\left(1 + e^{2}\right), & 2 < u < \infty \\ 0\ , \ \textit{elsewhere} \end{cases}$$

MISCELLANEOUS EXERCISE ON CHAPTER FIVE

1. 4 coins are tossed. Let X be the number of heads and Y be the number of heads minus the number of tails. Find the probability function of X, the probability function of Y and $P(-2 \le Y < 4)$.

Ans. Probability function of X is

Values of X, x	0	1	2	3	4
$p_1(x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

Probability function of Y is

Values of Y, y	4	2	0	−2	−4
$p_2(y)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

$$P(-2 \le Y < 4) = \frac{4 + 6 + 4}{16} = \frac{7}{8}.$$

2. A random process gives measurements X between 0 and 1 with a probability density function

$$f(x) = 12x^3 - 21x^2 + 10x, \quad 0 \le x \le 1$$

$$= 0, \quad \textit{elsewhere}$$

(i) Find $P(X \le \frac{1}{2})$ and $P(X > \frac{1}{2})$

(ii) Find a number k such that $P(X \le k) = \frac{1}{2}$.

Ans. *(i)* 9/16, 7/16, *(ii)* $k = 0{\cdot}452$.

3. Show that for the distribution

$$dF = y_o \left[1 - \frac{|x-b|}{a} \right] dx, \quad b-a < x < b+a$$
$$= 0, \quad \textit{otherwise},$$
$$y_o = \frac{1}{a}, \quad \textit{mean} = b \text{ and } \textit{variance} = a^2/6$$

4. A ray of light is sent in a random direction towards the x-axis from a station Q (0, 1) on the y-axis and the ray meets the x-axis at a point P. Find the probability density function of the abscissa of P.

[Calcutta Univ. B.Sc.(Hons.), 1982]

5. Let X be a continuous variate with p.d.f.

$$f(x) = k(x - x^2) \; ; \quad a < x < b, \; k > 0$$

What are the possible values of a and b and what is k ?

[Delhi Univ. B.Sc.(Maths Hons.), 1989]

6. Pareto distribution with parameters r and A is given by the probability density function

$$f(x) = rA^r \frac{1}{x^{r+1}}, \quad \text{for } x \geq A$$
$$= 0, \; x < A, \; r > 0$$

Show that it has a finite nth moment if and only if $n < r$. Find the mean and variance of the distribution.

7. For a continuous random variable X, defined in the range $(0 \leq x < \infty)$, the probability distribution is such that

$$P(X \leq x) = 1 - e^{-\beta x^2}, \; \textit{where } \beta > 0$$

Find the median of the distribution. Also if m, m_o and σ denote the mean, mode and standard deviation respectively of the distribution, prove that

$$2m_o^2 - m^2 = \sigma^2 \text{ and } m_o = m\sqrt{2/\pi}$$

What is the sign of skewness of the distribution ?

8. *(a)* Two dice are rolled, $S = \{(a, b) \mid a, b = 1, 2 \ldots, 6\}$. Let X denote the sum of the two faces and Y the absolute value of their difference, *i.e.*, X is distributed over the integers 2, 3, ..., 12 and Y over 0, 1, 2, ..., 5. Assuming the dice are fair, find the probabilities that *(i)* $X = 5 \cap Y = 1$, *(ii)* $X = 7 \cap Y \geq 3$, *(iii)* $X = Y$, and *(iv)* $X + Y = 4 \cap X - Y = 2$.

Ans. *(i)* 1/8, *(ii)* 1/9, *(iii)* 0 and *(iv)* 1/18.

9. The joint probability density function of the two-dimensional variable (X, Y) is of the form

$$f(x, y) = k e^{-(x+y)}, \; 0 \leq y < x < \infty$$
$$= 0, \; \textit{elsewhere}$$

(i) Determine the constant k. *(ii)* Find the conditional probability density function $f_1(x \mid y)$ and *(iii)* Compute $P(Y \geq 3)$.

[Sardar Patel Univ. B.Sc., 1986]

(iv) Find the marginal frequency function $f_1(x)$ of X.

(v) Find the marginal frequency function $f_2(y)$ of Y.

(vi) Examine if X, Y are independent.

(vii) Find the conditional frequency function of Y given $X = 2$.

Ans. (i) $k = 1$, (ii) $f_1(x \mid y) = e^{-x}$, (iii) e^{-3}.

10. Let

$$f(x, y) = \begin{cases} \binom{y}{x} p^x (1-p)^{y-x} \dfrac{e^{-\lambda} \lambda^y}{y!} \\ \qquad ; \; x = 0, 1, 2, \ldots; \; y = 0, 1, 2, \ldots; \; \textit{with } y \geq x \\ 0, \quad \textit{elsewhere} \end{cases}$$

Find the marginal density function of X and the marginal density function of Y. Also determine whether the random variables X and Y are independent.

[I.S.I., 1987]

11. Consider the following function :

$$f(x \mid y) = \begin{cases} \dfrac{y^x \, e^{-y}}{x!}, & x = 0, 1, 2, \ldots \\ 0, & \textit{otherwise} \end{cases}$$

(i) Show that $f(x \mid y)$ is the conditional probability function of X given Y; $y \geq 0$.

(ii) If the marginal p.d.f. of Y is

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda x}, & x > 0. \\ 0 & x \leq 0, \; \lambda > 0 \end{cases}$$

what is the joint p.d.f. of X and Y ?

(iii) Obtain the marginal probability function of X.

[Delhi Univ. M.A.(Econ.), 1989]

12. The probability density function of (x_1, x_2) is given as

$$f(x_1, x_2) = \begin{cases} \theta_1 \theta_2 e^{-\theta_1 x_1 - \theta_2 x_2} & \text{if } x_1, x_2 > 0 \\ 0 & \textit{otherwise}. \end{cases}$$

Find the density function of (y_1, y_2) where

$$y_1 = \frac{2x_1}{x_2} + 1, \quad y_2 = 3x_1 + x_2 \text{ almost everywhere.}$$

[Punjab Univ. M.A.(Econ.), 1992]

13. *(a)* Let X_1, X_2 be a random sample of size 2 from a distribution with probability density function,

$$f(x) = e^{-x}, \; 0 < x < \infty$$

$= 0$, *elsewhere*

Show

$$Y_1 = X_1 + X_2 \text{ and } Y_2 = \frac{X_1}{X_1 + X_2}$$

are independent. **[Sardar Patel Univ. B.Sc., Sept. 1986]**

(b) X_1, X_2, X_3 denote random sample of size 3 drawn from the distribution:

$$f(x) = e^{-x}, \quad 0 < x < \infty$$
$$= 0, \quad \textit{elsewhere}$$

Show that

$$Y_1 = \frac{X_1}{X_1 + X_2}, \quad Y_2 = \frac{X_1 + X_2}{X_1 + X_2 + X_3} \text{ and } Y_3 = X_1 + X_2 + X_3$$

are mutually independent.

14. If the probability density function of the random varaibles X and $Y|X$ is given by

$$f(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & \textit{elsewhere} \end{cases}$$

and $f_{Y|X}(y|x) = \begin{cases} \dfrac{e^{-x}x^y}{y!}, & y \geq 0 \\ 0, & \textit{elsewhere} \end{cases}$

respectively, find the probability density function of the random variable Y.

[Jiwaji Univ. M.Sc., 1987]

15. *(a)* The random variable X and Y have a joint p.d.f. $f(x, y)$ given by

$$f(x, y) = g(x + y), \quad x > 0, \ y > 0$$
$$= 0, \quad \text{otherwise.}$$

Obtain the distribution function $H(z)$ of $Z = X + Y$ and hence show that its p.d.f. is

$$h(z) = z\, g(z), \quad z > 0$$
$$= 0 \quad z \leq 0.$$

(b) The joint density function of two random variables is given by

$$f(x, y) = e^{-(x+y)}; \ x > 0, \ y > 0.$$ Show that the p.d.f. of

$$U = \frac{X + Y}{2} \text{ is } g(u) = 4u e^{-2u}$$

[Calicut Univ. B.Sc., 1986]

16. The time X taken by a garage to repair a car is a continuous random variable with probability density function

$$f_1(x) = \begin{cases} \frac{3}{4}x(2-x), & 0 \le x \le 2 \\ 0, & \textit{elsewhere} \end{cases}$$

If, on leaving his car, a motorist goes to keep an engagement lasting for a time Y, where Y is a continuous random variable, independent of X, with probability function

$$f_2(y) = \begin{cases} \frac{1}{2}y, & 0 \le y \le 2 \\ 0, & \textit{elsewhere}; \end{cases}$$

determine the probability that the car will not be ready on his return.

[Calcutta Univ. B.A.(Hons.), 1988]

17. If X and Y are two independent random variables such that

$$f(x) = e^{-x}, \; x \ge 0 \quad \text{and} \quad g(y) = 3e^{-3y}, \; y \ge 0;$$

find the probability distribution of $Z = X/Y$.

[Madurai Univ. B.Sc., Oct. 1987]

18. The random variables X and Y are independent and their probability density functions are, respectively given by

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{\sqrt{1+x^2}}, \; |x| < 1 \quad \text{and} \quad g(y) = y e^{-y^2/2}, \; y > 0.$$

Find the joint probability density of Z and W where $Z = XY$ and $W = X$. Deduce the probability density of Z. **[Calcutta Univ. B.Sc.(Hons.), 1985]**

CHAPTER SEVEN

Theoretical Discrete Probability Distributions

7·0. Introduction. In the previous chapters we have discussed in detail the frequency distributions. In the present chapter we will discuss theoretical discrete distributions in which variables are distributed according to some definite probability law which can be expressed mathematically. The present study will also enable us to fit a mathematical model or a function of the form $y = p(x)$ to the observed data.

We have already defined distribution function, mathematical expectation, m.g.f., characteristic function and moments. This prepares us for a study of theoretical distributions. This chapter is devoted to the study of univariate (except for the multinomial) distributions like Binomial, Poisson, Negative binomial Geometric, Hypergeometric, Multinomial and Power-series distributions.

7·1. Bernoulli Distribution. A random variable X which takes two values 0 and 1, with probabilities q and p respectively, *i.e.*, $P(X = 1) = p$, $P(X = 0) = q$, $q = 1 - p$ is called a *Bernoulli variate* and is said to have a Bernoulli distribution.

Remark. Sometimes, the two values are +1, −1 instead of 1 and 0.

7·1·1. Moments of Bernoulli distribution. The r^{th} moment about origin is

$$\mu_r' = E(X^r) = 0^r \cdot q + 1^r \cdot p = p \; ; r = 1, 2, \ldots \qquad \ldots(7\cdot1^*)$$

$$\mu_1' = E(X) = p, \quad \mu_2' = E(X^2) = p$$

$$\mu_2 = \text{Var}(X) = p - p^2 = pq$$

The m.g.f. of Bernoulli variate is given by :

$$M_X(t) = e^{0\cdot t} \times P(X = 0) + e^{1\cdot t} \cdot P(X = 1) = q + pe^t \qquad \ldots(7\cdot1a)$$

Remark. Degenerate Random Variable. Sometimes we may come across a variate X which is degenerate at a point 'c', say, so that : $P(X = c) = 1$ and $= 0$ otherwise, *i.e.*, the whole mass of the variable is concentrated at a single point 'c'.

$$\text{Since } P(X = c) = 1, \text{ Var}(X) = 0.$$

Thus a degenerate r.v. X is characterised by Var $(X) = 0$.

M.g.f. of degenerate r.v. is given by

$$M_X(t) = E(e^{tX}) = e^{tc} P(X = c) = e^{ct} \qquad \ldots(7\cdot1b)$$

7·2. Binomial Distribution. Binomial distribution was discovered by James Bernoulli (1654-1705) in the year 1700 and was first published posthumously in 1713, eight years after his death). Let a random experiment be performed repeatedly and let the occurrence of an event in a trial be called a success and its non-occurrence a failure. Consider a set of n independent Bernoullian trials (n

being finite), in which the probability 'p' of success in any trial is constant for each trial. Then $q = 1 - p$, is the probability of failure in any trial.

The probability of x successes and consequently $(n - x)$ failures in n independent trials, in a specified order (say) *SSFSFFFS...FSF* (where S represents success and F failure) is given by the compound probability theorem by the expression :

$$P(SSFSFFFS...FSF) = P(S)P(S)P(F)P(S)P(F)P(F)P(F)P(S) \times \ldots \times P(F)P(S)P(F)$$

$$= p \cdot p \cdot q \cdot p \cdot q \cdot q \cdot q \cdot p \ldots q \cdot p \cdot q$$

$$= \underbrace{p \cdot p \cdot p \ldots p}_{x \text{ factors}} \quad \underbrace{q \cdot q \cdot q \ldots q}_{(n-x) \text{ factors}} = p^x q^{n-x}$$

But x successes in n trials can occur in $\binom{n}{x}$ ways and the probability for each of these ways is $p^x q^{n-x}$. Hence the probability of x successes in n trials in *any order* whatsoever is given by the addition theorem of probability by the expression:

$$\binom{n}{x} p^x q^{n-x}$$

The probability distribution of the number of successes, so obtained is called the *Binomial probability distribution*, for the obvious reason that the probabilities of 0, 1, 2, ..., n successes, *viz.*,

$q^n, \binom{n}{1} q^{n-1} p, \binom{n}{2} q^{n-2} p^2, \ldots, p^n$, are the successive terms of the binomial expansion $(q + p)^n$.

Definition. *A random variable X is said to follow binomial distribution if it assumes only non-negative values and its probability mass function is given by*

$$P(X = x) = p(x) = \begin{cases} \binom{n}{x} p^x q^{n-x}; x = 0,1,2,\ldots, n; q = 1 - p \\ 0, \textit{otherwise} \end{cases} \quad \ldots(7\cdot2)$$

The two independent constants n and p in the distribution are known as the *parameters* of the distribution. 'n' is also, sometimes, known as the degree of the binomial distribution.

Binomial distribution is a discrete distribution as X can take only the integral values, *viz.*, 0, 1, 2,..., n. Any variable which follows binomial distribution is known as *binomial variate.*

We shall use the notation $X \sim B(n, p)$ to denote that the random variable X follows binomial distribution with parameters n and p.

The probability $p(x)$ in (7·2) is also sometimes denoted by $b(x, n, p)$.

Remarks 1. This assignment of probabilities is permissible because

$$\sum_{x=0}^{n} p(x) = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (q + p)^n = 1$$

2. Let us suppose that n trials constitute an experiment. Then if this experiment is repeated N times, the *frequency function* of the binomial distribution is given by

$$f(x) = Np(x) = N\binom{n}{x} p^x q^{n-x}; x = 0, 1, 2, ..., n \quad ...(7.3)$$

and the expected frequencies of 0, 1, 2, ..., n successes are the successive terms of the binomial expansion, $N(q+p)^n$, $q + p = 1$.

3. Binomial distribution is important not only because of its wide applicability, but because it gives rise to many other probability distributions. Tables for $p(x)$ are available for various values of n and p.

4. **Physical conditions for Binomial Distribution.** We get the binomial distribution under the following experimental conditions.

(i) Each trial results in two mutually disjoint outcomes, termed as success and failure.

(ii) The number of trials 'n' is finite.

(iii) The trials are independent of each other.

(iv) The probability of success 'p' is constant for each trial.

The problems relating to tossing of a coin or throwing of dice or drawing cards from a pack of cards with replacement lead to binomial probability distribution.

Example 7·1. *Ten coins are thrown simultaneously. Find the probability of getting at least seven heads.*

Solution. p = Probability of getting a head = $\frac{1}{2}$

q = Probability of not getting a head = $\frac{1}{2}$

The probability of getting x heads in a random throw of 10 coins is

$$p(x) = \binom{10}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{10-x} = \binom{10}{x}\left(\frac{1}{2}\right)^{10}; x = 0, , 2, ..., 10$$

$\therefore$ Probability of getting at least seven heads is given by

$$P(X \geq 7) = p(7) + p(8) + p(9) + p(10)$$

$$= \left(\frac{1}{2}\right)^{10}\left\{\binom{10}{7} + \binom{10}{8} + \binom{10}{9} + \binom{10}{10}\right\}$$

$$= \frac{120 + 45 + 10 + 1}{1024} = \frac{176}{1024}.$$

Example 7·2. *A and B play a game in which their chances of winning are in the ratio 3 : 2. Find A's chance of winning at least three games out of the five games played.* **[Burdwan Univ. B.Sc. (Hons.), 1993]**

Solution. Let p be the probability that 'A' wins the game. Then we are given $p = 3/5 \Rightarrow q = 1 - p = 2/5$.

Hence, by binomial probability law, the probability that out of 5 games played, A wins 'r' games is given by :

$$P(X = r) = p(r) = \binom{5}{r} \cdot (3/5)^r (2/5)^{5-r}; r = 0, 1, 2, ..., 5$$

The required probability that 'A' wins at least three games is given by :

$$P(X \geq 3) = \sum_{r=3}^{5} \binom{5}{r} \frac{3^r \cdot 2^{5-r}}{5^5}$$

$$= \frac{3^3}{5^5}\left[\binom{5}{3} 2^2 + \binom{5}{4} . 3 \times 2 + 1.3^2 \times 1\right] = \frac{27 \times (40 + 30 + 9)}{3125} = 0{\cdot}68$$

Example 7·3. *If m things are distributed among 'a' men and 'b' women, show that the probability that the number of things received by men is odd, is*

$$\frac{1}{2}\left[\frac{(b + a)^m - (b - a)^m}{(b + a)^m}\right]$$

(Nagpur Univ B.Sc., 1989, '93)

Solution. p = Probability that a thing is received by man = $\frac{a}{a + b}$, then $q = 1 - p = 1 - \frac{a}{a + b} = \frac{b}{a + b}$, is the probability that a thing is received by woman.

The probability that out of m things exactly x are received by men and the rest by women, is given by

$$p(x) = {}^m C_x p^x q^{m-x}; \; x = 0, 1, 2, ..., m$$

The probability P that the number of things received by men is odd is given by

$$P = p(1) + p(3) + p(5) + ... = {}^mC_1 \cdot q^{m-1} \cdot p + {}^m C_3 \cdot q^{m-3} \cdot p^3 + {}^mC_5 \cdot q^{m-5} \cdot p^5 + ...$$

Now

$$(q + p)^m = q^m + {}^mC_1 \cdot q^{m-1} \cdot p + {}^mC_2 \cdot q^{m-2} p^2 + {}^mC_3 \cdot q^{m-3} \cdot p^3 + {}^mC_4 \cdot q^{m-4} \cdot p^4 + ...$$

and

$$(q - p)^m = q^m - {}^mC_1 \cdot q^{m-1} \cdot p + {}^mC_2 \cdot q^{m-2} \cdot p^2 - {}^mC_3 \cdot q^{m-3} \cdot p^3 + {}^mC_4 \cdot q^{m-4} \cdot p^4 - ...$$

$$\therefore \quad (q + p)^m - (q - p)^m = 2\left[{}^mC_1 \cdot q^{m-1} \cdot p + {}^mC_3 \cdot q^{m-3} \cdot p^3 + ...\right] = 2P$$

But $q + p = 1$ and $q - p = \frac{b - a}{b + a}$

$$\therefore \quad 1 - \left(\frac{b - a}{b + a}\right)^m = 2P \Rightarrow P = \frac{1}{2}\left[\frac{(b + a)^m - (b - a)^m}{(b + a)^m}\right]$$

Example 7·4 *An irregular six faced die is thrown and the expectation that in 10 throws it will give five even numbers is twice the expectation that it will give four even numbers. How many times in 10,000 sets of 10 throws each, would you expect it to give no even number.* **(Gujarat Univ. B.Sc. 1988)**

Solution. Let p be the probability of getting an even number in a throw of a die. Then the probability of getting x even numbers in ten throws of a die is

$$P(X = x) = \binom{10}{x} p^x q^{10-x}; x = 0, 1, 2...10$$

We are given that

$$P(X = 5) = 2P(X = 4)$$

$$i.e., \quad \binom{10}{5} p^5 q^5 = 2 \binom{10}{4} p^4 q^6$$

$$\Rightarrow \quad \frac{10\,!\,p}{5\,!\,5\,!} = 2\,\frac{10\,!\,q}{4\,!\,6\,!}$$

$$\Rightarrow \quad \frac{p}{5} = \frac{2q}{6} = \frac{q}{3}$$

$$\therefore 3p = 5q = 5(1-p) \quad \Rightarrow \quad 8p = 5 \quad \Rightarrow \quad p = 5/8 \text{ and } q = 3/8$$

$$\therefore \quad P(X = x) = \binom{10}{x} \left(\frac{5}{8}\right)^x \left(\frac{3}{8}\right)^{10-x}$$

Hence the required number of times that in 10,000 sets of 10 throws each, we get no even number

$$= 10{,}000 \times P(X = 0) = 10{,}000 \times \left(\frac{3}{8}\right)^{10} = 1 \text{ (approx.)}$$

Example 7·5 *In a precision bombing attack there is a 50% chance that any one bomb will strike the target. Two direct hits are required to destroy the target completely. How many bombs must be dropped to give a 99% chance or better of completely destroying the target ?* **[Gauhati Univ. M.A., 1992]**

Solution. We have :

p = Probability that the bomb strikes the target = 50% = $\frac{1}{2}$. Let n be the number of bombs which should be dropped to ensure 99% chance or better of completely destroying the target. This implies that "probability that out of n bombs, at least two strike the target, is greater than 0·99".

Let X be a r.v. representing the number of bombs striking the target. Then $X \sim B(n, p = \frac{1}{2})$ with

$$p(x) = P(X = x) = \binom{n}{x}\left(\frac{1}{2}\right)^x \cdot \left(\frac{1}{2}\right)^{n-x} = \binom{n}{x}\left(\frac{1}{2}\right)^n ; x = 0, 1, \ldots, n$$

We should have :

$$P(X \geq 2) \geq 0{\cdot}99$$

$$\Rightarrow \quad [1 - P(X \leq 1)] \geq 0{\cdot}99$$

$$\Rightarrow \quad [1 - \{p(0) + p(1)\}] \geq 0{\cdot}99$$

$$\Rightarrow \quad 1 - \left\{\binom{n}{0} + \binom{n}{1}\right\}\left(\frac{1}{2}\right)^n \geq 0{\cdot}99$$

$$\Rightarrow \quad 0{\cdot}01 \geq \frac{1+n}{2^n} \quad \Rightarrow \quad 2^n \times (0{\cdot}01) \geq 1 + n$$

$$\Rightarrow \quad 2^n \geq 100 + 100\,n \qquad \ldots(*)$$

By trial method, we find that the inequality (*) is satisfied by $n = 11$. Hence the minimum number of bombs needed to destroy the target completely is 11.

Example 7·6. *A department in a works has 10 machines which may need adjustment from time to time during the day. Three of these machines are old, each having a probability of 1/11 of needing adjustment during the day, and 7 are new, having corresponding probabilities of 1/21.*

Assuming that no machine needs adjustment twice on the same day, determine the probabilits that on a particular day

(i) just 2 old and no new machines need adjustment.

(ii) If just 2 machines need adjustment, they are of the same type.

(Nagpur Univ. B.E., 1989)

Solution. Let p_1 = Probability that an old machine needs adjustment

$= 1/11$

$\therefore \quad q_1 = 1 - p_1 = 10/11$

and p_2 = Probability that a new machine needs adjustment = 1/21

$q_2 = 1 - p_2 = 20/21$

Then $P_1(r)$ = Probability that 'r' old machines need adjustment

$= {}^3C_r\, p_1^r\, q_1^{3-r} = {}^3C_r\,(10/11)^{3-r}\,(1/11)^r$

and $P_2(r)$ = Probability that 'r' new machine need adjustment

$= {}^7C_r\, p_2^r\, q_2^{7-r} = {}^7C_r\,(1/21)^r\,(20/21)^{7-r}$

(i) The probability that just two old machines and no new machine need adjustment is given (by the compound probability theorem) by the expression :

$$P_1(2) \cdot P_2(0) = {}^3C_2(1/11)^2 \cdot (10/11) \cdot (20/21)^7 = 0{\cdot}016$$

(ii) Similarly the probability that just 2 new machines and no old machine need adjustment is

$$P_1(0) \cdot P_2(2) = (10/11)^3 \cdot {}^7C_2\,(1/21)^2 \cdot (20/21)^5 = 0{\cdot}028$$

The probability that "If just two machines need adjustment, they are of the same type" is the same as the probability that "either just 2 old and no new or just 2 new and no old machines need adjustment".

$\therefore$ Required probability = 0·016 + 0·028 = 0·044

7·2·1 Moments. The first four moments about origin of binomial distribution are obtained as follows :

$$\mu_1' = E(X) = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} = np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} q^{n-x} *$$

$$= np(q + p)^{n-1} = np \qquad (\because q + p = 1)$$

Thus the mean of the binomial distribution is np.

---

* $\binom{n}{x} = \frac{n}{x} \cdot \binom{n-1}{x-1} = \frac{n}{x} \cdot \frac{n-1}{x-1} \cdot \binom{n-2}{x-2}$

$= \frac{n}{x} \cdot \frac{n-1}{x-1} \cdot \frac{n-2}{x-2} \binom{n-3}{x-3}$, and so on.

$$\mu_2' = E(X^2) = \sum_{x=0}^{n} x^2 \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} \{x(x-1) + x\} \frac{n(n-1)}{x(x-1)} \cdot \binom{n-2}{x-2} p^x q^{n-x}$$

$$= n(n-1)p^2 \left[\sum_{x=2}^{n} \binom{n-2}{x-2} p^{x-2} q^{n-x} \right] + np$$

$$= n(n-1)p^2 (q+p)^{n-2} + np = n(n-1)p^2 + np$$

$$\mu_3' = E(X^3) = \sum_{x=0}^{n} x^3 \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} \{x(x-1)(x-2) + 3x(x-1) + x\} p^x q^{n-x}$$

$$= n(n-1)(n-2)p^3 \sum_{x=3}^{n} \binom{n-3}{x-3} p^{x-3} q^{n-x}$$

$$+ 3n(n-1)p^2 \sum_{x=2}^{n} \binom{n-2}{x-2} p^{x-2} q^{n-x} + np$$

$$= n(n-1)(n-2)p^3 (q+p)^{n-3} + 3n(n-1)p^2 (q+p)^{n-2} + np$$

$$= n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np$$

Similarly

$$x^4 = x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x$$

Let $x^4 = Ax(x-1)(x-2)(x-3) + Bx(x-1)(x-2) + Cx(x-1) + x$

By giving to x the values 1, 2 and 3 respectively, we find the values of arbitrary constants A, B and C. Therefore,

$$\mu_4' = E(X^4) = \sum_{x=0}^{n} x^4 \binom{n}{x} p^x p^{n-x}$$

$$= n(n-1)(n-2)(n-3)p^4 + 6n(n-1)(n-2)p^3 + 7n(n-1)p^2 + np$$

[On simplification]

Central Moments of Binomial Distribution :

$$\mu_2 = \mu_2' - \mu_1'^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1-p) = npq$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= \{n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np\} - 3\{n(n-1)p^2 + np\} np + 2(np)^3$$

$$= np[-3np^2 + 3np + 2p^2 - 3p + 1 - 3npq]$$

$$= np[3np(1-p) + 2p^2 - 3p + 1 - 3npq]$$

$$= np[2p^2 - 3p + 1] = np(2p^2 - 2p + q) = npq(1 - 2p)$$
$$= npq[q + p - 2p] = npq(q - p)$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = npq[1 + 3(n-2)pq]$$

[On simplification]

Hence

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{n^2p^2q^2(q-p)^2}{n^3p^3q^3} = \frac{(q-p)^2}{npq} = \frac{(1-2p)^2}{npq} \quad ...(7\cdot4)$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{npq\{1 + 3(n-2)pq\}}{n^2p^2q^2} = \frac{1 + 3(n-2)pq}{npq} = 3 + \frac{1-6pq}{npq} \quad ...(7\cdot5)$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}} = \frac{1-2p}{\sqrt{npq}}, \quad \gamma_2 = \beta_2 - 3 = \frac{1-6pq}{npq} \quad ...(7\cdot5\,a)$$

Example 7·7 *Comment on the following :*

The mean of a binomial distribution is 3 and variance is 4'.

Solution. If the given binomial distribution has parameters n and p, then we are given

$$\text{Mean} = np = 3 \quad ...(*)$$

and $$\text{Variance} = npq = 4 \quad ...(**)$$

Dividing (**) by (*), we get $q = 4/3$,

which is impossible, since probability cannot exceed unity. Hence the given statement is wrong.

Example 7·8. *The mean and variance of binomial distribution are 4 and $\frac{4}{3}$ respectively. Find $P(X \geq 1)$.* **(Sardar Patel Univ. B.Sc. 1993)**

Solution. Let $X \sim B(n, p)$. Then we are given

$$\text{Mean} = E(X) = np = 4$$

and $$\text{Var}(X) = npq = \tfrac{4}{3} \quad ...(*)$$

Dividing, we get

$$q = \tfrac{1}{3} \qquad \Rightarrow \qquad p = \tfrac{2}{3}$$

Substituting in (*), we get

$$n = \frac{4}{p} = \frac{4 \times 3}{2} = 6.$$

$$P(X \geq 1) = 1 - P(X = 0) = 1 - q^n = 1 - (1/3)^6 = 1 - (1/729)$$
$$= 1 - 0{\cdot}00137 = 0{\cdot}99863$$

Example 7·9 *If $X \sim B(n, p)$, show that :*

$$E\left(\frac{X}{n} - p\right)^2 = \frac{pq}{n}; \quad Cov\left(\frac{X}{n}, \frac{n-X}{n}\right) = -\frac{pq}{n}$$

(Delhi Univ. B.Sc., 1989)

Solution. Since $X \sim B(n, p)$, $E(X) = np$ and $\text{Var}(X) = npq$

$$\therefore \quad E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = p\,; \ \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \cdot \text{Var}(X) = \frac{pq}{n}$$

(i) $$E\left(\frac{X}{n} - p\right)^2 = E\left[\frac{X}{n} - E\left(\frac{X}{n}\right)\right]^2 = \text{Var}\left(\frac{X}{n}\right) = \frac{pq}{n}$$

(ii) $$\text{Cov}\left(\frac{X}{n}, \frac{n-X}{n}\right) = E\left[\left\{\frac{X}{n} - E\left(\frac{X}{n}\right)\right\}\left\{\frac{n-X}{n} - E\left(\frac{n-X}{n}\right)\right\}\right]$$

$$= E\left[\left(\frac{X}{n} - p\right)\left\{\left(1 - \frac{X}{n}\right) - (1-p)\right\}\right]$$

$$= E\left[\left(\frac{X}{n} - p\right)\left\{-\left(\frac{X}{n} - p\right)\right\}\right]$$

$$= -E\left(\frac{X}{n} - p\right)^2 = -\text{Var}\left(\frac{X}{n}\right) = -\frac{pq}{n}$$

7·2·2 Recurrence Relation for the moments of Binomial Distribution. *(Renovsky Formula)*

By def.,

$$\mu_r = E\{X - E(X)\}^r = \sum_{x=0}^{n} (x - np)^r \binom{n}{x} p^x q^{n-x}$$

Differentiating with respect to p, we get

$$\frac{d\mu_r}{dp} = \sum_{x=0}^{n} \binom{n}{x} \left[-nr(x-np)^{r-1} p^x q^{n-x} + (x-np)^r \{x p^{x-1} q^{n-x} - (n-x) p^x q^{n-x-1}\}\right]$$

$$= -nr \sum_{x=0}^{n} \binom{n}{x} (x-np)^{r-1} p^x q^{n-x} + \sum_{x=0}^{n} \binom{n}{x} (x-np)^r p^x q^{n-x} \left\{\frac{x}{p} - \frac{n-x}{q}\right\}$$

$$= -nr \sum_{x=0}^{n} (x-np)^{r-1} p(x) + \sum_{x=0}^{n} (x-np)^r p(x) \frac{(x-np)}{pq}$$

$$= -nr \sum_{x=0}^{n} (x-np)^{r-1} p(x) + \frac{1}{pq} \sum_{x=0}^{n} (x-np)^{r+1} p(x)$$

$$\therefore \quad \frac{d\mu_r}{dp} = -nr\,\mu_{r-1} + \frac{1}{pq}\,\mu_{r+1}$$

$$\Rightarrow \quad \mu_{r+1} = pq\left[nr\,\mu_{r-1} + \frac{d\mu_r}{dp}\right] \qquad \ldots(7\cdot 6)$$

Putting $r = 1, 2$ and 3 successively in (7·6), we get

$$\mu_2 = pq\left[n\mu_o + \frac{d\mu_1}{dp}\right] = npq \qquad (\because \mu_0 = 1 \text{ and } \mu_1 = 0)$$

$$\mu_3 = pq\left[2n\mu_1 + \frac{d\mu_2}{dp}\right] = pq \cdot \frac{d(npq)}{dp} = npq\frac{d}{dp}\{p(1-p)\}$$

$$= npq\frac{d}{dp}(p - p^2) = npq(1-2p) = npq(q-p)$$

and $$\mu_4 = pq\left[3n\mu_2 + \frac{d\mu_3}{dp}\right] = pq\left[3n \cdot npq + \frac{d}{dp}\{npq(q-p)\}\right]$$

$$= pq\left[3n^2pq + n\frac{d}{dp}\{p(1-p)(1-2p)\}\right]$$

$$= pq\left[3n^2pq + n\frac{d}{dp}(p - 3p^2 + 2p^3)\right]$$

$$= pq[3n^2pq + n(1 - 6p + 6p^2)] = pq[3n^2pq + n(1-6pq)]$$

$$= npq[3npq + 1 - 6pq] = npq[1 + 3pq(n-2)]$$

Example 7·10 *Show that the rth moment μ_r' about the origin of the binomial distribution of degree n is given by:*

$$\mu_r' = \left(p\frac{\partial}{\partial p}\right)^r (q+p)^n \qquad ...(*)$$

[Patna Univ. B.Sc. (Hons.), 1993]

Solution. We shall prove this result by using the principle of mathematical induction. We have

$$(q+p)^n = \sum_{x=0}^{n}\binom{n}{x} p^x q^{n-x} \Rightarrow \frac{\partial}{\partial p}(q+p)^n = \sum_{x=0}^{n}\binom{n}{x} q^{n-x}xp^{x-1}$$

$$\therefore \frac{\partial}{\partial p}(q+p)^n = p\sum_{x=0}^{n}\binom{n}{x} q^{n-x}xp^{x-1} = \sum_{x=0}^{n}\binom{n}{x} p^x q^{n-x}x = \mu_1'$$

Thus the result (*) is true for $r = 1$.

Let us now assume that the result (*) is true for $r = k$, so that

$$\left(p\frac{\partial}{\partial p}\right)^k (q+p)^n = \mu_k' = \sum_{x=0}^{n}\binom{n}{x} p^x q^{n-x}x^k \qquad ...(**)$$

Differentiate (**) partially w.r. to p and multiply both sides by p to get:

$$p\left(\frac{\partial}{\partial p}\right)\left[\left(p\frac{\partial}{\partial p}\right)^k (q+p)^n\right] = \sum_{x=0}^{n}\binom{n}{x} p^x q^{n-x}x^{k+1} = E\left(X^{k+1}\right)$$

$$\Rightarrow \left(p\frac{\partial}{\partial p}\right)^{k+1}(q+p)^n = \mu_{k+1}'$$

Hence if the result (*) is true for $r = k$, it is also true for $r = k + 1$. It is already shown to be true for $k = 1$. Hence by the principle of mathematical induction, (*) is true for all positive integral values of r.

7·2·3. Factorial Moments of Binomial Distribution. The rth factorial moment of the Binomial distribution is:

$$\mu_{(r)}' = E\left[X^{(r)}\right] = \sum_{x=0}^{n} x^{(r)} p(x) = \sum_{x=0}^{n} x^{(r)} \frac{n!}{x!\,(n-x)!} p^x q^{n-x}$$

$$= n^{(r)} p^r \sum_{x=r}^{n} \frac{(n-r)!}{(x-r)!\,(n-x)!} p^{x-r} q^{n-x} = n^{(r)} p^r (q+p)^{n-r}$$

$$= n^{(r)} p^r \qquad \ldots(7\cdot 7)$$

$$\mu_{(1)}' = E\left[X^{(1)}\right] = np = \text{Mean}$$

$$\mu_{(2)}' = E\left[X^{(2)}\right] = n^{(2)} p^2 = n(n-1)p^2$$

$$\mu_{(3)}' = E\left[X^{(3)}\right] = n^{(3)} p^3 = n(n-1)(n-2)p^3$$

Now $\mu_{(2)} = \mu_{(2)}' - \mu_{(1)}'^2 + \mu_{(1)}' = n^2p^2 - np^2 - n^2p^2 + np = npq$

$$\mu_{(3)} = \mu_{(3)}' - 3\mu_{(2)}'\mu_{(1)}' + 2\mu_{(1)}'^3 - 2\mu_{(1)}'$$

$$= n(n-1)(n-2)p^3 - 3n(n-1)p^2 np + 2n^3p^3 - 2np = -2npq(1+p)$$

[On simplification]

7·2·4. Mean Deviation About Mean of Binomial Distribution. The mean deviation η about the mean np of the binomial distribution is given by

$$\eta = \sum_{x=0}^{n} |x - np|\, p(x) = \sum_{x=0}^{n} |x - np| \binom{n}{x} p^x q^{n-x},$$

(x being an integer)

$$= \sum_{x=0}^{np} -(x - np)\binom{n}{x} p^x q^{n-x} + \sum_{x=np}^{n} (x - np)\binom{n}{x} p^x q^{n-x}$$

$$= 2\sum_{x=np}^{n} (x - np)\binom{n}{x} p^x q^{n-x} *$$

$$= 2\sum_{\mu}^{n} (x - np)\binom{n}{x} p^x q^{n-x},$$

where μ is the greatest integer contained in $np + 1$.

$$= 2\sum_{\mu}^{n} \left[\{xq - (n-x)p\}\binom{n}{x} p^x q^{n-x}\right]$$

$$= 2\sum_{\mu}^{n} \left[\frac{n!}{(x-1)!\,(n-x!} p^x q^{n-x+1} - \frac{n!}{x!\,(n-x-1)!} p^{x+1} q^{n-x}\right]$$

---

* $\because \quad \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} = np$

$\Rightarrow \quad \sum_{x=0}^{n} (x - np)\binom{n}{x} p^x q^{n-x} = 0$

$$= 2 \sum_{x=\mu}^{n} \left[t_{x-1} - t_x \right], \text{ where } t_x = \frac{n!}{x!\,(n-x-1)!} p^{x+1} q^{n-x}$$

$$= 2 \left[t_{\mu-1} - t_n \right] = 2\, t_{\mu-1}$$

This is obtained by summing over x and using $t_n = o$

$$\therefore \eta = 2t_{\mu-1} = 2 \frac{n!}{(\mu-1)!\,(n-\mu)!} \cdot p^{\mu} q^{n-\mu+1}$$

$$= 2npq \binom{n-1}{\mu-1} p^{\mu-1} q^{n-\mu} \qquad \ldots(7\cdot8)$$

7·2·5. Mode of the Binomial Distribution. We have

$$\frac{p(x)}{p(x-1)} = \binom{n}{x} p^x q^{n-x} \Big/ \binom{n}{x-1} p^{x-1} q^{n-x+1}$$

$$= \frac{n!}{(n-x)!\,x!} p^x q^{n-x} \Big/ \frac{n!}{(x-1)!\,(n-x+1)!} p^{x-1} q^{n-x+1}$$

$$= \frac{(n-x+1)p}{xq} = \frac{xq + (n-x+1)p - xq}{xq}$$

$$= 1 + \frac{(n+1)p - x(p+q)}{xq} = 1 + \frac{(n+1)p - x}{xq} \qquad \ldots(7\cdot9)$$

Mode is the value of x for which $p(x)$ is maximum.

We discuss the following two cases :

Case 1. When $(n+1)p$ is not an integer

Let $(n+1)p = m + f$, where m is an integer and f is fractional such that $0 < f < 1$. Substituting in (7·9), we get

$$\frac{p(x)}{p(x-1)} = 1 + \frac{(m+f) - x}{xq} \qquad \ldots(*)$$

From (*), it is obvious that

$$\frac{p(x)}{p(x-1)} > 1 \text{ for } x = 0, 1, 2, \ldots, m$$

and $$\frac{p(x)}{p(x-1)} < 1 \text{ for } x = m+1, m+2, \ldots, n$$

$$\Rightarrow \quad \frac{p(1)}{p(0)} > 1, \frac{p(2)}{p(1)} > 1, \ldots, \frac{p(m)}{p(m-1)} > 1,$$

and $$\frac{p(m+1)}{p(m)} < 1, \frac{p(m+2)}{p(m+1)} < 1, \ldots, \frac{p(n)}{p(n-1)} < 1,$$

$$\therefore p(0) < p(1) < p(2) < \ldots < p(m-1) < p(m) > p(m+1) > p(m+2) > p(m+3) \ldots > p(n),$$

Thus in this case there exists unique modal value for binomial distribution and it is m, the integral part of $(n+1)p$.

Case II. *When $(n + 1)p$ is an integer.*

Let $(n + 1)p = m$ *(an integer).*

Substituting in (7·9), we get

$$\frac{p(x)}{p(x-1)} = 1 + \frac{m - x}{xq} \qquad \ldots(**)$$

From (**) it is obvious that

$$\left.\begin{aligned} \frac{p(x)}{p(x-1)} \end{aligned}\right\} \begin{aligned} &> 1 \text{ for } x = 1, 2, \ldots, m-1 \\ &= 1 \text{ for } x = m \\ &< 1 \text{ for } x = m+1, m+2, \ldots, n \end{aligned}$$

Now proceeding as in case 1, we have :

$$p(0) < p(1) < \ldots < p(m-1) = p(m) > p(m+1) > p(m+2) > \ldots > p(n)$$

Thus in this case the distribution is bimodal and the two modal values are m and $m - 1$.

Example 7·11. *Determine the binomial distribution for which the mean is 4 and variance 3 and find its mode.* **(Madurai Kamraj Univ B.Sc. 1993)**

Solution, Let $X \sim B(n, p)$, then we are given that

$$E(X) = np = 4 \qquad ..(*)$$

and $$\text{Var}(X) = npq = 3 \qquad ..(**)$$

Dividing (**) by (*), we get

$$q = \tfrac{3}{4} \quad \Rightarrow \quad p = 1 - q = \tfrac{1}{4}$$

Hence from (*), $$n = \frac{4}{p} = 16$$

Thus the given binomial distribution has parameters $n = 16$ and $p = 1/4$.

Mode. We have $(n + 1)p = 4{\cdot}25$, which is not an integer. Hence the unique mode of the binomial distribution is 4, the integral part of $(n + 1)p$.

Example 7·12. *Show that for $p = 0{\cdot}50$, the binomial distribution has a maximum probability at $X = \frac{1}{2}n$, if n is even, and at $X = \frac{1}{2}(n - 1)$ as well as $X = \frac{1}{2}(n + 1)$, if n is odd.* **(Mysore Univ., B. Sc. 1991)**

Solution. Here we have to find the mode of the binomial distribution.

(i) Let n be even $= 2m$, (say), $m = 1, 2, \ldots$

$\therefore$ If $p = 0{\cdot}5$, then $(n + 1)p = (2m + 1) \times \left(\frac{1}{2}\right) = m + 0{\cdot}5$

Hence in this case, the distribution is unimodal, the unique mode being at $X = m = n/2$.

(ii) Let n be odd $= (2m + 1)$, say. Then

$$(n + 1)p = (2m + 2) \times \tfrac{1}{2} = m + 1 \text{ (Integer)}$$

$$= \frac{n-1}{2} + 1 = \frac{n+1}{2}$$

Since $(n + 1)p$ is an integer, the distribution is bimodal, the two modes being $\frac{1}{2}(n + 1)$ and $\frac{1}{2}(n+1)-1 = \frac{1}{2}(n-1)$.

7·2·6. Moment Generating Function of Binomial Distribution. Let X be a variable following binomial distribution, then

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} (pe^t)^x q^{n-x} \binom{n}{x} = (q + pe^t)^n \quad ...(7·10)$$

M.G.F. about Mean of Binomial Distribution :

$$E\left\{e^{t(X-np)}\right\} = E(e^{tx} e^{-tnp}) = e^{-tnp} \cdot E(e^{tX}) = e^{-tnp} \cdot M_X(t)$$

$$= e^{-tnp} \cdot (q + pe^t)^n = (qe^{-pt} + pe^{tq})^n \quad ...(7·11)$$

$$= \left[q\left\{1 - pt + \frac{p^2t^2}{2!} - \frac{p^3t^3}{3!} + \frac{p^4t^4}{4!} - \ldots\right\} + p\left\{1 + tq + \frac{t^2q^2}{2!} + \frac{t^3q^3}{3!} - \ldots\right\}\right]^n$$

$$= \left[1 + \frac{t^2}{2!}pq + \frac{t^3}{3!}pq(q^2 - p^2) + \frac{t^4}{4!}pq(q^3 + p^3) + \ldots\right]^n$$

$$= \left[1 + \left\{\frac{t^2}{2!} \cdot pq + \frac{t^3}{3!} \cdot pq(q - p) + \frac{t^4}{4!}pq(1 - 3pq) + \ldots\right\}\right]^n$$

$$= \left[1 + \binom{n}{1}\left\{\frac{t^2}{2!} \cdot pq + \frac{t^3}{3!}pq(q - p) + \frac{t^4}{4!}pq(1 - 3pq) + \ldots\right\} + \binom{n}{2}\left\{\frac{t^2}{2!}pq + \frac{t^3}{3!}pq(q - p) + \ldots\right\}^2 + \ldots\right.$$

Now μ_2 = Coefficient of $\frac{t^2}{2!} = npq$

μ_3 = Coefficient of $\frac{t^3}{3!} = npq(q - p)$

μ_4 = Coefficient of $\frac{t^4}{4!} = npq(1 - 3pq) + 3n(n - 1)p^2q^2$

$= npq(1 - 3pq) + 3n^2p^2q^2 - 3np^2q^2$

$= 3n^2p^2q^2 + npq(1 - 6pq)$

Example 7·13 *X is binomially distributed with parameters n and p. What is the distrbution of $Y = n - X$?* **[Delhi Univ. B.Sc. (Maths Hons.), 1990]**

Solution. $X \sim B(n, p)$, represents the number of successes in n independent trials with constant probability p of success for each trial.

$\therefore$ $Y = n - X$, represents the number of failures in n independent trial with constant probability 'q' of failure for each trial. Hence $Y = n - X \sim B(n, q)$

Aliter Since $X \sim B(n, p)$, $M_X(t) = E(e^{tX}) = (q + pe^t)^n$

$$\therefore \quad M_Y(t) = E(e^{tY}) = E(e^{t(n-X)})$$
$$= e^{nt} \cdot E(e^{-tX}) = e^{nt} M_X(-t)$$
$$= e^{nt} \cdot (q + pe^{-t})^n$$
$$= \left[e^t \left(q + pe^{-t}\right)\right]^n = \left(p + qe^t\right)^n$$

Hence by uniqueness theorem of m.g.f., $Y = n - X \sim B(n, q)$

Example 7·14. *The m.g.f. of a r.v. X is* $\left(\frac{2}{3} + \frac{1}{3}e^t\right)^9$. *Show that :*

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = \sum_{x=1}^{5} \binom{9}{x} \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{9-x}$$

[Delhi Univ. B.Sc. (Maths Hons.), 1989]

Solution. Since $M_X(t) = \left(\frac{2}{3} + \frac{1}{3}e^t\right)^9 = \left(q + pe^t\right)^n$,

by uniqueness theorem of m.g.f. $X \sim B\left(n = 9, p = \frac{1}{3}\right)$

Hence $E(X) = \mu_x = np = 3$; $\sigma_X^2 = npq = 9 \times \frac{1}{3} \times \frac{2}{3} = 2$

$\mu \pm 2\sigma = 3 \pm 2 \times \sqrt{2} = 3 \pm 2 \times 1{\cdot}4 = (0{\cdot}2, 5{\cdot}8)$

$\therefore$ $P(\mu - 2\sigma < X < \mu + 2\sigma) = P(0{\cdot}2 < X < 5{\cdot}8) = P(1 \le X \le 5)$

$$= \sum_{x=1}^{5} p(x) = \sum_{x=1}^{5} {}^nC_x p^x q^{n-x}$$
$$= \sum_{x=1}^{5} {}^9C_x (1/3)^x (2/3)^{9-x}$$

7·2·7. Additive Property of Binomial Distribution. Let $X \sim B(n_1, p_1)$ and $Y \sim B(n_2, p_2)$ be independent random variables. Then

$$M_X(t) = (q_1 + p_1 e^t)^{n_1}, \; M_Y(t) = (q_2 + p_2 e^t)^{n_2} \qquad \ldots(*)$$

What is the distribution of $X + Y$?

We have

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) \; [\because X \text{ and } Y \text{ are independent}]$$
$$= (q_1 + p_1 e^t)^{n_1} \cdot (q_2 + p_2 e^t)^{n_2} \qquad \ldots(**)$$

Since (**) cannot be expressed in the form $(q + p e^t)^n$, from uniqueness theorem of m.g.f.'s it follows that $X + Y$ is not a binomial variate. Hence, *in general the sum of two independent binomial variates is not a binomial variate.*

In other words, binomial distribution does not possess the additive or reproductive property.

However, if we take $p_1 = p_2 = p$ (say), then from (**), we get

$$M_{X+Y}(t) = (q + p e^t)^{n_1 + n_2},$$

which is the m.g.f. of a binomial variate with parameters $(n_1 + n_2, p)$. Hence, by uniqueness theorem of m.g.f.'s $X + Y \sim B(n_1 + n_2, p)$. Thus the *binomial distribution possesses the additive or reproductive property if* $p_1 = p_2$.

Generalisation. *If* X_i, $(i = 1, 2, ..., k)$ *are independent binomial variates with parameters* (n_i, p), $(i = 1, 2, ..., k)$ *then their sum* $\sum_{i=1}^{k} X_i \sim B\left(\sum_{i=1}^{k} n_i,\ p\right)$.

The proof is left as an exercise to the reader.

Example 7·15. *If the independent random variables X, Y are binomially distributed, respectively with* $n = 3$, $p = 1/3$, *and* $n = 5$, $p = 1/3$, *write down the probability that* $X + Y \geq 1$.

Solution. We are given

$$X \sim B(3, \tfrac{1}{3}) \text{ and } Y \sim B(5, \tfrac{1}{3}).$$

Since X and Y are independent binomial random variables, with $p_1 = p_2 = \frac{1}{3}$, by the additive property of binomial distribution, we get

$$X + Y \sim B(3 + 5, \tfrac{1}{3}), \text{ i.e., } X + Y \sim B(8, \tfrac{1}{3})$$

$$\therefore \quad P(X + Y = r) = {}^8C_r (\tfrac{1}{3})^r (\tfrac{2}{3})^{8-r} \qquad ...(*)$$

Hence $P(X + Y \geq 1) = 1 - P(X + Y < 1)$

$$= 1 - P(X + Y = 0)$$

$$= 1 - (\tfrac{2}{3})^8$$

7·2·8. Characteristic Function of Binomial Distribution.

$$\varphi_X(t) = E(e^{itX}) = \sum_{x=0}^{n} e^{itx} p(x) = \sum_{x=0}^{n} e^{itx} \binom{n}{x} p^x q^{n-x}$$

$$= \sum_{x=0}^{n} e^{itx} \binom{n}{x} (pe^{it})^x q^{n-x} = (q + pe^{it})^n \qquad ...(7·12)$$

7·2·9. Cummulants of the Binomial Distribution. Cumulant generating function is

$$K_X(t) = \log M_X(t) = \log (q + pe^t)^n = n \log (q + pe^t)$$

$$= n \log \left[q + p \left(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \ldots \right) \right]$$

$$= n \log \left[1 + p \left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \ldots \right) \right]$$

$$= n\left[p\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \ldots\right) - \frac{p^2}{2}\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right)^2 + \frac{p^3}{3}\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right)^3 - \frac{p^4}{4}\left(t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right)^4 + \ldots\right]$$

Mean $= \kappa_1 =$ Coefficient of t in $K_X(t) = np$

$\mu_2 = \kappa_2 =$ Coefficient of $\frac{t^2}{2!}$ in $K_X(t) = n(p - p^2) = np(1 - p) = npq$

The coefficient of t^3 in $K_X(t)$

$$= n\left[\frac{p}{3!} - \frac{p^2}{2!} \cdot 2 \cdot \frac{1}{2!} + \frac{p^3}{3}\right] = \frac{np}{3!}(1 - 3p + 2p^2)$$

$\therefore \kappa_3 =$ Coefficient of $\frac{t^3}{3!}$ in $K_X(t) = np(1 - 3p + 2p^2)$

$$= np(1 - p)(1 - 2p) = npq(1 - p - p) = npq(q - p)$$

$\therefore \quad \mu_3 = \kappa_3 = npq(q - p)$

The Coefficient of t^4 in $K_X(t)$

$$= n\left[\frac{p}{4!} - \frac{p^2}{2!}\left(\frac{2}{3!} + \frac{1}{4}\right) + \frac{p^3}{3} \cdot \frac{3}{2!} - \frac{p^4}{4}\right]$$

$$= \frac{np}{4!}[1 - 7p + 12p^2 - 6p^3]$$

$\therefore \kappa_4 =$ Co efficient of $\frac{t^4}{4!}$ in $K_X(t) = np(1 - p)(1 - 6p + 6p^2)$

$$= npq[1 - 6p(1 - p)] = npq(1 - 6pq)$$

$\therefore \mu_4 = \kappa_4 + 3\kappa_2^2 = npq(1 - 6pq) + 3n^2p^2q^2$

$$= npq(1 - 6pq + 3npq) = npq[1 + 3pq(n - 2)]$$

7·2·10. Recurrence Relation for Cumulants of Binomial Distribution. By def.,

$$\kappa_r = \left[\frac{d^r}{dt^r} \log M_X(t)\right]_{t=0} = n\left[\frac{d^r}{dt^r} \log (q + pe^t)\right]$$

$$\frac{d\kappa_r}{dp} = n\left[\frac{d^r}{dt^r} \cdot \frac{d}{dp} \log (q + pe^t)\right]_{t=0} = n\left[\frac{d^r}{dt^r} \cdot \frac{(-1 + e^t)}{q + pe^t}\right]_{t=0}$$

$$\kappa_{r+1} = n\left[\frac{d^{r+1}}{dt^{r+1}} \log (q + pe^t)\right]_{t=0}$$

$$= n\left[\frac{d^r}{dt^r} \cdot \frac{d}{dt} \log (q + pe^t)\right]_{t=0} = n\left[\frac{d^r}{dt^r}\left(\frac{pe^t}{q + pe^t}\right)\right]_{t=0}$$

$$= n \left[\frac{d^r}{dt^r} \left(1 - \frac{q}{q + pe^t} \right) \right]_{t=0} = - nq \left[\frac{d^r}{dt^r} \left(\frac{1}{q + pe^t} \right) \right]_{t=0}$$

Hence

$$\kappa_{r+1} - pq \frac{d\kappa_r}{dp} = - nq \left[\frac{d^r}{dt^r} \left(\frac{1}{q + pe^t} \right) \right]_{t=0} - npq \left[\frac{d^r}{dt^r} \left(\frac{e^t - 1}{q + pe^t} \right) \right]_{t=0}$$

$$= - nq \left[\frac{d^r}{dt^r} \left\{ \frac{1 + pe^t - p}{q + pe^t} \right\} \right]_{t=0}$$

$$= - nq \left[\frac{d^r}{dt^r} \left\{ \frac{q + pe^t}{q + pe^t} \right\} \right]_{t=0} = - nq \left[\frac{d^r}{dt^r} (1) \right]_{t=0} = 0$$

$$\therefore \qquad \kappa_{r+1} = pq \frac{d\kappa_r}{dp} \qquad \ldots(7\cdot13)$$

In particular,

$$\kappa_2 = pq \cdot \frac{d\kappa_1}{dp} = pq \cdot \frac{d}{dp} (np) = npq \qquad (\because \kappa_1 = \text{mean} = np)$$

$$\kappa_3 = pq \cdot \frac{d\kappa_2}{dp} = pq \cdot \frac{d(npq)}{dp} = npq(q - p)$$

$$\kappa_4 = pq \cdot \frac{d\kappa_3}{dp} = pq \cdot \frac{d}{dp} \{ npq (q - p) \}$$

$$= npq \frac{d}{dp} \{ p (1 - p) (1 - 2p) \}$$

$$= npq \cdot \frac{d}{dp} (p - 3p^2 + 2p^3) = npq (1 - 6p + 6p^2)$$

$$= npq [1 - 6p (1 - p)] = npq (1 - 6pq)$$

7·2·11. Probability Generating Function of Binomial Distribution

$$P(s) = \sum_{\kappa=0}^{n} P(X = k) s^{\kappa} = \sum_{\kappa=0}^{n} \binom{n}{k} (ps)^{\kappa} q^{n-k} = (ps + q)^n \qquad \ldots(7\cdot13\,a)$$

The fact that this generating function is nth power of $(q + ps)$ shows that $p(x) = \{ b(x; n, p) \}$ is the distribution of the sum $S_n = X_1 + X_2 + \ldots + X_n$ of n random variables with the common generating function $(q + ps)$. Each variable X_i assumes the value 0 with probability q and 1 with probability p.

Thus $$\{ b(k; n, p) \} = \{ b(k; 1, p) \}^{n *} \qquad \ldots(7\cdot13b)$$

Let X and Y be two independent random variables having $b(k; m, p)$ and $b(k; n, p)$ as their distributions, then

$$P_X(s) = (q + ps)^m \text{ and } P_Y(s) = (q + ps)^n$$

$$\therefore \qquad P_{X+Y}(s) = (q + ps)^m (q + ps)^n = (q + ps)^{m+n}$$

$$\therefore \quad \{ b(k; m, p) \} * \{ b(k; n, p) \} = \{ b(k; m + n, p) \} \qquad \ldots(7\cdot13c)$$

Also $\mu_{(1)}' = [\, n(q + ps)^{n-1} p \,]_{s=1} = np$

$\mu_{(2)}' = [\, n(n-1)(q + ps)^{n-2} p^2 \,]_{s=1} = n(n-1)p^2$ and so on.

$\mu_{(r)}' = [n(n-1)\ldots(n-r+1)(q + ps)^{n-r} p^r]_{s=1}$

$= n(n-1)\ldots(n-r+1)p^r$

Example 7·16 *Show that*

$$E\left(\frac{1}{X + a}\right) = \int_0^1 t^{a-1} G(t)\, dt, \quad a > 0 \qquad \ldots(*)$$

where $G(t)$ is the probability generating function of X.

Find it when $X \sim B(n, p)$, and $a = 1$

[Delhi Univ. (Stat Hons.) Spl Course, 1988]

Solution. R.H.S. $= \int_0^1 t^{a-1} \cdot G(t)\, dt = \int_0^1 t^{a-1} \left(Et^X\right) dt$

$$= \int_0^1 \left\{ t^{a-1} \left(\sum_x p_x t^x \right) \right\} dt = \sum_x \left[p_x \int_0^1 t^{x+a-1}\, dt \right]$$

$$= \sum_x p_x \cdot \frac{1}{(x + a)} = E\left(\frac{1}{X + a}\right)$$

If $X \sim B(n, p)$, then $G(t) = \sum_{x=0}^{n} t^x p_x = (q + pt)^n$ $\qquad \ldots(**)$

Hence taking $a = 1$ in (*) and using (**), we get :

$$E\left[\frac{1}{(X + a)}\right] = \int_0^1 (q + pt)^n\, dt = \left| \frac{(q + pt)^{n+1}}{(n+1)p} \right|_0^1 = \frac{1 - q^{n+1}}{(n+1)p}$$

7·2·12. Recurrence Relation for the Probabilities of Binomial Distribution. *(Fitting of Binomial Distribution).*

We have

$$\frac{p(x+1)}{p(x)} = \frac{\binom{n}{x+1} p^{x+1} q^{n-x-1}}{\binom{n}{x} p^x q^{n-x}}$$

$$= \frac{n - x}{x + 1} \cdot \frac{p}{q} \qquad \text{(On simplification)}$$

$$p(x+1) = \left\{ \frac{n - x}{x + 1} \cdot \frac{p}{q} \right\} p(x), \qquad \ldots(7 \cdot 14)$$

which is the required recurrence formula.

This formula provides us a very convenient method of graduating the given data by a binomial distribution. The only probability we need to calculate is $p(0)$

which is given by $p(0) = q^n$, where q is estimated from the given data by equating the mean $\bar{x}$ of the distribution to np, the mean of the binomial distribution. Thus $\hat{p} = \bar{x}/n$.

The remaining probabilities, *viz.*, $p(1), p(2), \ldots$ can now be easily obtained from (7·14) as explained below :

$$p(1) = [p(x+1)]_{x=0} = \left(\frac{n-x}{x+1} \cdot \frac{p}{q}\right)_{x=0} p(0)$$

$$p(2) = [p(x+1)]_{x=1} = \left(\frac{n-x}{x+1} \cdot \frac{p}{q}\right)_{x=1} p(1)$$

$$p(3) = [p(x+1)]_{x=2} = \left(\frac{n-x}{x+1} \cdot \frac{p}{q}\right)_{x=2} p(2)$$

and so on.

Example 7·17. *Seven coins are tossed and number of heads noted. The experiment is repeated 128 times and the following distribution is obtained:*

No. of heads	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*Total*
Frequencies	*7*	*6*	*19*	*35*	*30*	*23*	*7*	*1*	*128*

Fit a Binomial distribution assuming

(i) The coin is unbaised,

(ii) The nature of the coin is not known.

(iii) Probability of a head for four coins is 0·5 and for the remaining three coins is 0·45.

Solution. In fitting Binomial distribution, first of all the mean and variance of the data are equated to np and npq respectively. Then the expected frequencies are calculated from these values of n and p. Here $n = 7$ and $N = 128$.

Case I. When the coin is unbaised

$$p = q = \tfrac{1}{2}, \quad (p/q = 1)$$

Now $$p(0) = q^n = (\tfrac{1}{2})^7 = (1/128)$$

$$f(0) = Nq^n = 128\,(\tfrac{1}{2})^7 = 1$$

Using the recurrence formula, the various probabilities, *viz.*, $p(1)$, $p(2)$,... can be easily calculated as shown below.

x	$\frac{n-x}{x+1}$	$\frac{n-x}{x+1} \cdot \frac{p}{q}$	*Expected frequency* $f(x) = Np(x)$
0	7	7	$f(0) = Np(0) = 1$
1	3	3	$f(1) = 1 \times 7 = 7$

2	$\frac{5}{3}$	$\frac{5}{3}$	$f(2) = 7 \times 3 = 21$
3	1	1	$f(3) = 21 \times \frac{5}{3} = 35$
4	$\frac{3}{5}$	$\frac{3}{5}$	$f(4) = 35 \times 1 = 35$
5	$\frac{1}{3}$	$\frac{1}{3}$	$f(5) = 35 \times \frac{3}{5} = 21$
6	$\frac{1}{7}$	$\frac{1}{7}$	$f(6) = 21 \times \frac{1}{3} = 7$
7			$f(7) = 7 \times \frac{1}{7} = 1$

Case II. When the nature of the coin is not known, then

$$np = \frac{1}{N} \sum_{i=1}^{n} f_i x_i = \frac{433}{128} = 3{\cdot}3828\,; n = 7$$

$\therefore$ $p = 0{\cdot}48326$ and $q = 0{\cdot}51674$, $(p/q = 0{\cdot}93521)$

$f(0) = Nq^7 = 128\,(0{\cdot}5167)^7 = 1{\cdot}2593$ (using logarithms)

x	$\frac{n-x}{x+1}$	$\frac{n-x}{x+1}\cdot\frac{p}{q}$	Expected frequency $f(x) = Np(x)$
0	7	6·54647	$f(0) = Np(0) = 1{\cdot}2593 \simeq 1$
1	3	2·80563	$f(1) = 1{\cdot}2593 \times 6{\cdot}54647 = 8{\cdot}2438 \simeq 8$
2	$\frac{5}{3}$	1·55868	$f(2) = 2{\cdot}80563 \times 8{\cdot}2438 = 23{\cdot}129 \simeq 23$
3	1	0·93521	$f(3) = 1{\cdot}55868 \times 23{\cdot}129 = 36{\cdot}05 \simeq 36$
4	$\frac{3}{5}$	0·56113	$f(4) = 0{\cdot}93521 \times 36{\cdot}05 = 33{\cdot}715 \simeq 34$
5	$\frac{1}{3}$	0·31174	$f(5) = 0{\cdot}56113 \times 33{\cdot}715 = 18{\cdot}918 \simeq 19$
6	$\frac{1}{7}$	0·13360	$f(6) = 0{\cdot}31174 \times 18{\cdot}918 = 5{\cdot}897 \simeq 6$
7			$f(7) = 0{\cdot}13360 \times 5{\cdot}897 = 0{\cdot}788 \simeq 1$

The probability generating functions (p.g.f.), say $P_X(s)$ for the 4 coins and $P_Y(S)$ for the remaining 3 coins are given by,

$$P_X(s) = (0{\cdot}50 + 0{\cdot}50\,s)^4,\ P_Y(s) = (0{\cdot}55 + 0{\cdot}45s)^3 \quad \ldots[\text{c.f. } 7{\cdot}13\,(a)]$$

Since all the throws are independent, the p.g.f. $P_{X+Y}(s)$ for the whole experiment is given by

$$P_{X+Y}(s) = P_X(s)\,P_Y(s) \qquad \text{...[c.f. 7·13 (b)]}$$

$$= (0{\cdot}50 + 0{\cdot}50\,s)^4\,(0{\cdot}55 + 0{\cdot}45\,s)^3$$

$$= (0{\cdot}0625 + 0{\cdot}25\,s + 0{\cdot}375\,s^2 + 0{\cdot}25\,s^3 + 0{\cdot}0625\,s^4)$$

$$\times\,(0{\cdot}166375 + 0{\cdot}408375\,s + 0{\cdot}334125\,s^2 + 0{\cdot}091125\,s^3)$$

Now $f(x) = N \times$ coefficient *of* t^x in $P_{X+Y}(t)$

$\therefore$ $f(0) = 128 \times {\cdot}0625 \times {\cdot}16637 = 1{\cdot}13310$

$f(1) = 128\,\{{\cdot}25 + {\cdot}166375 + {\cdot}408375 \times {\cdot}0625\} = 8{\cdot}5910$

$f(2) = 128\,\{{\cdot}28396\} = 36{\cdot}3470$ $\quad f(5) = 128\,\{{\cdot}14602\} = 18{\cdot}6934$

$f(3) = 128\,\{{\cdot}184117\} = 23{\cdot}5669$ $\quad f(6) = 128\,\{{\cdot}04366\} = 5{\cdot}5889$

$f(4) = 128\,\{{\cdot}260570\} = 33{\cdot}3529$ $\quad f(7) = 128\,\{{\cdot}005695\} = {\cdot}72896$

Example 7·18. *Let X and Y be independent binomial variates, each with parameters n and p. Find P (X − Y = k).* **(Calcutta Univ. B.Sc., 1993)**

Solution. Since each of the variables X and Y takes the values $0,1,2,..,n$, $Z = X - Y$ takes on the values $-n, -(n-1), ..., -1, 0, 1 ..., n$

$$P(Z = k) = \sum_{r=0}^{n} P(X = k + r \cap Y = r)$$

$$= \sum_{r=0}^{n} P(X = k + r) \cdot P(Y = r) \quad (\because X \text{ and Y are independent}).$$

$$= \sum_{r=0}^{n} \binom{n}{k+r} p^{k+r} \cdot q^{n-k-r} \binom{n}{r} p^r q^{n-r}$$

$$= \sum_{r=0}^{n} \binom{n}{k+r} \binom{n}{r} p^{2r+k}\, q^{2n-2r-k} \qquad \text{...(*)}$$

where $k = -n, -(n-1), ..., -2, -1, 0, 1, 2, ..., n$; and $q = 1 - p$.

In particular, we have :

$$P(Z = 0) = \sum_{r=0}^{n} \binom{n}{r}^2 \cdot p^{2r}\, q^{2n-2r}$$

$$P(Z = -n) = \sum_{r=0}^{n} \binom{n}{-n+r} \binom{n}{r} p^{2r-n}\, q^{3n-2r} = p^n q^n,$$

because we get the result when $r = n$ and for other values of $r < n$, $\binom{n}{-n+r}$ is not defined and hence taken as 0.

Example 7·19. *Find the m.g.f. of standard binomial variate* $(X - np)/\sqrt{npq}$ *and obtain its limiting form as* $n \to \infty$. *Also interpret the result.*

[Delhi Univ. B.Sc. (Stat. Hons.) 1990, 85]

Solution. We know that if $X \sim B(n, p)$, then

$$M_X(t) = (q + p\,e^t)^n$$

The m.g.f. of standard binomial variate.

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{X - \mu}{\sigma}, \text{ (say)}$$

where $\mu = np$ and $\sigma^2 = npq$, is given by

$$M_Z(t) = e^{-\mu t/\sigma} M_X(t/\sigma)$$

$$= e^{-npt/\sqrt{npq}} \cdot (q + p e^{t/\sqrt{npq}})^n \qquad \text{[From (**)]}$$

$$= \left[e^{-pt/\sqrt{npq}} (q + p e^{t/\sqrt{npq}}) \right]^n$$

$$= \left[q e^{-pt/\sqrt{npq}} + p e^{qt/\sqrt{npq}} \right]^n$$

$$= \left[q \left\{ 1 - \frac{pt}{\sqrt{npq}} + \frac{p^2 t^2}{2npq} + 0'(n^{-3/2}) \right\} \right.$$

$$\left. + p \left\{ 1 + \frac{qt}{\sqrt{npq}} + \frac{q^2 t^2}{2npq} + 0''(n^{-3/2} \right\} \right]^n$$

where $0'(n^{-3/2})$ and $0''(n^{-3/2})$ involve terms containing $n^{3/2}$ and higher powers of n in the denominator.

$$\therefore M_Z(t) = \left[(q + p) + \frac{t^2 pq}{2npq} (p + q) + 0(n^{-3/2}) \right]^n$$

$$= \left[1 + \frac{t^2}{2n} + 0(n^{-3/2}) \right]^n$$

where $0(n^{-3/2})$ involves terms with $n^{3/2}$ and higher powers of n in the denominator.

$$\therefore \log M_Z(t) = n \log \left[1 + \frac{t^2}{2n} + 0(n^{-3/2}) \right]$$

$$= n \left[\left\{ \frac{t^2}{2n} + 0(n^{-3/2}) \right\} - \frac{1}{2} \left\{ \frac{t^2}{2n} + 0(n^{-3/2}) \right\}^2 + \ldots\ldots \right]$$

$$= \frac{t^2}{2} + 0'''(n^{-1/2})$$

where $0'''(n^{-1/2})$ involve terms with $n^{1/2}$ and higher powers of n in the denominator. Proceeding to the limit as $n \to \infty$, we get

$$\lim_{n \to \infty} \log M_Z(t) = \frac{t^2}{2}$$

$$\Rightarrow \lim_{n \to \infty} M_Z(t) = \exp(t^2/2) \qquad \ldots(**)$$

Interpretation. (**) is the m.g.f. of standard normal variate [c.f. Remark to § 8·2·5]. Hence by uniqueness theorem of moment generating functions,

standard binomial variate tends to standard normal variate as $n \to \infty$. In other words, binomial distribution tends to normal distribution as $n \to \infty$.

Example 7·20. *A drunk performs a random walk over positions 0, ± 1, ± 2, ..., as follows. He starts at 0. He takes successive one unit steps, going to the right with probability p and to the left with probability (1 − p). His steps are independent. Let X denote his position after n steps. Find the distribution of (X + n)/2 and find E (X).* **(I.I.T. B.Tech., Dec. 1991)**

Solution. With the ith step of the drunk, let us associate a variable X_i defined as follows :

$X_i = 1$, if he takes the step to the right

$= -1$ if he takes the step to the left

Then $X = X_1 + X_2 + \ldots + X_n$, gives the position of the drunkard after n steps.

Define $Y_i = (X_i + 1)/2$

Then $Y_i = (1 + 1)/2 = 1$, with probability p

$= (-1 + 1)/2 = 0$, with probability $1 - p = q$, (say).

Since the n steps of drunkard are independent, Y_i's, $(i = 1, 2, \ldots n)$ are i.i.d. Bernonlli variates with parameter p.

Hence $\sum_{i=1}^{n} Y_i \sim B(n, p)$

$$\Rightarrow \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \left(\frac{X_i + 1}{2} \right) = \frac{1}{2} \left[\sum_{i=1}^{n} X_i + n \right] = \frac{X + n}{2} \sim B(n, p)$$

where $X = \sum_{i=1}^{n} X_i$, is the position of the drunkard after n steps.

Since $(X + n)/2 \sim B(n, p)$, we have

$$E \left[\frac{X + n}{2} \right] = np \Rightarrow \frac{1}{2} E(X + n) = np$$

$$\Rightarrow E(X) + n = 2np \Rightarrow E(X) = n(2p - 1)$$

Example 7·21. *Suppose that the r.v. X is uniformly distributed on (0,1)*

i.e., $f_X(x) = 1 ; 0 \le x \le 1$. ...(*)

Assume that the conditional distributional $Y|X = x$ has a binomial distribution with parameters n and p = x, i.e.,

$$P(Y = y | X = x) = \binom{n}{y} x^y (1 - x)^{n-y} ; \; y = 0, 1, 2, \ldots, n \qquad (**)$$

Find (a) E (Y)

(b) Find the distribution of Y. **(Punjab P.C.S., 1990)**

Solution. (a) We are given that the conditional distribution of

$Y|X = x \sim B(n, x)$ ...(i)

$\therefore E(Y|X = x) = nx$ ...(ii)

We have :

$$E(Y) = E[E(Y|X)] = E[nX] = nE(X) \qquad \text{[On using (ii)]}$$

$$\text{Now } E(X) = \int_0^1 x f(x)\,dx = \int_0^1 x\,dx = \frac{1}{2}.$$

$$\therefore\ E(Y) = n \times (\tfrac{1}{2}) = \frac{1}{2} n$$

(b) We have : $f_{X,Y}(x, y) = f_X(x) \cdot f_{Y|X}(y|x)$

Since X has (continuous) uniform distribution on (0,1) marginal distribution of Y is given by.

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,dx = \int_0^1 f_{Y|X}(y|x) \cdot f_X(x)\,dx$$

$$= \int_0^1 {}^nC_y \cdot x^y (1 - x)^{n-y} \cdot 1 \cdot dx \qquad \text{[using (*) and (**)]}$$

$$= {}^nC_y \int_0^1 x^y (1 - x)^{n-y}\,dx$$

$$= {}^nC_y \cdot \beta(y + 1, n - y + 1) = \frac{n!}{y!\,(n - y)!} \frac{\Gamma(y + 1)\,\Gamma(n - y + 1)}{\Gamma(n + 2)}$$

$$= \frac{n!}{y!\,(n - y)!} \times \frac{y!\,(n - y)!}{(n + 1)!}$$

$$= \frac{1}{n + 1} \quad ; \quad y = 0, 1, 2, \ldots, n$$

Since Y takes the values 0, 1, 2, ..., n each with equal probability $1/(n + 1)$, Y has discrete uniform distribution.

Remark We could find $E(Y)$ on using the distribution of Y in (*b*).

$$E(Y) = \sum_{y=0}^{n} y p(y) = \frac{1}{n + 1} \sum_{y=0}^{n} y$$

$$= \frac{1}{n + 1} [0 + 1 + 2 + \ldots + n] = \frac{n}{2},$$

as in Part (a).

Example 7·22. *If $K(t)$ is the cumulative function about the origin of the Binomial Distribution of size n, show that*

$$\frac{d}{dt} K(t) = n \left\{1 + e^{-(z+t)}\right\}^{-1}, \text{ where } z = \log_e (p/q)$$

(b) *By expanding the R.H.S. in powers of t by Taylor's Theorem, show that*

$$\kappa_r = n \frac{d^{r-1} p}{dz^{r-1}}, \text{ where } \kappa_r \text{ is the rth cumulant.}$$

(c) *Hence or otherwise obtain the recurrence relation*

$$\kappa_{r+1} = pq \cdot \frac{d\kappa_r}{dp}, \quad r > 1$$

[Baroda Univ. B.Sc. 1993; Delhi Univ. B.Sc. (Stat. Hons.) 1992]

(d) *Prove that* $\kappa_{r+1} = \frac{d\kappa_r}{dz}$, *where* $z = \log_e (p/q)$

Solution. For binomial distribution with parameters n and p, we have

$$K(t) = \log M(t) = n \log (q + pe^t)$$

(a) $$\frac{d}{dt} K(t) = \frac{npe^t}{q + pe^t} = n \left(1 + \frac{q}{p} e^{-t} \right)^{-1}$$

if $z = \log_e (p/q) \Rightarrow (p/q) = e^z \Rightarrow (q/p) = e^{-z}$, then

$$\frac{d}{dt} K(t) = n [\, 1 + e^{-(z+t)} \,]^{-1} \qquad \ldots(*)$$

(b) $$\kappa_r = \left[\frac{d^r}{dt^r} \kappa(t) \right]_{t=0} = \left[\frac{d^{r-1}}{dt^{r-1}} \cdot \frac{d}{dt} \kappa(t) \right]_{t=0}$$

$$= n \left[\frac{d^{r-1}}{dt^{r-1}} \left\{ 1 + e^{-(z+t)} \right\}^{-1} \cdot \right]_{t=0} = n \left[\frac{d^{r-1}}{dt^{r-1}} \left(\frac{e^{z+t}}{1 + e^{z+t}} \right) \right]_{t=0} \qquad \ldots(**)$$

By summetry of the function $e^{z+t}/(1 + e^{z+t})$ in t and z we have

$$\frac{d}{dt} \left(\frac{e^{z+t}}{1 + e^{z+t}} \right) = \frac{d}{dz} \left(\frac{e^{z+t}}{1 + e^{z+t}} \right)$$

$$\Rightarrow \quad \frac{d^{r-1}}{dt^{r-1}} \left(\frac{e^{z+t}}{1 + e^{z+t}} \right) = \frac{d^{r-1}}{dz^{r-1}} \left(\frac{e^{z+t}}{1 + e^{z+t}} \right)$$

Substituting in (**), we get

$$\kappa_r = n \left[\frac{d^{r-1}}{dz^{r-1}} \left(\frac{e^{z+t}}{1 + e^{z+t}} \right) \right]_{t=0} = n \frac{d^{r-1}}{dz^{r-1}} \left(\frac{e^z}{1 + e^z} \right)$$

$$= n \frac{d^{r-1}}{dz^{r-1}} (1 + e^{-z})^{-1} = n \frac{d^{r-1}}{dz^{r-1}} \left(1 + \frac{q}{p} \right)^{-1}$$

$$= n \frac{d^{r-1} p}{dz^{r-1}} \qquad \ldots(***)$$

(c) $$\frac{d\kappa_r}{dp} = n \frac{d}{dp} \left(\frac{d^{r-1} p}{dz^{r-1}} \right) = n \frac{d}{dz} \left(\frac{d^{r-1} p}{dz^{r-1}} \right) \frac{dz}{dp}$$

$$= n \frac{d^r p}{dz^r} \cdot \frac{1}{pq} \qquad [\because z = \log_e (p/q)]$$

$$= \frac{1}{pq} \cdot \kappa_{r+1} \qquad [\text{From } (***)]$$

(d) $\dfrac{d\kappa_r}{dz} = \dfrac{d\kappa_r}{dp} \cdot \dfrac{dp}{dz} = \dfrac{d\kappa_r}{dp} \Big/ \dfrac{dz}{dp} = \dfrac{d\kappa_r}{dp} \Big/ \dfrac{1}{pq} = pq \cdot \dfrac{d\kappa_r}{dp}$

$\therefore \quad \dfrac{d\kappa_r}{dz} = \kappa_{r+1}$ [c.f. part (*c*)]

Example 7·23. *If* $b(r; n, p) = \binom{n}{r} p^r q^{n-r}$ *is the binomial probability in the usual notation and if*

$$B(k; n, p) = P(X \le k) = \sum_{r=0}^{k} b(r; n, p),$$

then prove that

$$B(k; n, p) = (n - k)\binom{n}{k} \int_0^q t^{n-k-1}(1 - t)^k \, dt; \quad q = 1 - p$$

Solution. $B(k; n, p) = \sum_{r=0}^{k} b(r; n, p) = \sum_{r=0}^{k} \binom{n}{r} p^r q^{n-r}$

Differentiating w.r. to q and noting that $q = 1 - p \Rightarrow \dfrac{dq}{dp} = -1$, we get:

$$\frac{d}{dq} \cdot B(k; n, p) = \sum_{r=0}^{k} \left[\binom{n}{r} \left\{ r p^{r-1} (-1) \cdot q^{n-r} + p^r \cdot (n - r) q^{n-r-1} \right\} \right]$$

$$= \sum_{r=0}^{k} \left[\frac{n!(-r)}{r!(n-r)!} p^{r-1} q^{n-r} + \frac{n!(n-r)}{r!(n-r)!} p^r q^{n-r-1} \right]$$

$$= \sum_{r=0}^{k} \left[-\frac{n(n-1)!}{(r-1)!(n-r)!} p^{r-1} q^{n-r} + \frac{n(n-1)!}{r!(n-r-1)!} p^r q^{n-r-1} \right]$$

$$= \sum_{r=0}^{k} \left[n \cdot \binom{n-1}{r} p^r q^{n-r-1} - n \binom{n-1}{r-1} p^{r-1} q^{n-r} \right]$$

$$= \sum_{r=0}^{k} \left[n \left\{ t_r - t_{r-1} \right\} \right] \qquad \ldots(**)$$

where $t_r = \binom{n-1}{r} p^r q^{n-r-1}$ $\qquad \ldots(***)$

$$= n \left[\left(t_0 - t_{-1} \right) + \left(t_1 - t_0 \right) + \left(t_2 - t_1 \right) + \ldots + \left(t_k - t_{k-1} \right) \right]$$

$$= n t_k \qquad \left[\because t_{-1} = 0, \text{From } (***) \right]$$

$$\therefore \frac{d}{dq} \cdot B(k, n, p) = n \binom{n-1}{k} p^k \cdot q^{n-k-1}, \quad p = 1 - q$$

On integration, we get

$$B(k; n, p) = n \cdot \binom{n-1}{k} \int_0^q (1-u)^k \cdot u^{n-k-1}\, du.$$

But $n \cdot \binom{n-1}{k} = \frac{n \cdot (n-1)!}{k!\,(n-1-k)!} = \frac{n!\,(n-k)}{k!\,(n-k)!} = (n-k)\binom{n}{k}$

$$\therefore \quad B(k; n, p) = (n-k)\binom{n}{k} \int_0^q (1-u)^k \cdot u^{n-k-1}\, du$$

as desired.

Remarks. 1. We further get :

$$\beta(k+1,\ n-k) = \frac{\Gamma(k+1)\,\Gamma(n-k)}{\Gamma(n+1)} = \frac{k!\,(n-k-1)!}{n!}$$

$$\Rightarrow \frac{1}{\beta(k+1, n-k)} = \frac{n!}{k!\,(n-k-1)!} = (n-k)\binom{n}{k}$$

Hence the result may be written as :

$$B(k; n, p) = P(X \le k) = \frac{1}{\beta(k+1, n-k)} \int_0^q (1-u)^k u^{n-k-1}\, du$$

This result is of great practical utility. It enables us to represent the cumulative Binomial Probabilities (which are generally quite tedious and time consuming to compute) in terms of Incomplete Beta Functions which are tabulated in Karl Pearson's Tables of the Incomplete Beta Functions.

2 Let us now work out the probability :

$$P(X \ge k) = \sum_{r=k}^{n} \binom{n}{r} p^r\, q^{n-r}$$

Differentiating w.r. to p, and proceeding similarly, we shall get :

$$\frac{d}{dp} P(X \ge k) = -n \sum_{r=k}^{n} \left(T_r - T_{r-1}\right) \qquad \text{(Try it)}$$

where $\quad T_r = \binom{n-1}{r} p^r\, q^{n-r-1}, \quad (T_n = 0)$

$$\therefore \frac{d}{dp} P(X \ge k) = n\, T_{k-1} = n \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \quad (\because q = 1-p)$$

On integration, we shall get :

$$P(X \ge k) = n \binom{n-1}{k-1} \int_0^p u^{k-1} (1-u)^{n-k}\, du$$

$$P(X \ge k) = \frac{1}{\beta(k, n-k+1)} \int_0^p u^{k-1} (1-u)^{n-k}\, du$$

This is quite an important result and should be committed to memory. We shall use it in *'Order Statistics'*.

This result can be stated as follows :

If $X \sim B(n, p)$ and Y has Beta distribution with parameters k and $n - k + 1$ (c.f. Chapter 8), then

$$P(Y \leq p) = P(X \geq k) = 1 - P(X \leq k - 1)$$

$$\Rightarrow \quad F_Y(p) = 1 - F_X(k - 1)$$

EXERCISE 7 (a)

1. (*a*) Describe the probability model from which the binomial distribution can be generated. Hence find the first four central moments.

(*b*) If p is the probability of 'success' at a single trial, obtain the probability of r 'successes' out of n independent trails. Determine the mode of the resulting distribution.

2. (*a*) Define the binomial distribution with parameters p and n, and give a situation in real life where the distribution is likely to be realized. Obtain the moment generating function of the binomial distribution and hence or otherwise obtain the mean, variance, skewness and kurtosis of the distribution.

(*b*) Obtain the Moment Generating Function of the Binomial Distribution. Derive from it the result that the sum of two binomial variates is a binomial variate if the variates are independent and have the same probability of success.

3. The mean and variance of a binomial variate X with parameters n and p are 16 and 8. Find

(*i*) $P(X = 0)$, (*ii*) $P(X = 1)$, (*iii*) $P(X \geq 2)$.

4. For a Binomial distribution the mean is 6 and the standard deviation is $\sqrt{2}$. Write out all the terms of the distribution.

Ans. $n = 9,\ p = 2/3,\ q = 1/3;\ p(r) = (1/3)^9 \cdot \binom{9}{r} 2^r;\ r = 0, 1, 2, \ldots, 9$

5. (*a*) A perfect cube is thrown a large number of times in sets of 8. The occurrence of a 2 or 4 is called a success. In what proportion of the sets would you expect 3 successes.

Ans 27·31%

(*b*) In eight throws of a die, 5 or 6 is considered a success. Find the mean number of successes and the standard deviation. **(Ans. 2·66, 1·33)**

(*c*) A man tosses a fair coin 10 times. Find the probability that he will have

(*i*) heads on the first five tosses and tails on the next five tosses

(*ii*) heads on tosses 1, 3, 5, 7, 9 and tails on tosses 2, 4, 6, 8, 10.

(*iii*) 5 heads and 5 tails

(*iv*) at least 5 heads

(*v*) not more than 5 heads. **[Madras Univ.B.Sc.(Main Stat) Nov. 1991]**

Ans. (*i*) $(1/2)^{10}$, (*ii*) $(1/2)^{10}$, (*iii*) ${}^{10}C_5 (1/2)^{10}$

(*iv*) $\sum_{x=5}^{10} {}^{10}C_x \left(\frac{1}{2}\right)^{10}$ (*v*) $\sum_{x=0}^{5} {}^{10}C_x \left(\frac{1}{2}\right)^{10}$

6. (*a*) In 256 sets of twelve tosses of a fair coin, in how many cases may one expect eight heads and four tails ?

(Ans. 31) **(Delhi Univ. B.Sc. Oct. 1992)**

(*b*) In 100 sets of ten tosses of an unbaised coin, in how many cases should we expect

(*i*) Seven heads and three tails, (*ii*) at least seven heads ?

Ans. (*i*) 12, (*ii*) 17

7. (*a*) During war 1 ship out of 9 was sunk on an average in making a certain voyage. What was the probability that exactly 3 out of a convoy of 6 ships would arrive safely ? **(Madras Univ. B.Sc., 1992)**

Ans. $^6C_3\ (8/9)^3\ (1/9)^3$

(*b*) In the long run 3 vessels out of every 100 are sunk. If 10 vessels are out, what is the probability that

(*i*) exactly 6 will arrive safely, and

(*ii*) at least 6 will arrive safely ?

Hint. The probability 'p' that a vessel will arrive safely is

$$P = 97/100 = 0{\cdot}97 \text{ and } q = 0{\cdot}03$$

The probability that out of 10 vessels, x vessels will arrive safely is

$$p(x) = {}^{10}C_x p^x q^{10-x} = {}^{10}C_x (0{\cdot}97)^x (0{\cdot}03)^{10-x}$$

(*i*) Required probability $= p(6) = {}^{10}C_6 (0{\cdot}97)^6 (0{\cdot}03)^x$.

(*ii*) Required probability $= P(X \geq 6)$

8. (*a*) A student takes a true-false examination consisting of 10 questions. He is completely unprepared so he plans to guess each answer. The guesses are to be made at random. For example, he may toss a fair coin and use the outcome to determine his guess.

(*i*) Compute the probability that he guesses correctly at least five times.

(*ii*) Compute the probability that he guesses correctly at least 9 times.

(*iii*) What is the smallest n that the probability of guessing at least n correct answers is less than 1/2. **(Dibrugarh Univ. M.A., 1993)**

Ans. (*i*) 319/512; (*ii*) 11/1024; (*iii*) 6.

(*b*) A multiple choice test consists of 8 questions and 3 answers to each question, of which only one is correct. If a student answers each question by rolling a balanced die and checking the first answer if he gets 1 or 2, the second answer if he gets 3 or 4, and the third answer if he gets 5 or 6, find the probability of getting:

(*i*) exactly 3 correct answers,

(*ii*) no correct answer,

(*iii*) at least 6 correct answers. **[Gauhati Univ. M.A. (Econ.), 1993]**

9. (*a*) The incidence of occupational disease in an industry is such that the workers have a 20% chance of suffering from it. What is the probability that out of six workers chosen at random, four or more will suffer from the disease.

Ans. 52/3125

(*b*). (*a*) In a binomial distribution consisting of 5 independent trails, probabilities of 1 and 2 successes are 0·4096 and 0·2048 respectively. Find the parameter p of the distribution. **(Ans. 0·2)**

10. (*a*) With the usual notations, find p for a binomial random variable X if $n = 6$ and if $9P(X = 4) = P(X = 2)$. **(Ans. 0·25)**

(Mysore Univ. B.Sc. April 1992)

(*b*) X is a random variable following binomial distribution with mean 2·4 and variance 1·44. Find $P(X \geq 5)$, $P(1 < X \leq 4)$.

11. (*a*) In a certain town 20% of the population is literate, and assume that 200 investigators take a sample of ten individuals each to see whether they are literate. How many investigators would you expect to report that three people or less are literates in the sample ? **(Shivaji Univ. B.Sc., Oct. 1992)**

(*b*) A lot contains 1 per cent of defective items. What should be the number (n) of items in a random sample so that the probability of finding at least one defective in it, is at least 0·95 ? **(Ans. 68)**

12. (*a*) If on the average rain falls on ten days in every thirty days, find the probability

(*i*) that rain falls on at least three days of a given week,

(*ii*) that first three days of a given week will be dry and the remaining wet.

Ans. (*i*) $\sum_{x=3}^{7} {}^{7}C_x (1/3)^x (2/3)^{7-x}$, (*ii*) $(2/3)^3 \cdot (1/3)^4$.

(*b*) Suppose that weather records show that on the average 5 out of 31 days in October are rainy days. Assuming a binomial distribution with each day of October as an independent trial, find the probability that the next October will have at most three rainy days.

Ans. 0·2403

13. The probability of a man hitting a target is 1/4. (*i*) If he fires 7 times, what is the probability p of his hitting the target at least twice ? (*ii*) How many times must he fire so that the probability of his hitting the target at least once is greater than 2/3 ? [**Ans.** (*i*) 4547/8192, (*ii*) 4]

Hint. (*ii*) $p = \frac{1}{4}$, $q = \frac{3}{4}$. We want n such that

$$1 - q^n > \frac{2}{3} \Rightarrow q^n < \frac{1}{3} \Rightarrow \left(\frac{3}{4}\right)^n < \frac{1}{3} \Rightarrow n = 4$$

14. (*a*) The probability of a man hitting a target is 1/3. How many times must be fire so that the probability of hitting the target at least once is more than 90%. **Ans.** 6. **(Shivaji Univ. B.Sc., 1991)**

(*b*) Eight mice are selected at random and they are divided into two groups of 4 each. Each mouse in group A is given a dose of certain poison 'a' which is expected to kill one in four; each mouse in group B is given a dose of certain poison 'b' which is expected to kill one or two. Show that nevertheless, there may be fewer deaths in group A and find the probability of this happening.

Ans. 525/4096

15 (*a*) A card is drawn and replaced in an ordinary deck of 52 cards. How many times must a card be drawn so that (*i*) there is at least an even chance of drawing a heart, (*ii*) the probability of drawing a heart is greater than 3/4 ?

Ans. (*i*) 3, (*ii*) 5

(*b*) Five coins are tossed. What is the variance of the number of heads per toss of the five coins:

(*i*) if each coins is unbiased,

(*ii*) if the probability of a head appearing is 0·75 for each coin, and

(*iii*) if four coins are unbaised and for the fifth the probability of a head appearing is 0·75 ?

Hint (*iii*) Use generating function. [See Ex. 7·17 (*iii*)]

16. An owner of a small hotel with five rooms is considering buying television sets to rent to room occupants. He expects that about half of his customers would be willing to rent sets, and finally he buys three sets. Assuming 100% occupancy at all times :

(*i*) What fraction of the evenings will there be more request than T.V. sets?

(*ii*) What is the probability that a customer who requests a television set will receive one?

(*iii*) If the owner's cost per set per day is C, what rent R must he charge in order to break even (neither gain nor lose) in the long run ?

Hint. (*i*) Let the random variable X denote the daily number of requests. Then required probability is

$$P(X \geq 4) = P(X = 4) + P(X = 5) = \binom{5}{4}\left(\frac{1}{2}\right)^5 + \binom{5}{5}\left(\frac{1}{2}\right)^5$$

(*ii*) The customer can get a T.V. in the following mutually exclusive ways,

(*a*) There are no other requests that night.

(*b*) There is one other request.

(*c*) There are two other requests.

(*d*) There are three other requests and his request precedes at least one of them.

(*e*) There are four other requests, and his request preceedes at least two of them.

The probability of the desired event

$$= (0\cdot5)^4 \left\{1 + {}^4C_1 + {}^4C_2 + \tfrac{3}{4}\cdot{}^4C_3 + \tfrac{3}{5}\,{}^4C_4\right\}$$

(*iii*) Mean revenue

$$= (0\cdot5)^5 \cdot 0 + {}^5C_1\,(0\cdot5)^5 R + {}^5C_2\,(0\cdot5)^5 2R + \left\{{}^5C_3\,(0\cdot5)^5 + {}^5C_4\,(0\cdot5)^5 + {}^5C_5\,(0\cdot5)^5\right\} 3R$$

$$= \frac{73}{32}R$$

The break-even rental is the value of R for which

$$\frac{73}{32}R = 3C \Rightarrow R = 1\cdot315\,C$$

17. A manufacturer claims that at most 10 per cent of his product is defective. To test this claim, 18 units are inspected and his claim is accepted if among these 18 units, at most 2 are defective. Find the probability that the manufacturer's claim will be accepted if the actual probability that a unit is defective is

(*a*) 0·05 (*b*) 0·10 (*c*) 0·15 and (*d*) 0·20.

Ans. (*a*) 0·9410 (*b*) 0·9326 (*c*) 0·4445 (*d*) 0·2715

18. (*a*) A set of 8 symmetrical coins was tossed 256 times and the frequencies of throws observed were as follows :

Number of heads :	0	1	2	3	4	5	6	7	8
Frequency of throws:	2	6	24	63	64	50	36	10	1

Fit a binomial distribution and find mean and standard deviation of fitted distribution.

(*b*) A set of 6 similar coins is tossed 640 times with the following results:

Number of heads :	0	1	2	3	4	5	6
Frequency :	7	64	140	210	132	75	12

Calculate the binomial frequencies on the assumption that the coins are symmetrical.

19. (*a*) The following data due to Weldon shows the results of throwing 12 dice 4096 times, a throw of 4, 5 or 6 being called a success (x).

x:	0	1	2	3	4	5	6	7	8	9	10	11	12	Total
f:	—	7	60	198	430	731	948	847	536	257	71	11	—	4096

Fit the binomial distribution and calculate the expected frequencies. Compare the actual mean and S.D. with those of the expected ones for the distribution.

Ans. Expected freq. : 1, 12, 66, 220 495 792, 924, 792, 495, 220, 66, 12, 0; mean = 6, variance = 1·71.

(*b*) In 103 litters of 4 mice, the number of litters which contained 0, 1, 2, 3, 4 females are recorded below :

Number of female mice	0	1	2	3	4	Total
Number of litters	8	32	34	24	5	103

(*i*) If the chance of obtaining a female in a single trial is assumed constant, estimate the constant but unknown probability.

(*ii*) If the size of the litter 4 had not been given, how could it be estimated from the data ?

20. X is random variable distributed according to the Binomial law :

$$b(x\,;n,p) = \binom{n}{x} p^x q^{n-x};\ x = 0, 1, 2, \ldots, n$$

Obtain the recurrence formula :

$$b(x+1\,;n,p) = \frac{n-x}{x+1} \cdot \frac{p}{q} \cdot b(x\,;n,p)$$

Use this as a reduction formula and get the theoretical frequencies when an unbaised coin is tossed 8 times and the experiment is repeated 256 times.

(Madras Univ. B. Sc. April 1992)

21. (*a*) By differentiating the following identity with respect to p and then multiplying by p,

$$\sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (q + p)^n, q = 1 - p$$

prove that $\mu_1' = np$ and $\mu_2 = npq$.

22. (*a*) Let $X \sim b(x; n, p)$ and r be a non-negative integer. If the rth moment about the origin is denoted by $\mu_r' = E(X^r)$, prove that

$$\mu_{r+1}' = np\,\mu_r' + p(1 - p)\,\frac{d\mu_r'}{dp}$$

[Delhi Univ. B.Sc. (Hons. Subs.), 1993, '88]

(*b*) Show that for the binomial distribution $B(n, p)$,

$$\mu_{r+1} = pq\left(nr\,\mu_{r-1} + \frac{d}{dp}\mu_r\right), \quad p + q = 1,$$

where symbols have their usual meanings.

[Delhi Univ. B.Sc. (Stat. Hons), 1989]

(c) If $X \sim B(n, p)$, obtain the recurrence relation for its central moments and hence find values of β_1 and β_2.

[Calcutta Univ. B.Sc. (Hons.), 1992]

23. (*a*) The following results were obtained when 100 batches of seeds were allowed to germinate on damp filter paper in a laboratory :

$$\beta_1 = \frac{1}{15} \text{ and } \beta_2 = \frac{89}{30}$$

Determine the binomial distribution and calculate the frequency for $X = 8$, considering $p > q$.

Hint. We have $\beta_1 = \dfrac{(q - p)^2}{npq} = \dfrac{1}{15}$ ...(*i*)

and $\beta_2 = 3 + \dfrac{1 - 6pq}{npq} = \dfrac{89}{30}$ ...(*ii*)

From (*i*) and (*ii*), we can find the value of n, p and q

(*b*) Between a Binomial distribution with $n = 5$ and $p = \frac{1}{2}$ and a distribution with frequency function

$$f(x) = 6x(1 - x),\ 0 \le x \le 1;$$

determine which is more skewed.

24. (*a*) $x = r$ is the unique mode of Binomial Distribution having mean np and variance $np(1 - p)$. Show that

$$(n + 1)p - 1 < r < (n + 1)p$$

Find the mode of the binomial distribution with $p = \frac{1}{2}$ and $n = 7$.

[Delhi Univ. B.Sc. (Stat. Hons.) 1991, '84]

Ans. 4, 3 (Bimodal).

(*b*) Show that if np be a whole number, the mean of the binomial distribution coincides with the greatest term.

(*c*) Compute the mode of a binomial distribution $b\left(7, \frac{1}{2}\right)$.

[Delhi Univ. B.Sc. (Maths. Hons.), 1989]

Ans. 1, 2 (Bimodal).

(*d*) Define Bernoulli trials and state the binomial law of probability. Find the bounds for the most probable number of successes in a sequence of n Bernoulli trials.

One workers can manufacture 120 articles during a shift, another worker 140 articles, the probabilities of the articles being of a high quality are 0·94 and 0·80 respectively. Determine the most probable number of high quality articles manufactured by each worker. **[Calcutta Univ. B.Sc. (Maths. Hons.), 1988]**

25. Show that if two symmetrical binomial distributions $(p = q = \frac{1}{2})$ of degree n (and of the same number of observations) are so superimposed that the rth term of one coincides with the $(r + 1)$th term of the other, the distribution formed by adding superimposed terms is a symmetrical binomial of degree $(n + 1)$. **[Bhagalpur Univ. B.Sc., 1993]**

26. (*a*) Let X denote a binomially distributed random variable. Show that

$$E\left(\frac{X-np}{\sqrt{npq}}\right) = 0,\ E\left(\frac{X-np}{\sqrt{npq}}\right)^2 = 1, \text{ and}$$

$$E\left[\exp.\left\{t\left(\frac{X-np}{\sqrt{npq}}\right)\right\}\right] = \left[(1-p)\exp\left\{-t\sqrt{\left(\frac{p}{nq}\right)}\right\} + p\exp\left\{t\sqrt{\left(\frac{q}{np}\right)}\right\}\right]$$

(*b*) Obtain the characteristic function of the standard binomial variate $(X - np)/\sqrt{npq}$, where X is the number of successes obtained in n independent trials, each with constant probability p of success, $q = 1 - p$. Obtain the limit of this function as $n \to \infty$. **[Delhi Univ. B.Sc.(Maths. Hons.), 1991]**

(*c*) If $X \sim B(n, p)$, prove that

$$\kappa_{r+1} = pq \cdot \frac{d}{dp}(\kappa_r),$$

where κ_r is the rth cumulant.

Hence deduce the values of κ_2 and κ_3.

[Delhi Univ. B.Sc. (Stat. Hons.), 1991, '87]

27. (*a*) If X and Y are two independent identically distributed binomial variates, obtain the probability that the absolute difference $|X - Y|$ equals a given value, say r.

(*b*) (*i*) If X and Y are independent binomial variates, with parameters p_1 and p_2 and indices n_1 and n_2 respectively, obtain the probability that $X + Y$ equals 'r'.

(*ii*) In the above if $p_1 = p_2$, what is the distribution of $X + Y$?

[Poona Univ. B.Sc., 1988]

(*c*) If X and Y are two independent binomial variates with parameters $n_1 = 6$, $p = 1/2$ and $n_2 = 4$, $p = 1/2$ respectively, evaluate,

(*i*) $P(X + Y = r)$, (*ii*) $P(X + Y \geq 3)$

(Gujarat Univ. B. Sc. Oct. 1992)

Hint $X + Y \sim B(6 + 4, 1/2) = B(10, 1/2)$

Ans. (*i*) $P(X + Y = r) = p(r) = {}^{10}C_r\,(1/2)^r$; $r = 0, 1, ..., 10$

(*ii*) $P(X + Y \geq 3) = 1 - [p(0) + p(1) + p(2)] = 0{\cdot}945$

(*d*) If X and Y are two independent binomial variates with parameters $(n_1 = 3, p = 0{\cdot}4)$ and $(n_2 = 4, p = 0{\cdot}4)$ respectively, find:

(*i*) $P(X = Y)$, (*ii*) $P(X + Y \leq 2)$, (*iii*) $P(X = 3 \mid X + Y = 4)$

Hint. $X + Y \sim B(3 + 4, 0{\cdot}4) = B(7, 0{\cdot}4)$

(*i*) $P(X = Y) = \sum_{r=0}^{3} P(X = r \cap Y = r) = \sum_{r=0}^{3} P(X = r)\,P(Y = r) = 0{\cdot}2871$

(*ii*) $P(X + Y \leq 2) = \sum_{r=0}^{2} \binom{7}{r} (0{\cdot}4)^r (0{\cdot}6)^{7-r} = 0{\cdot}420$

(*iii*) $P(X = 3 \mid X + Y = 4) = \dfrac{P(X = 3 \cap X + Y = 4)}{P(X + Y = 4)} = \dfrac{P(X = 3 \cap Y = 1)}{P(X + Y = 4)} = 0{\cdot}1141$

28. (*a*) Obtain the moment generating function of Binomial distribution with $n = 7$ and $p = 0{\cdot}6$. Find the first three moments of the distribution.

[Poona Univ.B. Sc. 1992]

Ans. $(0{\cdot}4 + 0{\cdot}6\,e^t)^7$: mean $= 4{\cdot}2$, $\mu_2 = 1{\cdot}68$, $\mu_3 = -0{\cdot}336$.

(*b*) Suppose that the m.g.f. of a random variable X is of the form

$$M_X(t) = (0{\cdot}4\,e^t + 0{\cdot}6)^8$$

What is the m.g.f. of the random variable $Y = 3X + 2$? Evaluate $E(X)$.

Ans. $E(X) = 3{\cdot}2$, $M_Y(t) = e^{2t}\,(0{\cdot}6 + 0{\cdot}4\,e^{3t})^8$

(*c*) Obtain the moment generating function of the binomial distribution. Hence or otherwise obtain the mean, variance and skewness of the distribution.

29. Show that the factorial moment generating function $w(t)$ of the binomial distribution $b(x; n, p)$ is $(1 + pt)^n$. Hence or otherwise show that

$$\mu_{(r)}' = n^{(r)}\,p^r$$

Hint. Factorial moment generating function $\omega(t)$ is defined as

$$\omega(t) = E(1 + t)^X = \sum_x (1 + t)^x\,p(x) = \sum_x {}^nC_x \left\{p(1 + t)\right\}^x q^{n-x}$$

$$\mu_{(r)}' = \text{co efficient of } \frac{t^r}{r!} \text{ in } \omega(t) = {}^nC_r\,r!\,p^r = n^{(r)}\,p^r$$

30. Show that

(*i*) $b(n, p; k) = b(n, 1 - p; n - k)$

(*ii*) $\sum_{k=r}^{n} b(n, p; k) = 1 - \sum_{k=n-r+1}^{n} b(n, 1 - p; k)$

(*iii*) $b(n + 1, p; k) = p \cdot b(n, p; k - 1) + q \cdot b(n, p; k)$

Hint. (*i*) $b(n, 1-p; n-k) = \binom{n}{n-k} (1-p)^{n-k} p^{n-(n-k)}$

(*ii*) $\sum_{k=r}^{n} b(n, p; k) = \sum_{k=r}^{n} b(n, 1-p; n-k) = \sum_{k=0}^{n-r} b(n, 1-p; k)$

31. For a binomial distribution, let

$$F_n(y) = \sum_{x=0}^{y} \binom{n}{x} p^x q^{n-x},$$

where $q = 1 - p$,

prove that

(*i*) $F_{n+1}(y) = p F_n(y-1) + q F_n(y)$

(*ii*) Cov $(X, n-X) = -npq$ **(Bombay Univ. B.Sc., April 1990)**

32. (*a*) Random variable X follows binomial distribution with parameters $n = 40$ and $p = \frac{1}{4}$. Use Chebychev's inequality to find bounds for

(*i*) $P[|X - 10| < 8]$; (*ii*) $P[|X - 10| > 10]$

Compare these values with the actual values (**Hint** : Use Normal approximation for the Binomial). **(Madras Univ. B.Sc. (Main Stat.), 1988)**

Ans. (*i*) 11 3/128 (lower bound), (*ii*) 0.075 (upper bound).

(*b*) X follows binomial distribution with $n = 40$, $p = \frac{1}{2}$. Use Chebychev's lemma to

(*i*) find k such that

$P\{|X - 20| > 10k\} \leq 0{\cdot}25$, and

(*ii*) obtain a lower limit for $P\{|X - 20| \leq 5\}$.

[Delhi Univ. B.Sc. (Maths. Hons.), 1984]

Ans. (*i*) $2\sqrt{10}$, (*ii*) 3/5

(*c*) How many trials must be made of an event with binomial probability of success $\frac{1}{2}$ in each trial, in order to be assured with probability of at least 0·9 that the relative frequency of success will be between 0·48 and 0·52 ? **(Ans. 6250)**

Hint. Use Chebychev's Inequality.

33. (*a*) Show that if a coin is tossed n times, the probability of not more than k heads is :

$$\left[\binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{k}\right] \left(\frac{1}{2}\right)^n$$

[South Gujarat Univ. B.Sc., 1988]

(*b*) If X has binomial distribution with paramctes n and p, then prove that $P[X \text{ is even}] = \frac{1}{2}[1 + (q-p)^n]$. **[Delhi Univ. B.Sc. (Stat. Hons.), 1988]**

34. If the probability of hitting a target is 1/5 and if 10 shots are fired, what is the conditional probability of the target being hit at least twice assuming that at least one hit is already scored ?

[Nagpur Univ. B.Sc., 1988, '93]

Hint. Let X donote the number of times a target is hit when 10 shots are fired. Then $X \sim B(10, 0{\cdot}2)$. The required probability is :

$$P(X \geq 2 \mid X \geq 1) = \frac{P[(X \geq 2) \cap (X \geq 1)]}{P(X \geq 1)} = \frac{P(X \geq 2)}{P(X \geq 1)}$$

$$= \frac{1 - [P(X = 0) + P(X = 1)]}{1 - [P(X = 0)]} = \frac{0{\cdot}625}{0{\cdot}893} = 0{\cdot}6999$$

35. (*a*) Let X be a $B(2, p)$ and Y be a $B(4, p)$. If $P(X \geq 1) = 5/9$, find $P(Y \geq 1)$ **[Kerala Univ. B. Sc., 1989]**

Hint. $P(X \geq 1) = 1 - P(X = 0) = 1 - q^2 = 5/9 \Rightarrow q = 2/3, p = 1/3.$

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - q^4 = 65/81.$$

36. Let B denote the number of boys in a family with five children. If p denotes the probability that a boy is there in a family, find the least value of p such that

$$P(B = 0) > P(B = 1)$$ **(Shivaji Univ. B. Sc., 1990)**

Ans. $q^5 > 5pq^4 \Rightarrow q > 5p \Rightarrow p < \frac{1}{6}$.

37. (*a*) Suppose $X \sim B(n, p)$. with $E(X) = 5$, Var $(X) = 4$. Find n and p. (Ans. $n = 25$, $p = 1/5$)

(*b*) Let $X \sim B(n, p)$. For what p is variance (X) maximised if we assume n is fixed.

Ans. Var $X = npq = n(p - p^2) = f(p)$, (say) ; $f'(p) = 0$, $f''(p) < 0$; $p = 1/2 = q$

38. (*a*) $X \sim B(n = 100, p = 0{\cdot}1)$. Find $P(X \leq \mu_x - 3\sigma_x)$

Ans. $\mu = 10$, $\sigma = 3$, $P(X \leq \mu_x - 3\sigma_x) = P(X \leq 1) = 10{\cdot}9 \times (0{\cdot}9)^{99}$

(*b*) If $X \sim B(25, 0{\cdot}2)$, find $P(X < \mu_x - 2\sigma_x)$

[Delhi Univ. B.A. (Stat. Hons.) Spl. Course 1989]

39. For one half of n events, the chance of success is p, and the chance of failure is q, whilst for the other half the chance of success is q, and the chance of failure is p. Show that the S.D. of the number of successes is the same as if the chance of success were p in all the cases i.e. $\sqrt{npq}$, but that the mean of the number of successes is $n/2$ and not np. **(Delhi Univ. B.A. 1992)**

Hint. $X \sim B(n/2, p)$ and $Y \sim B(n/2, q)$ are independent. Let $Z = X + Y$. Now prove that Var $(Z) = npq$ and $E(Z) = n/2$.

40. The discrete density of X is given by $f_X(x) = x/3$, for $x = 1, 2$ and $f_{Y|X}(y \mid x)$ is binomial with parameters x and $\frac{1}{2}$ i.e.,

$$F_{Y|X}(y \mid x) = P(Y = y \mid X = x) = \binom{x}{y} \cdot \left(\tfrac{1}{2}\right)^x;$$

for $y = 0, 1, ..., x$ and $x = 1, 2$.

(*a*) Find $E(X)$ and Var (X); (*b*) Find $E(Y)$

(c) Find the joint distribution of X and Y.

Hint. Proceed as in Example 7·21.

Ans. (*a*) $E(X) = 5/3$, $\text{Var}(X) = 2/9$, (*b*) $E(Y) = 5/6$.

(*c*) $f(x,y) = \binom{x}{y} \cdot \left(\frac{x}{3}\right) \cdot \left(\frac{1}{2}\right)^x$; $n = 1, 2, ;\ y = 0, 1, ..., x$.

41. Two dice are thrown n times. Let X denote the number of throws in which the number on the first dice exceeds the number on the second dice. What is the distribution of X?

Ans. $X \sim B(n, p = 15/36)$

Hint. p is the probability that the number on the first dice exceeds the number on the second dice in a throw of two dice.

42. Let $X_1 \sim B(n, p_1)$ and $X_2 \sim B(n, p_2)$.

If $p_1 < p_2$, prove that:

$$P(X_1 \le k) \ge P(X_2 \le k) \text{ for } k = 0, 1, ..., n.$$

Hint. Use Example 7·23.

43. If $X \sim B(n, p)$, show that

$$P(X \le k) = \lambda \int_{p/q}^{\infty} \frac{y^k}{(1+y)^{n+1}}\, dy$$

where $$\lambda^{-1} = \int_0^{\infty} \frac{y^k}{(1+y)^{n+1}}\, dy = \beta(k+1, n-k)$$

Hint. $\frac{d}{dq} P(X \le k) = n \binom{n-1}{k} \cdot p^k \cdot q^{n-k-1} = A_k$, (*say*)

[See Example 7·23]

Find $\frac{d}{dq}(\text{RHS}) = \lambda \cdot \frac{d}{dq}\left(\int_{p/q}^{\infty} \frac{y^k}{(1+y)^{n+1}}\, dy\right) = \lambda \frac{(p/q)^k}{[1+(p/q)]^{n+1}} \left(\frac{1}{q^2}\right)$

$$= \frac{1}{\beta(k+1, n-k)} \cdot p^k \cdot q^{n-k-1} = A_k$$

(On simplification)

44. If X $B(n,p)$ and Y has beta distribution with parameters k and $n - k + 1$, (See Chapter 8), then prove that

$P(Y \le p) = P(X \ge k)$ *i.e.*, $F_Y(p) = 1 - F_X(k-1)$

45. If a fair coin is tossed an even number $2n$ times, show that the probability of obtaining more heads than tails is

$$\frac{1}{2}\left\{1 - {}^{2n}C_n \left(\frac{1}{2}\right)^{2n}\right\}$$

Hint. X : No. of heads; Y = No. of tails; No. of trials = $2n$

$P(X > Y) + P(X < Y) + P(X = Y) = 1$

$$\Rightarrow \qquad 2P(X > Y) = 1 - P(X = Y)$$

[$\because$ By symmetry, $p = q = \frac{1}{2} \Rightarrow P(X > Y) = P(X < Y)$]

$$= 1 - {}^{2n}C_n p^n \cdot q^n = 1 - {}^{2n}C_n \left(\tfrac{1}{2}\right)^{2n}$$

$$\Rightarrow \qquad P(X > Y) = \frac{1}{2}\left[1 - {}^{2n}C_n \left(\tfrac{1}{2}\right)^{2n}\right]$$

7·3·0. Poisson Distribution (as a limiting case of Binomial Distribution). Poisson distribution was discovered by the French mathematician and physicist Simeon Denis Poisson (1781—1840) who published it in 1837. Poisson distribution is a limiting case of the binomial distribution under the following conditions:

(*i*) n, the number of trials is indefinitely large, *i.e.*, $n \to \infty$.

(*ii*) p, the constant probability of success for each trial is indefinitely small, *i.e.*, $p \to 0$.

(*iii*) $np = \lambda$, (say), is finite. Thus $p = \lambda/n$, $q = 1 - \lambda/n$, where λ is a positive real number.

The probability of x successes in a series of n independent trials is

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}; x = 0, 1, 2, ..., n \qquad ...(*)$$

We want the limiting form of (*) under the above conditions. Hence

$$\lim_{n \to \infty} b(x; n, p) = \lim_{n \to \infty} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \cdot \left[1 - \frac{\lambda}{n}\right]^{n-x}$$

Using Stirling's approximation for $n!$ as $n \to \infty$ *viz.*,

$$\lim_{n \to \infty} n! \approx \sqrt{2\pi}\, e^{-n} n^{n+(1/2)},$$ we get

$$\lim_{n \to \infty} b(x; n, p) = \lim_{n \to \infty} \left[\frac{\sqrt{2\pi}\, e^{-n} \cdot n^{n+(1/2)}}{x!\sqrt{2\pi}\, e^{-(n-x)} \cdot (n-x)^{n-x+(1/2)}}\right] \left(\frac{\lambda}{n}\right)^x \left[1 - \frac{\lambda}{n}\right]^{n-x}$$

$$= \frac{\lambda^x}{e^x \cdot x!} \cdot \lim_{n \to \infty} \frac{n^{n-x+(1/2)}}{(n-x)^{n-x+(1/2)}} \cdot \left[1 - \frac{\lambda}{n}\right]^{n-x}$$

$$= \frac{\lambda^x}{e^x x!} \lim_{n \to \infty} \frac{\left(1 - \frac{\lambda}{n}\right)^{n-x}}{\left(1 - \frac{x}{n}\right)^{n-x+(1/2)}}$$

$$= \frac{\lambda^x}{e^x x!} \cdot \frac{\lim\limits_{n \to \infty} \left[1 - \frac{\lambda}{n}\right]^n \cdot \lim\limits_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^{-x}}{\lim\limits_{n \to \infty} \left[1 - \frac{x}{n}\right]^n \lim\limits_{n \to \infty} \left[1 - \frac{x}{n}\right]^{-x+(1/2)}}$$

But we know that

and
$$\left.\begin{aligned} &\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n} = e^{-\lambda}, \\ &\lim_{n\to\infty}\left[1-\frac{\lambda}{n}\right]^{\alpha} = 1,\ \alpha \text{ is not a function of } n \end{aligned}\right\} \quad ...(**)$$

Therefore

$$\lim_{n\to\infty} b(x;n,p) = \frac{\lambda^x}{e^x\cdot x!}\cdot\frac{e^{-\lambda}\cdot 1}{e^{-x}\cdot 1} = \frac{e^{-\lambda}\cdot\lambda^x}{x!};\ x = 0, 1, 2, ..., \infty;$$

[Using (**)]

which is the required probability function of the Poisson distribution. 'λ' is known as the parameter of Poisson distribution.

Aliter. Poisson distribution can also be derived without using Stirling's approximation as follows :

$$b(x;n,p) = \binom{n}{x}p^x(1-p)^{n-x} = \binom{n}{x}\left[\frac{p}{1-p}\right]^x(1-p)^n$$

$$= \frac{n(n-1)(n-2)\ldots(n-x+1)}{x!}\cdot\frac{\left(\frac{\lambda}{n}\right)^x}{\left[1-\frac{\lambda}{n}\right]^x}\left[1-\frac{\lambda}{n}\right]^n$$

$$= \frac{\left[1-\frac{1}{n}\right]\left[1-\frac{2}{n}\right]\ldots\left[1-\frac{x-1}{n}\right]}{x!\left[1-\frac{\lambda}{n}\right]^x}\lambda^x\left[1-\frac{\lambda}{n}\right]^n$$

$$\therefore \lim_{n\to\infty} b(x;n,p) = \frac{e^{-\lambda}\lambda^x}{x!};\ x = 0, 1, 2, \ldots \qquad \text{[From (**)]}$$

Definition. *A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its probability mass function is given by*

$$p(x,\lambda) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!};\ x = 0, 1, 2, \ldots;\ \lambda > 0$$
$$= 0, \textit{ otherwise} \qquad ...(7\cdot 14)$$

Here λ is known as the parameter of the distribution.

We shall use the notation $X \sim P(\lambda)$ to denote that X is a Poisson variate with paramter λ.

Remarks 1. It should be noted that

$$\sum_{x=0}^{\infty} P(X = x) = e^{-\lambda} \sum_{x=0}^{\infty} \lambda^x / x! = e^{-\lambda} e^{\lambda} = 1$$

2. The corresponding distribution function is:

$$F(x) = P(X \leq x) = \sum_{r=0}^{x} p(r) = e^{-\lambda} \sum_{r=0}^{x} \lambda^r / r! \; ; x = 0, 1, 2, \ldots.$$

3. Poisson distribution occurs when there are events which do not occur as outcomes of a definite number of trials (unlike that in binomial) of an experiment but which occur at random points of time and space wherein our interest lies only in the number of occurrences of the event, not in its non-occurrences.

4. Following are some instances where Poisson distribution may be successfully employed·

(1) Number of deaths from a disease (not in the form of an epidemic) such as heart attack or cancer or due to snake bite.

(2) Number of suicides reported in a particular city.

(3) The number of defective material in a packing manufactured by a good concern.

(4) Number of faulty blades in a packet of 100.

(5) Number of air accidents in some unit of time.

(6) Number of printing mistakes at each page of the book.

(7) Number of telephone calls received at a particular telephone exchange in some unit of time or connections to wrong numbers in a telephone exchange.

(8) Number of cars passing a crossing per minute during the busy hours of a day.

(9) The number of fragments received by a surface area 't' from a fragment atom bomb.

(10) The emission of radioactive (alpha) particles.

7·3·1. The Poisson Process. The Poisson distribution may also be obtained independently (*i.e.*, without considering it as a limiting form of the Binomial distribution) as follows :

Let X_t be the number of telephone calls received in time interval 't' on a telephone switch board. Consider the following experimental conditions :

(1) The probability of getting a call in small time interval $(t, t + dt)$ is $\lambda\, dt$, where λ is a positive constant and dt denotes a small increment in time 't'.

(2) The probability of getting more than one call in this time interval is very small, *i.e.*, is of the order of $(dt)^2$ *i.e.*, $0[(dt)^2]$ such that

$$\lim_{dt \to 0} \frac{0(dt)^2}{dt} = 0$$

(3) The probability of any particular call in the time interval $(t, t + dt)$ is independent of the actual time t and also of all previous calls.

Under these conditions it can be shown that the probability of getting x calls in time 't', say, $P_x(t)$ is given by

$$P_x(t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}; \; x = 0, 1, 2, ..., \infty$$

which is a Poisson distribution with parameter λt.

Proof. Let $P_x(t) = P$ {of getting x calls in a time interval of length 't'}. Also P {of at least one call during $(t, t + dt)$} $= \lambda\, dt + 0\,[(dt)^2]$

and P {of more than one call during $(t, t + dt)$} $= 0\,[(dt)^2]$.

The event of getting exactly x calls in time $t + dt$ can materialise in the following two mutually exclusive ways :

(*i*) x calls in $(0, t)$ and none during $(t, t + dt)$ and the probability of this event is $P_x(t)\left\{1 - [\lambda\, dt + 0\,(dt)^2]\right\}$,

(*ii*) exactly $(x - 1)$ calls during $(0, t)$ and one call in $(t, t + dt)$ and the probability of this event is $P_{x-1}(t)\,(\lambda\, dt)$.

Hence by the addition theorem of probability, we get

$$P_x(t + dt) = P_x(t)\left\{1 - \lambda\, dt - 0\,(dt)^2\right\} + P_{x-1}(t)\,\lambda\, dt.$$

$$= P_x(t)(1 - \lambda\, dt) + P_{x-1}(t)\,\lambda\, dt + 0\,(dt)^2 P_x(t) \quad ...(1)$$

$$\Rightarrow \frac{P_x(t + dt) - P_x(t)}{dt} = -\lambda P_x(t) + \lambda P_{x-1}(t) + \frac{0\,(dt)^2}{dt} P_x(t)$$

Proceeding to the limit as $dt \to 0$, we get

$$\lim_{dt \to 0} \frac{P_x(t + dt) - P_x(t)}{dt} = -\lambda P_x(t) + \lambda P_{x-1}(t)$$

$$\therefore \quad P_x'(t) = -\lambda P_x(t) + \lambda P_{x-1}(t), x \geq 1 \quad ...(2)$$

where $(')$ denotes differentiation w.r. to 't'.

For $x = 0, P_{x-1}(t) = P_{-1}(t) = P\{(-1) \text{ calls in time 't'}\} = 0$

Hence from (1), we get

$$P_0(t + dt) = P_0(t)\left\{1 - \lambda\, dt\right\} + 0\,(dt)^2$$

which on taking the limit $dt \to 0$, gives

$$P_0'(t) = -\lambda P_0(t) \Rightarrow \frac{P_0'(t)}{P_0(t)} = -\lambda$$

Integrating w.r. to. 't', we get

$$\log P_0(t) = -\lambda t + C,$$

where C is an arbitrary constant to be determined from the condition

$$P_0(0) = 1$$

Hence $\quad \log 1 = C \Rightarrow C = 0$

$$\therefore \quad \log P_0(t) = -\lambda t \Rightarrow P_0(t) = e^{-\lambda t}$$

Substituting this value of $P_0(t)$ in (2), we get, with $x = 1$

$$P_1'(t) = -\lambda P_1(t) + \lambda e^{-\lambda t}$$

$$\Rightarrow \quad P_1'(t) + \lambda P_1(t) = \lambda e^{-\lambda t}$$

This is an ordinary linear differential equation whose integrating factor is $e^{\lambda t}$. Hence its solution is

$$e^{\lambda t} P_1(t) = \lambda \int e^{\lambda t} e^{-\lambda t}\, dt + C_1 = \lambda t + C_1,$$

where C_1 is an arbitrary constant to be determined from $P_1(0) = 0$, which gives $C_1 = 0$.

$$\therefore \qquad P_1(t) = e^{-\lambda t} \lambda t$$

Again substituting this in (2) with $x = 2$, we get

$$P_2'(t) + \lambda P_2(t) = \lambda e^{-\lambda t}\, \lambda t$$

Integrating factor of this equation is $e^{\lambda t}$ and its solution is

$$P_2(t)\, e^{\lambda t} = \lambda^2 \int t e^{-\lambda t}\, e^{\lambda t}\, dt + C_2 = \frac{\lambda^2 t^2}{2} + C_2$$

where C_2 is an arbitrary constant to be determined from $P_2(0) = 0$, which gives $C_2 = 0$. Hence

$$P_2(t) = e^{-\lambda t} \frac{(\lambda t)^2}{2}$$

Proceeding similarly step by step, we shall get

$$P_x(t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}; \; x = 0, 1, 2, \ldots, \infty.$$

7·3·2. Moments of the Poisson Distribution

$$\mu_1' = E(X) = \sum_{x=0}^{\infty} x\, p(x, \lambda)$$

$$= \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda} \lambda^x}{x!} = \lambda e^{-\lambda} \left[\sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \right]$$

$$= \lambda e^{-\lambda} \left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots \right) = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$$

Hence the mean of the Poisson distribution is λ.

$$\mu_2' = E(X^2) = \sum_{x=0}^{\infty} x^2 p(x, \lambda) = \sum_{x=0}^{\infty} \{x(x-1) + x\} \frac{e^{-\lambda} \lambda^x}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} x(x-1) \frac{\lambda^x}{x!} + \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!}$$

$$= \lambda^2 e^{-\lambda} \left[\sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} \right] + \lambda = \lambda^2 e^{-\lambda} e^{\lambda} + \lambda = \lambda^2 + \lambda$$

$$\mu_3' = E(X^3) = \sum_{x=0}^{\infty} x^3 p(x, \lambda)$$

$$= \sum_{x=0}^{\infty} \{x(x-1)(x-2) + 3x(x-1) + x\} \frac{e^{-\lambda} \lambda^x}{x!}$$

$$= \sum_{x=0}^{\infty} x(x-1)(x-2) \frac{e^{-\lambda} \lambda^x}{x!} + 3 \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda} \lambda^x}{x!} + \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!}$$

$$= e^{-\lambda}\lambda^3\left[\sum_{x=3}^{\infty}\frac{\lambda^{x-3}}{(x-3)!}\right] + 3e^{-\lambda}\lambda^2\left[\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!}\right] + \lambda$$

$$= e^{-\lambda}\lambda^3 e^{\lambda} + 3e^{-\lambda}\lambda^2 e^{\lambda} + \lambda = \lambda^3 + 3\lambda^2 + \lambda$$

$$\mu_4' = E(X^4) = \sum_{x=0}^{\infty} x^4 \cdot p(x;\lambda)$$

$$= \sum_{x=0}^{\infty}\{x(x-1)(x-2)(x-3) + 6x(x-1)(x-2) + 7x(x-1) + x\}\frac{e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda}\lambda^4\left[\sum_{x=4}^{\infty}\frac{\lambda^{x-4}}{(x-4)!}\right] + 6e^{-\lambda}\lambda^3\left[\sum_{x=3}^{\infty}\frac{\lambda^{x-3}}{(x-3)!}\right]$$

$$+ 7e^{-\lambda}\lambda^2\left[\sum_{x=2}^{\infty}\left(\frac{\lambda^{x-2}}{(x-2)!}\right]\right. + \lambda$$

$$= \lambda^4(e^{-\lambda}e^{+\lambda}) + 6\lambda^3(e^{-\lambda}e^{+\lambda}) + 7\lambda^2(e^{-\lambda}e^{+\lambda}) + \lambda = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda$$

The four central moments are now obtained as follows :

$$\mu_2 = \mu_2' - \mu_1'^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda$$

Thus the mean and the variance of the Poisson distribution are each equal to λ.

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 = (\lambda^3 + 3\lambda^2 + \lambda) - 3\lambda(\lambda^2 + \lambda) + 2\lambda^3 = \lambda.$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= (\lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda) - 4\lambda(\lambda^3 + 3\lambda^2 + \lambda) + 6\lambda^2(\lambda^2 + \lambda) - 3\lambda^4 = 3\lambda^2 + \lambda$$

Co-efficients of skewness and kurtosis are given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\lambda^2}{\lambda^3} = \frac{1}{\lambda} \text{ and } \gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{\lambda}}$$

Also $$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{1}{\lambda} \text{ and } \gamma_2 = \beta_2 - 3 = \frac{1}{\lambda} \qquad ...(7\cdot 15)$$

Hence the Poisson distribution is always a skewed distribution.

Proceeding to the limit as $\lambda \to \infty$, we get

$$\beta_1 = 0 \text{ and } \beta_2 = 3$$

7·3·3. Mode of the Poisson Distribution

$$\frac{p(x)}{p(x-1)} = \frac{\dfrac{e^{-\lambda}\lambda^x}{x!}}{\dfrac{e^{-\lambda}\lambda^{x-1}}{(x-1)!}} = \frac{\lambda}{x} \qquad ..(7\cdot 16)$$

We discuss the following cases :

Case I. *When λ is not an integer.*

Let us suppose that S is the intergral part of λ.

$$\frac{p(1)}{p(0)} > 1, \quad \ldots, \frac{p(S-1)}{p(S-2)} > 1, \frac{p(S)}{p(S-1)} > 1,$$

and $$\frac{p(S+1)}{p(S)} < 1, \frac{p(S+2)}{p(S+1)} < 1, \ldots$$

Combining the above expressions into a single expression, we get

$p(0) < p(1) < p(2) \ldots < p(S-2) < p(S-1) < p(S) > p(S+1) > p(S+2) > \ldots$, which shows that $p(S)$ is the maximum value. Hence in this case the distribution is unimodal and the integral part of λ is the unique modal value.

Case II. *When $\lambda = k$ (say) is an integer.* Here we have

$$\frac{p(1)}{p(0)} > 1, \frac{p(2)}{p(1)} > 1, \ldots, \frac{p(k-1)}{p(k-2)} > 1$$

and $$\frac{p(k)}{p(k-1)} = 1, \frac{p(k+1)}{p(k)} < 1, \frac{p(k+2)}{p(k+1)} < 1, \ldots$$

$\therefore\ p(0) < p(1) < p(2) < \ldots < p(k-2) < p(k-1) = p(k) > p(k+1) > p(k+2) \ldots$

In this case we have two maximum values, *viz.*, $p(k-1)$ and $p(k)$ and thus the distribution is bimodal and two modes are at $(k-1)$ and k, *i.e.*, at $(\lambda - 1)$ and λ, (since $k = \lambda$).

7·3·4. Recurrence Relation for the Moments of the Poisson Distribution. By def.,

$$\mu_r = E\{X - E(X)\}^r = \sum_{x=0}^{\infty} (x-\lambda)^r\, p(x,\lambda)$$

$$= \sum_{x=0}^{\infty} (x-\lambda)^r \frac{e^{-\lambda}\lambda^x}{x!}$$

Differentiating with respect to λ, we get

$$\frac{d\mu_r}{d\lambda} = \sum_{x=0}^{\infty} r(x-\lambda)^{r-1}(-1)\frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty} \frac{(x-\lambda)^r}{x!}\left\{x\lambda^{x-1}e^{-\lambda} - \lambda^x e^{-\lambda}\right\}$$

$$= -r\sum_{x=0}^{\infty} (x-\lambda)^{r-1}\cdot\frac{e^{-\lambda}\lambda^x}{x!} + \sum_{x=0}^{\infty} \frac{(x-\lambda)^r}{x!}\left\{\lambda^{x-1}e^{-\lambda}(x-\lambda)\right\}$$

$$= -r\sum_{x=0}^{\infty} (x-\lambda)^{r-1}\frac{e^{-\lambda}\lambda^x}{x!} + \frac{1}{\lambda}\sum_{x=0}^{\infty} (x-\lambda)^{r+1}\cdot\frac{e^{-\lambda}\lambda^x}{x!}$$

$$\therefore \quad \frac{d\mu_r}{d\lambda} = -r\mu_{r-1} + \frac{1}{\lambda}\mu_{r+1}$$

$$\Rightarrow \quad \mu_{r+1} = r\lambda\mu_{r-1} + \lambda\frac{d\mu_r}{d\lambda} \qquad \ldots(7\cdot17)$$

Putting $r = 1, 2$ and 3 successively, we get

$$\mu_2 = r\mu_0 + \lambda\frac{d\mu_1}{d\lambda} = \lambda \qquad (\because \mu_0 = 1,\ \mu_1 = 0)$$

$$\mu_3 = 2\lambda\mu_1 + \lambda\frac{d\mu_2}{d\lambda} = \lambda,\ \mu_4 = 3\lambda\mu_2 + \lambda\frac{d\mu_3}{d\lambda} = 3\lambda^2 + \lambda$$

7·3·5. Moment Generating Function of the Poisson Distribution

$$M_X(t) = \sum_{x=0}^{\infty} e^{tx}\cdot\frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=0}^{\infty}\frac{e^{-\lambda}(\lambda e^t)^x}{x!}$$

$$= e^{-\lambda}\left\{1 + \lambda e^t + \frac{(\lambda e^t)^2}{2!} + \ldots\right\} = e^{-\lambda}\cdot e^{\lambda e^t} = e^{\lambda(e^t-1)} \qquad \ldots(7\cdot 18)$$

7·3·6. Characteristic Function of the Poisson Distribution

$$\phi_X(t) = \sum_{x=0}^{\infty} e^{itx}\cdot p(x) = \sum_{x=0}^{\infty} e^{itx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^{it})^x}{x!}$$

$$= e^{-\lambda}\, e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)} \qquad \ldots(7\cdot 19)$$

7·3·7. Cumulants of the Poisson Distribution

$$K_X(t) = \log M_X(t) = \log\left[e^{\lambda(e^t-1)}\right] = \lambda(e^t - 1)$$

$$= \lambda\left[\left(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots + \frac{t^r}{r!} + \ldots\right) - 1\right]$$

$$= \lambda\left[t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots + \frac{t^r}{r!} + \ldots\right]$$

κ_r = rth cumulant = co-efficient of $\frac{t^r}{r!}$ in $K_X(t) = \lambda$

$$\Rightarrow \qquad \kappa_r = \lambda;\ r = 1, 2, 3, \ldots \qquad \ldots(7\cdot 19a)$$

Hence all the cumulants of the Poisson distribution are equal, each being equal to λ. In particular, we have

$$\text{Mean} = \kappa_1 = \lambda,\ \mu_2 = \kappa_2 = \lambda,\ \mu_3 = \kappa_3 = \lambda \text{ and } \mu_4 = \kappa_4 + 3\kappa_2^2 = \lambda + 3\lambda^2$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\lambda^2}{\lambda^3} = \frac{1}{\lambda} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\lambda + 3\lambda^2}{\lambda^2} = \frac{1}{\lambda} + 3$$

Remark. If m is the mean and σ is the s.d. of Poisson distribution with parameter λ, then

$$m\sigma\gamma_1\gamma_2 = \lambda\cdot\sqrt{\lambda}\cdot\sqrt{\beta_1}\,(\beta_2 - 3)$$

$$= \lambda\cdot\sqrt{\lambda}\cdot\frac{1}{\sqrt{\lambda}}\cdot\frac{1}{\lambda} = 1.$$

7·3·8. Additive or Reproductive Property of Independent Poisson Variates. *Sum of independent Poisson variates is also a Poisson variate. More elaborately, if X_i, $(i = 1, 2, \ldots, n)$ are independent Poisson variates with param-*

ters λ_i; $i = 1, 2, ..., n$ *respectively, then* $\sum_{i=1}^{n} X_i$ *is also a Poisson variate with parameter* $\sum_{i=1}^{n} \lambda_i$.

Proof. $M_{X_i}(t) = e^{\lambda_i (e^t - 1)}$; $i = 1, 2, ..., n$

$$M_{X_1 + X_2 + ... + X_n}(t) = M_{X_1}(t)\, M_{X_2}(t) \ldots M_{X_n}(t),$$

[since X_i; $i = 1, 2, ..., n$ are independent]

$$= e^{\lambda_1 (e^t - 1)}\, e^{\lambda_2 (e^t - 1)} \ldots e^{\lambda_n (e^t - 1)}$$

$$= e^{(\lambda_1 + \lambda_2 + ... + \lambda_n)(e^t - 1)}$$

which is the m.g.f. of a Poisson variate with parameter $\lambda_1 + \lambda_2 + \ldots + \lambda_n$. Hence by uniqueness theorem of m.g.f.'s, $\sum_{i=1}^{n} X_i$ is also a Poisson variate with parameter $\sum_{i=1}^{n} \lambda_i$.

Remarks 1. In fact, the converse of the above result is also true *i.e., If* $X_1, X_2, ..., X_n$ *are independent and* $\sum_{i=1}^{n} X_i$ *has a Poisson distribution, then each of the random variables* $X_1, X_2, ..., X_n$ *has a Poisson distribution.*

Let X_1 and X_2 be independent *r.v.'s* so that $X_1 \sim P(\lambda_1)$ and $X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$. Then we want to prove that $X_2 \sim P(\lambda_2)$.

Proof. Since X_1 and X_2 are independent, we have

$$M_{X_1 + X_2}(t) = M_{X_1}(t)\, M_{X_2}(t)$$

$$\Rightarrow \quad e^{(\lambda_1 + \lambda_2)(e^t - 1)} = e^{\lambda_1 (e^t - 1)} \cdot M_{X_2}(t)$$

$$\Rightarrow \quad M_{X_2}(t) = e^{\lambda_2 (e^t - 1)}$$

$\Rightarrow$ $X_2 \sim P(\lambda_2)$, by uniqueness theorem of m.g.f.

2. The difference of two independent Poisson variates is not a Poisson variate.

$$M_{X_1 - X_2}(t) = M_{X_1 + (-X_2)}(t) = M_{X_1}(t) \cdot M_{(-X_2)}(t),$$

(since X_1 and X_2 are independent).

$$\therefore \quad M_{X_1 - X_2}(t) = M_{X_1}(t)\, M_{X_2}(-t) \qquad [\because M_{cX}(t) = M_X(ct)]$$

$$= e^{\lambda_1 (e^t - 1)} \cdot e^{\lambda_2 (e^{-t} - 1)} = e^{\lambda_1 (e^t - 1) + \lambda_2 (e^{-t} - 1)}$$

which cannot be put in the form $e^{\lambda (e^t - 1)}$. Hence $(X_1 - X_2)$ is not a Poisson variate.

Moreover the difference $(X_1 - X_2)$ cannot be a Poisson variate is evident from the fact that it may have positive as well as negative values, while a Poisson variate is always non-negative.

7·3·9. Probability Generating Function of Poisson Distribution

$$\text{P.G.F. of } X = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \cdot s^k = \sum_{k=0}^{\infty} e^{-\lambda} \frac{(\lambda s)^k}{k!} = e^{-\lambda} e^{\lambda s} = e^{\lambda(s-1)}$$

...(7·20)

Example 7·24. *A car hire firm has two cars which it fires out day by day. The number of demands for a car on each day is distributed as Poisson variate with mean 1·5. Calculate the proportion of days on which (i) neither car is used, and (ii) some demand is refused.* **[Meerut Univ. B.Sc. 1993]**

Solution. The proportion of days on which there are x demands for a car

$$= P\{\text{of } x \text{ demands in a day}\}$$

$$= \frac{e^{-1\cdot5}(1\cdot5)^x}{x!},$$

since the number of demands for a car on any day is a Poisson variate with mean 1·5. Thus

$$P(X = x) = \frac{e^{-1\cdot5}(1\cdot5)^x}{x!}; \quad x = 0, 1, 2, \ldots$$

(*i*) Proportion of days on which neither car is used is given by

$$P(X = 0) = e^{-1\cdot5}$$

$$= \left[1 - 1\cdot5 + \frac{(1\cdot5)^2}{2!} - \frac{(1\cdot5)^3}{3!} + \frac{(1\cdot5)^4}{4!} - \ldots\right]$$

$$= 0\cdot2231$$

(*ii*) Proportion of days on which some demand is refused is

$$P(X > 2) = 1 - P(X \leq 2)$$

$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$

$$= 1 - e^{-1\cdot5}\left[1 + 1\cdot5 + \frac{(1\cdot5)^2}{2!}\right]$$

$$= 1 - 0\cdot2231 \times 3\cdot625 = 0\cdot19126$$

Example 7·25. *A manufacturer of cotter pins knows that 5% of his product is defective. If he sells cotter pins in boxes of 100 and guarantees that not more than 10 pins will be defective, what is the approximate probability that a box will fail to meet the guaranteed quality ?* **[Kanpur Univ. B.Sc. 1993]**

Solution. We are given $n = 100$.

Let p = Probability of a defective pin = 5% = 0·05

$\therefore$ λ = Mean number of defective pins in a box of 100

$= np = 100 \times 0\cdot05 = 5$

Since 'p' is small, we may use Poisson distribution.

Probability of x defective pins in a box of 100 is

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-5}5^x}{x!};\ x = 0, 1, 2, \ldots$$

Probability that a box will fail to meet the guaranteed quality is

$$P(X > 10) = 1 - P(X \leq 10) = 1 - \sum_{x=0}^{10} \frac{e^{-5}5^x}{x!} = 1 - e^{-5} \sum_{x=0}^{10} \frac{5^x}{x!}$$

Example 7·26. *Six coins are tossed 6,400 times. Using the Poisson distribution, find the approximate probability of getting six heads r times.*

Solution. The probability of obtaining six heads in one throw of six coins (*a* single trial), is $p = (1/2)^6$, assuming that head and tail are equally probable.

$$\therefore \qquad \lambda = np = 6400 \times (1/2)^6 = 100.$$

Hence, using Poisson probability law, the required probability of getting 6 heads r times is given by :

$$P(X = r) = \frac{e^{-\lambda} \cdot \lambda^r}{r!} = \frac{e^{-100} \cdot (100)^r}{r!};\ r = 0, 1, 2, \ldots$$

Example 7·27. *In a book of 520 pages, 390 typo-graphical errors occur. Assuming Poisson law for the number of errors per page, find the probability that a random sample of 5 pages will contain no error.*

[Patna Univ. B.Sc. (Hons.), 1988]

Solution. The average number of typographical errors per page in the book is given by $\lambda = (390/520) = 0{\cdot}75$

Hence using Poisson probability law, the probability of x errors per page is given by : $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0{\cdot}75}(0{\cdot}75)^x}{x!};\ x = 0, 1, 2, \ldots$

The required probability that a random sample of 5 pages will contain no error is given by : $[P(X = 0)]^5 = \left(e^{-0{\cdot}75}\right)^5 = e^{-3{\cdot}75}$

Example 7·28. *Suppose that the number of telephone calls coming into a telephone exchange between 10 A.M. and 11 A.M. say, X_1 is a random variable with Poisson distribution with parameter 2. Similarly the number of calls arriving between 11 A.M. and 12 noon say, X_2 has a Poisson distribution with parameter 6. If X_1 and X_2 are independent, what is the probability that more than 5 calls come in between 10 A.M. and 12 noon ?* **[Calicut U. B. Sc. Oct. 1992]**

Solution. Let $X = X_1 + X_2$. By the additive property of Poisson distribution, X is also a Poisson variate with parameter (say) $\lambda = 2 + 6 = 8$

Hence the probability of x calls in-between 10 A.M. and 12 noon is given by $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-8}8^x}{x!};\ x = 0, 1, 2, \ldots$

Probability that more than 5 calls come in between 10 A.M. and 12 noon is given by

$$P(X > 5) = 1 - P(X \leq 5) = 1 - \sum_{x=0}^{5} \frac{e^{-8} 8^x}{x!}$$
$$= 1 - 0{\cdot}1912 = 0{\cdot}8088$$

Example 7·29. *A Poisson distribution has a double mode at $x = 1$ and $x = 2$. What is the probability that x will have one or the other of these two values?*

Solution. We have proved that if the Poisson distribution is bimodal, then the two modes are at the points $x = \lambda - 1$ and $x = \lambda$. Since we are given that the two modes are at the points $x = 1$ and $x = 2$, we find that $\lambda = 2$.

$$\therefore \quad P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-2}2^x}{x!}; \; x = 0, 1, 2, \ldots$$

$$\Rightarrow \quad P(X = 1) = e^{-2}2$$

and $$P(X = 2) = \frac{e^{-2} \cdot 2^2}{2!} = e^{-2} \cdot 2$$

Required probability $= P(X = 1) + P(X = 2) = 2e^{-2} + 2e^{-2} = 0{\cdot}542$

Example 7·30. *If X is a Poisson variate such that*

$$P(X = 2) = 9P(X = 4) + 90P(X = 6) \qquad \ldots(*)$$

Find (i) λ, the mean of X, (ii) β, the coefficient of skewness.

[Delhi Univ. B. Sc. (Maths. Hons.) 1992, '87]

Solution. If X is a Poisson variate with parameter λ, then

$$P(X = x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}, \; x = 0, 1, 2, \ldots; \lambda > 0$$

Hence (*) gives

$$\frac{e^{-\lambda} \cdot \lambda^2}{2!} = e^{-\lambda}\left[9\frac{\lambda^4}{4!} + 90\frac{\lambda^6}{6!}\right]$$
$$= \frac{e^{-\lambda}\lambda^2}{8}[3\lambda^2 + \lambda^4]$$

$$\Rightarrow \quad \lambda^4 + 3\lambda^2 - 4 = 0$$

Solving as a quadratic in λ^2, we get

$$\lambda^2 = \frac{-3 \pm \sqrt{9 + 16}}{2} = \frac{-3 \pm 5}{2}$$

Since $\lambda > 0$, we get $\lambda^2 = 1 \quad \Rightarrow \quad \lambda = 1$

Hence mean $= \lambda = 1$, and $\mu_2 =$ Variance $= \lambda = 1$

Also $\beta_1 =$ Coefficient of skewness $= \frac{1}{\lambda} = 1.$

Example 7·31. *If X and Y are independent Poisson variates such that*

$$P(X = 1) = P(X = 2)$$

and $$P(Y = 2) = P(Y = 3) \qquad \ldots(*)$$

Find the varaince of $X - 2Y$.

Solution. Let $X \sim P(\lambda)$ and $Y \sim P(\mu)$.

Then we have

$$P(X = x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}, \; x = 0, 1, 2, \ldots ; \lambda > 0$$

and $$P(Y = y) = \frac{e^{-\mu} \cdot \mu^y}{y!}, \; y = 0, 1, 2, \ldots ; \; \mu > 0$$

Using (*), we get

$$\left.\begin{aligned} \lambda e^{-\lambda} &= \frac{\lambda^2 e^{-\lambda}}{2!} \\ \text{and} \quad \frac{\mu^2 e^{-\mu}}{2} &= \frac{\mu^3 e^{-\mu}}{3!} \end{aligned}\right\} \quad \ldots(**)$$

Solving (**), we get

$$\lambda e^{-\lambda} [\lambda - 2] = 0 \text{ and } \mu^2 e^{-\mu} [\mu - 3] = 0$$

$$\Rightarrow \quad \lambda = 2 \text{ and } \mu = 3, \text{ since } \lambda > 0, \; \mu > 0.$$

Now $$\text{Var}(X) = \lambda = 2, \text{ and } \text{Var}(Y) = \mu = 3 \quad \ldots(***)$$

$$\therefore \quad \text{Var}(X - 2Y) = 1^2 \text{ Var}(X) + (-2)^2 \cdot \text{Var}\, Y,$$

covariance term vanishes since X and Y are independent.

Hence, on using (***), we get

$$\text{Var}(X - 2Y) = 2 + 4 \times 3 = 14$$

Example 7·32. *If* X *and* Y *are independent Poisson variates with means* λ_1 *and* λ_2 *respectively, find the probability that*

(*i*) $X + Y = k$, (*ii*) $X = Y$ **[Delhi Univ. B. Sc. (Stat. Hons.), 1991]**

Solution. We have

$$P(X = x) = \frac{e^{-\lambda_1} \cdot \lambda_1^x}{x!}, \; x = 0, 1, 2, 3, \ldots ; \lambda_1 > 0$$

and $$P(Y = y) = \frac{e^{-\lambda_2} \cdot \lambda_2^y}{y!}, \; y = 0, 1, 2, 3 \ldots ; \; \lambda_2 > 0$$

$$(i) \quad P(X + Y = k) = \sum_{r=0}^{k} P(X = r \cap Y = k - r)$$

$$= \sum_{r=0}^{k} P(X = r) \, P(Y = k - r)$$

$$[\because X \text{ and } Y \text{ are independent}]$$

$$= \sum_{r=0}^{k} \frac{e^{-\lambda_1} \lambda_1^r}{r!} \cdot \frac{e^{-\lambda_2} \cdot \lambda_2^{k-r}}{(k - r)!}$$

$$= e^{-(\lambda_1 + \lambda_2)} \sum_{r=0}^{k} \frac{\lambda_1^r \cdot \lambda_2^{k-r}}{r! \, (k - r)!}$$

$$= e^{-(\lambda_1+\lambda_2)}\left[\frac{\lambda_2^k}{k!} + \frac{\lambda_1 \cdot \lambda_2^{k-1}}{1!\,(k-1)!} + \frac{\lambda_1^2 \cdot \lambda_2^{k-2}}{2!\,(k-2)!} + \ldots + \frac{\lambda_1^k}{k!}\right]$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\left[\lambda_2^k + {}^kC_1\,\lambda_2^{k-1}\cdot\lambda_1 + {}^kC_2\cdot\lambda_2^{k-2}\cdot\lambda_1^2 + \ldots + \lambda_1^k\right]$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \times (\lambda_1+\lambda_2)^k,\ k = 0, 1, 2, \ldots$$

which is the probability function of Poisson distribution with parameter $\lambda_1 + \lambda_2$.

Aliter. Since $X \sim P(\lambda_1)$ and $Y \sim P(\lambda_2)$ are independent, by the additive property of Poisson distribution $X + Y \sim P(\lambda_1 + \lambda_2)$. Hence

$$P(X+Y=k) = \frac{e^{-(\lambda_1+\lambda_2)} \times (\lambda_1+\lambda_2)^k}{k!};\ k = 0, 1, 2, \ldots$$

(ii) $$P(X=Y) = \sum_{r=0}^{\infty} P(X=r \cap Y=r)$$

$$= \sum_{r=0}^{\infty} P(X=r)\,P(Y=r)$$

[$\because X$ and Y are independent]

$$= e^{-(\lambda_1+\lambda_2)} \sum_{r=0}^{\infty} \frac{(\lambda_1\lambda_2)^r}{(r!)^2}$$

Example 7·33. *Show that in a Poisson distribution with unit mean, mean deviation about mean is (2/e) times the standard deviation.*

[Patna Univ. B. Sc. (Stat. Hons.) 1992; Delhi Univ. B.Sc. (Stat. Hons.), 1993]

Solution. Here we are given $\lambda = 1$.

$$\therefore \quad P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-1}\cdot 1}{x!} = \frac{e^{-1}}{x!};\ x = 0, 1, 2, \ldots$$

Mean deviation about mean 1 is

$$E(|X-1|) = \sum_{x=0}^{\infty} |x-1|\,p(x) = e^{-1} \sum_{x=0}^{\infty} \frac{|x-1|}{x!}$$

$$= e^{-1}\left[1 + \frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \ldots\right]$$

We have $$\frac{n}{(n+1)!} = \frac{(n+1)-1}{(n+1)!} = \frac{1}{n!} - \frac{1}{(n+1)!}$$

$\therefore$ Mean deviation about mean

$$= e^{-1}\left[1 + \left(1 - \frac{1}{2!}\right) + \left(\frac{1}{2!} - \frac{1}{3!}\right) + \left(\frac{1}{3!} - \frac{1}{4!}\right) + \ldots\right]$$

$$= e^{-1}(1 + 1) = \frac{2}{e} \times 1 = \frac{2}{e} \times \text{standard deviation},$$

since for the Poisson distribution, variance = mean = 1 (given).

Example 7·34. *Let $X_1, X_2, \ldots, X_n$ be identically and independently distributed Bin (1, p) variates. Let $S_n = \sum_{j=1}^{n} X_j$ and $M_n(t)$ be the m.g.f. of S_n. Find* $\lim_{n \to \infty} M_n(t)$, *using $np = \lambda$ (const.)* **[Delhi Univ. B. Sc. (Maths Hons.), 1989]**

Solution. Since X_i, $i = 1, 2, \ldots n$ are i.i.d. binomial variates $B(1, p)$,

$S_n = \sum_{j=1}^{n} X_j$, is a binomial $B(n, p)$ variate.

$$\therefore M_n(t) = M.g.f. \text{ of } S_n = \left(q + pe^t\right)^n = \left[1 + (e^t - 1)p\right]^n$$

If we take $np = \lambda \Rightarrow p = \lambda/n$ and let $n \to \infty$, we get

$$\lim_{n \to \infty} M_n(t) = \lim_{n \to \infty} \left[1 + \frac{(e^t - 1)\lambda}{n}\right]^n = \exp\left[\lambda(e^t - 1)\right],$$

which is the m.g.f. of Poisson distribution with parameter λ. Hence by uniqueness theorem of m.g.f., $S_n = \sum_{j=1}^{n} X_j \to P(\lambda)$, as $n \to \infty$, with $np = \lambda$ (fixed).

Example. 7·35. *(a) If X is a Poisson variate with mean m, show that the expectation of e^{-kX} is* $\exp\left[-m(1 - e^{-k})\right]$. **[Nagpur Univ. B.Sc. 1993]**

Hence show that, if $\bar{X}$ is the arithmetic mean of n independent random variables $X_1, X_2, \ldots, X_n$, each having Poisson distribution with parameter m, then $e^{-\bar{x}}$ as an estimate of e^{-m} is biased, although $\bar{X}$ is an unbaised estimate of m.

(b) If X is a Poisson variate with mean m, what would be the expectation of $e^{-kx}\,kX$, k being a constant.

Solution.

$$E(e^{-kX}) = \sum_{x=0}^{\infty} e^{-kx} p(x) = \sum_{x=0}^{\infty} e^{-kx} \cdot \frac{e^{-m} m^x}{x!} = e^{-m} \sum_{x=0}^{\infty} \frac{(me^{-k})^x}{x!}$$

$$= e^{-m}\left[1 + me^{-k} + \frac{(me^{-k})^2}{2!} + \ldots\right]$$

$$= e^{-m}\, e^{me^{-k}} = e^{-m(1-e^{-k})} \qquad ...(*)$$

We have

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$$

Since X_i; $i = 1, 2, ..., n$ is a Poisson variate with parameter m, $E(X_i) = m$.

$$\therefore\ E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} m = \frac{1}{n}\, nm = m$$

Hence $\bar{X}$ is an unbaised estimate of m.

Now
$$E(e^{-\bar{X}}) = E\left[\exp\left(\frac{-1}{n}\sum_{i=1}^{n} X_i\right)\right]$$
$$= E\left(e^{-X_1/n} \cdot e^{-X_2/n} \cdots e^{-X_n/n}\right),$$
$$= E\left(e^{-X_1/n}\right) E\left(e^{-X_2/n}\right) \ldots E\left(e^{-X_n/n}\right),$$
(since $X_1, X_2, ..., X_n$ are independent)

$$\therefore\qquad E(e^{-\bar{X}}) = \prod_{i=1}^{n} E\left(e^{-X_i/n}\right) \qquad ...(**)$$

Using (*) with $k = 1/n$, we get

$E(e^{-X_i/n}) = e^{-m(1-e^{-1/n})}$, (since X_i is a Poisson variate with parameter m)

$$\therefore\ E(e^{-\bar{X}}) = \prod_{i=1}^{n}\left[\exp\left\{-m(1-e^{-1/n})\right\} = \exp\left\{-m(1-e^{-1/n})\right\}\right]^{n}$$
$$= \exp\left\{-mn(1-e^{-1/n})\right\} \neq e^{-m}$$

Hence $e^{-\bar{X}}$ is not an unbaised estimated of e^{-m}, though $\bar{X}$ is an unbiased estimate of m.

(*b*) $$E(e^{-kX}\, kX) = \sum_{x=0}^{\infty} e^{-kx}\, kx \cdot p(x) = k\sum_{x=1}^{\infty} e^{-kx}\, x\, \frac{e^{-m} m^x}{x!}$$

$$= ke^{-m}\sum_{x=1}^{\infty}\frac{(me^{-k})^x}{(x-1)!} = ke^{-m}me^{-k}\sum_{x=1}^{\infty}\frac{(me^{-k})^{x-1}}{(x-1)!}$$

$$= mke^{-m-k}\left\{1 + me^{-k} + \frac{(me^{-k})^2}{2!} + \ldots\right\},$$

$$= mke^{-m-k}\cdot e^{me^{-k}} = mk\exp\left[\left\{m(e^{-k}-1)\right\} - k\right]$$

Example 7·36. *If X and Y are independent Poisson variates with means* m_1 and m_2 *respectively, prove that the probability that* $X - Y$ *has the value 'r' is the co-efficient of* t^r *in*

$$\exp\left\{m_1 t + m_2 t^{-1} - m_1 - m_2\right\}$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1991, '89]

Solution. Since X and Y are independent Poisson variates with means m_1 and m_2 respectively,

$$\left.\begin{aligned} & P(X = x) = \frac{e^{-m_1} m_1^x}{x!};\ x = 0, 1, 2, \ldots \infty \\ \text{and} \quad & P(Y = y) = \frac{e^{-m_2} m_2^y}{y!};\ y = 0, 1, 2, \ldots \infty \end{aligned}\right\} \quad \ldots(1)$$

$$P(X - Y = r) = \sum_{s=0}^{\infty} P(X = r + s \cap Y = s) = \sum_{s=0}^{\infty} P(X = r + s)\, P(Y = s)$$

$$= \sum_{s=0}^{\infty} \frac{e^{-m_1} \cdot m_1^{r+s}}{(r+s)!} \cdot \frac{e^{-m_2} m_2^s}{s!} \quad \ldots[\text{From (1)}]$$

$$= e^{-m_1 - m_2} \sum_{s=0}^{\infty} \frac{m_1^{r+s} m_2^s}{(r+s)!\, s!} \quad \ldots(2)$$

We have $e^{m_1 t + m_2 t^{-1}} = e^{m_1 t} \times e^{m_2 t^{-1}}$

$$= \left\{1 + m_1 t + \frac{(m_1 t)^2}{2!} + \ldots + \frac{(m_1 t)^{r+s}}{(r+s)!} + \ldots\right\}$$

$$\times \left\{1 + m_2 t^{-1} + \frac{(m_2 t^{-1})^2}{2!} + \ldots + \frac{(m_2 t^{-1})^s}{s!} + \ldots\right\}$$

$\therefore$ Co-efficient of t^r in $e^{m_1 t + m_2 t^{-1}} = \sum_{s=0}^{\infty} \frac{m_1^{r+s} m_2^s}{(r+s)!\, s!}$

Hence from (2), we get

$$P(X - Y = r) = e^{-m_1 - m_2} \times \text{Coefficient of } t^r \text{ in } e^{m_1 t + m_2 t^{-1}},$$

$$= \text{Coefficient of } t^r \text{ in } e^{-m_1 - m_2 + m_1 t + m_2 t^{-1}}$$

which is the required result.

Example 7·37. *If X is a Poisson variate with mean m, show that $\frac{X - m}{\sqrt{m}}$ is a variable with mean zero and variance unity. Find the M.G.F. for this variable and show that it approaches $e^{t^2/2}$ as $m \to \infty$. Also interpret the result.*

[Delhi Univ. B. Sc. (Stat. Hons.), 1987]

Solution. Let $Y = \frac{X - m}{\sqrt{m}}$

$$\therefore \quad E(Y) = E\left(\frac{X - m}{\sqrt{m}}\right) = \frac{1}{\sqrt{m}} E(X - m) = 0$$

$$V(Y) = E\left(\frac{X-m}{\sqrt{m}}\right)^2 = \frac{1}{m}E(X-m)^2 = \frac{1}{m}\mu_2 = 1$$

$$\text{M.G.F. of } Y = M_Y(t) = E\left(e^{tY}\right) = E\left[e^{t(X-m)/\sqrt{m}}\right]$$

$$= e^{-t\sqrt{m}}\left[E\left(e^{tX/\sqrt{m}}\right)\right]$$

$$= e^{-t\sqrt{m}}\sum_{x=0}^{\infty}\frac{e^{-m}m^x}{x!}\cdot e^{tx/\sqrt{m}}$$

$$= e^{-t\sqrt{m}}\cdot e^{-m}\sum_{x=0}^{\infty}\frac{\left(me^{t/\sqrt{m}}\right)^x}{x!}$$

$$= e^{-m-t\sqrt{m}}\left[1+\frac{me^{t/\sqrt{m}}}{1!}+\frac{\left(me^{t/\sqrt{m}}\right)^2}{2!}+\ldots\right]$$

$$= e^{-m-t\sqrt{m}}\cdot\exp\left(me^{t/\sqrt{m}}\right) = \exp\left[-m-t\sqrt{m}+me^{t/\sqrt{m}}\right]$$

$$= \exp\left[-m-t\sqrt{m}+m\left(1+\frac{t}{\sqrt{m}}+\frac{t^2}{2!m}+\frac{t^3}{3!m^{3/2}}+\ldots\right)\right]$$

$$= \exp\left[\frac{1}{2}t^2+\frac{1}{3!\sqrt{m}}t^2+\ldots\right]$$

Now proceeding to limit as $m \to \infty$, we get

$$\lim_{m\to\infty} M_Y(t) = e^{t^2/2} \qquad \ldots(*)$$

Interpretation. (*) is the m.g.f. of Standard Normal Variate [c.f. Remark to § 8·2·5]. Hence by uniqueness theorem of m.g.f.'s, standard Poisson variate tends to standard normal variate as $m \to \infty$. Hence Poisson distribution tends to Normal distribution for large values of parameter m.

Example 7·38. *Deduce the first four moments about the mean of the Poisson distribution from those of the Binomial distribution.*

Solution. The first four central moments of the binomial distribution are

$$\left.\begin{array}{ll} \mu_1 = 0, & \text{Mean} = np \\ \mu_2 = npq, & \mu_3 = npq(q-p) \text{ and} \\ \mu_4 = npq(1-6pq) + 3n^2p^2q^2 & \end{array}\right\} \qquad \ldots(*)$$

Poisson distribution is a limiting form of the binomial distribution under the following conditions :

(*i*) $n \to \infty$, (*ii*) $p \to 0$, *i.e.*, $q \to 1$, and (*iii*) $np = \lambda$, (say), is finite.

Using these conditions, we get from (*) the moments of the Poisson distribution as

$$\mu_1 = 0$$

$$\text{Mean} = \lim(np) = \lambda$$

$$\mu_2 = \lim(npq) = \lim(np)\cdot\lim(q) = \lambda\cdot 1 = \lambda$$

$$\mu_3 = \lim[npq(q-p)] = \lambda\cdot 1\,(1-0) = \lambda$$

$$\mu_4 = \lim[npq(1-6pq) + 3(np)^2q^2]$$

$$= [\lambda \cdot 1 (1 - 6 \cdot 0 \cdot 1) + 3\lambda^2 \cdot 1] = \lambda + 3\lambda^2$$

Example 7·39. *If X is a Poisson variate with parameter m and Y is another discrete variable whose conditional distribution for a given X is given by*

$$P(Y = r | X = x) = \binom{x}{r} p^r (1 - p)^{x-r}; \; 0 < p < 1, \; r = 0, 1, 2, \ldots, x$$

then show that the unconditional distribution of Y is a Poisson distribution with parameter mp.

[Delhi Univ. B.Sc. (Stat. Hons.), 1993, Shivaji U.B.Sc. Nov. 1992]

Solution. We are given that

$$P(X = x) = \frac{e^{-m} m^x}{x!}; \; x = 0, 1, 2, \ldots .$$

and $$P(Y = r | X = x) = \binom{x}{r} p^r (1 - p)^{x-r}; r \leq x$$

$$\therefore \quad P(X = x \cap Y = r) = P(X = x) P(Y = r | X = x)$$

$$= \frac{e^{-m} m^x}{x!} \binom{x}{r} p^r (1 - p)^{x-r}$$

$\therefore$ $P(Y = r)$ = The unconditional distribution of Y.

$$= \sum_{x=r}^{\infty} \left[\frac{e^{-m} m^x}{x!} \cdot \binom{x}{r} p^r (1 - p)^{x-r} \right]$$

$$= e^{-m} \left[\sum_{x=r}^{\infty} \binom{x}{r} \frac{p^r m^x (1 - p)^{x-r}}{x!} \right]$$

$$= e^{-m} \left[\sum_{x=r}^{\infty} \frac{m^x}{x!} \cdot \frac{x!}{r! (x - r)!} p^r (1 - p)^{x-r} \right]$$

$$= \frac{e^{-m}}{r!} \left[\sum_{x=r}^{\infty} \frac{m^x}{(x - r)!} p^r (1 - p)^{x-r} \right]$$

$$= \frac{e^{-m} (mp)^r}{r!} \left[\sum_{x=r}^{\infty} \frac{m^{x-r} (1 - p)^{x-r}}{(x - r)!} \right]$$

$$= \frac{e^{-m} (mp)^r}{r!} \left[\sum_{x=r}^{\infty} \frac{\{m(1 - p)\}^{x-r}}{(x - r)!} \right]$$

$$= \frac{e^{-m} (mp)^r}{r!} e^{m(1-p)} = \frac{e^{-mp} (mp)^r}{r!}; \; r = 0, 1, 2, \ldots$$

Hence Y is a Poisson variate with parameter mp.

Example 7·40. *If X and Y are independent Poisson variates, show that the conditional distribution of X given $X + Y$, is binomial.*

[Madras Univ. B.Sc. Main 1992; Delhi Univ. B. Sc. (Maths Hons.), 1988]

Solution. Let X and Y be independent Poisson variates with parameters λ and μ respectively. Then $X + Y$ is also a Poisson variate with parameter $\lambda + \mu$.

$$P[X = r | (X + Y = n)] = \frac{P(X = r \cap X + Y = n)}{P(X + Y = n)} = \frac{P(X = r \cap Y = n - r)}{P(X + Y = n)}$$

$$= \frac{P(X = r) P(Y = n - r)}{P(X + Y = n)} \quad \text{[since } X \text{ and } Y \text{ are indepdent]}$$

$$\therefore P[X = r | (X + Y = n)] = \frac{\dfrac{e^{-\lambda}\lambda^r}{r!} \cdot \dfrac{e^{-\mu}\mu^{n-r}}{(n-r)!}}{\dfrac{e^{-(\lambda+\mu)}(\lambda+\mu)^n}{n!}}$$

$$= \frac{n!}{r!(n-r)!}\left(\frac{\lambda}{\lambda+\mu}\right)^r \left(\frac{\mu}{\lambda+\mu}\right)^{n-r}$$

$$= \binom{n}{r} p^r q^{n-r}, \text{ where } p = \frac{\lambda}{\lambda+\mu}, q = 1 - p$$

Hence the conditional distribution of X given $X + Y = n$, is a binomial distribution with parameters n and $p = \lambda/(\lambda + \mu)$.

Example 7·41. *If X is a Poisson variate with parameter m and μ_r is the rth central moment, prove that*

$$m[{}^rC_1 \mu_{r-1} + {}^rC_2 \mu_{r-2} + \ldots + {}^rC_r \mu_0] = \mu_{r+1}.$$

[Delhi Univ. B.Sc. (Stat. Hons.) 1990]

Solution Since $X \sim P(m)$, its probability function is given by

$$p(x) = \frac{e^{-m} \cdot m^x}{x!}, \; x = 0, 1, 2, \ldots; m > 0$$

By definition,

$$\mu_{r+1} = E[X - E(X)]^{r+1} = E[X - m]^{r+1}$$

$$= \sum_{x=0}^{\infty} (x - m)^{r+1} p(x)$$

$$= \sum_{x=0}^{\infty} (x - m)^r (x - m) \frac{e^{-m} \cdot m^x}{x!}$$

$$= \sum_{x=0}^{\infty} \frac{x(x - m)^r e^{-m} m^x}{x!} - m \sum_{x=0}^{\infty} (x - m)^r \cdot \frac{e^{-m} m^x}{x!}$$

$$= \sum_{x=1}^{\infty} \frac{(x-m)^r e^{-m} m^x}{(x-1)!} - m\mu_r$$

$$= \sum_{y=0}^{\infty} \frac{(y-m+1)^r \cdot e^{-m} \cdot m^{y+1}}{y!} - m\mu_r, \qquad (x-1=y)$$

$$= m \cdot \sum_{y=0}^{\infty} (y-m+1)^r \cdot p(y) - m\mu_r$$

$$= m \sum_{y=0}^{\infty} \left[(y-m)^r + {}^rC_1 (y-m)^{r-1} + {}^rC_2 (y-m)^{r-2} \right.$$

$$\left. + \ldots + {}^rC_{r-1}(y-m) + 1 \right] p(y) - m\mu_r$$

$$= m[\mu_r + {}^rC_1 \mu_{r-1} + {}^rC_2 \mu_{r-2} + \ldots + {}^rC_r \mu_0] - m\mu_r$$

$$= m[{}^rC_1 \mu_{r-1} + {}^rC_2 \mu_{r-2} + \ldots + {}^rC_r \mu_0].$$

Example 7·42. *If X has a Poisson distribution with parameter λ, show that the distribution function of X is given by*

$$F(x) = \frac{1}{\Gamma(x+1)} \int_{\lambda}^{\infty} e^{-t} t^x \, dt; \; x = 0, 1, 2, \ldots$$

[Delhi Univ. M. Sc. (Stat) 1986]

Solution. If X is a Poisson variate, then

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}; \; x = 0, 1, 2, \ldots \qquad (*)$$

Consider the incomplete gamma integral;

$$I_x = \frac{1}{x!} \int_{\lambda}^{\infty} e^{-t} t^x \, dt; \quad (x \text{ is a positive integer})$$

$$= \left| -\frac{e^{-t} t^x}{x!} \right|_{\lambda}^{\infty} + \frac{1}{(x-1)!} \int_{\lambda}^{\infty} e^{-t} t^{x-1} \, dt$$

$$= \frac{e^{-\lambda} \lambda^x}{x!} + I_{x-1} \qquad (**)$$

which is a reduction formula for I_x.

Repeated applications of (**) gives

$$I_x = \frac{e^{-\lambda} \lambda^x}{x!} + \frac{e^{-\lambda} \lambda^{x-1}}{(x-1)!} + \ldots + \frac{e^{-\lambda} \lambda}{1!} + I_0$$

But $I_0 = \int_{\lambda}^{\infty} e^{-t} \, dt = \left| -e^{-t} \right|_{\lambda}^{\infty} = e^{-\lambda}$

$$\therefore \quad I_x = e^{-\lambda} + \lambda e^{-\lambda} + \frac{\lambda^2 e^{-\lambda}}{2!} + \ldots + \frac{\lambda^x}{x!} e^{-\lambda}$$

$$= P(X=0) + P(X=1) + \ldots + P(X=x) \quad \text{[From (*)]}$$

$$= P(X \leq x) = F(x)$$

where $F(\cdot)$ is the distribution function of the *r.v.* X.

$$\Rightarrow \qquad F(x) = \frac{1}{x!}\int_{\lambda}^{\infty} e^{-t}t^{x}\,dt = \frac{1}{\Gamma(x+1)}\int_{\lambda}^{\infty} e^{-t}t^{x}\,dt$$

$(\because \Gamma(x+1) = x!$, since x is a positive integer.)

Remark. This result is of great practical utility. It enables us to represent the cumulative Poisson probabilities (which are generally tedious to compute numerically) in terms of incomplete gamma integral, the values of which are tabulated for different values of λ by Karl Pearson in his Tables of Incomplete Γ-Functions.

7·3·10. Recurrence Formula for the Probabilities of Poisson Distribution. *(Fitting of Poisson Distribution).* For a Poisson distribution with parameter λ, we have

$$p(x) = \frac{e^{-\lambda}\cdot\lambda^{x}}{x!};\ x = 0, 1, 2, ..., \infty$$

and $$P(x+1) = \frac{e^{-\lambda}\lambda^{x+1}}{(x+1)!};\ x = 0, 1, 2, ..., \infty$$

$$\therefore \qquad \frac{p(x+1)}{p(x)} = \frac{\lambda}{(x+1)} \Rightarrow p(x+1) = \frac{\lambda}{x+1}p(x) \quad ...(17·20)$$

which is the required recurrence formula.

This formula provides us a very convenient method of graduating the given data by a Poisson distribution. The only probability we need to calculate is $p(0)$ which is given by $p(0) = e^{-\lambda}$, where λ is estimated from the given data. The other probabilities, *viz.*, $p(1), p(2)$.... can now be easily obtained as explained below:

$$p(1) = [p(x+1)]_{x=0} = \left[\frac{\lambda}{x+1}\right]_{x=0} p(0),$$

$$p(2) = [p(x+1)]_{x=1} = \left[\frac{\lambda}{x+1}\right]_{x=1} p(1),$$

$$p(3) = [p(x+1)]_{x=2} = \left[\frac{\lambda}{x+1}\right]_{x=2} p(2),$$

and so on.

Example 7·43. *After correcting 50 pages of the proof of a book, the proof reader finds that there are, on the average, 2 errors per 5 pages. How many pages would one expect to find with 0, 1, 2, 3 and 4 errors, in 1000 pages of the first print of the book ? (Given that* $e^{-0·4} = 0·6703$*)*

[Nagpur Univ. M.A. (Eco.), 1989]

Solution. Let the random variable X denote the number of errors per page. Then the mean number of errors per page is given by :

$$\lambda = 2/5 = 0{\cdot}4$$

Using Poisson probability law, probability of x errors per page is given by:

$$P(X = x) = p(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0{\cdot}4}(0{\cdot}4)^x}{x!}; x = 0, 1, 2, \ldots$$

Expected number of pages with x errors per page in a book of 1000 pages are :

$$1000 \times P(X = x) = 1000 \times \frac{e^{-0{\cdot}4}(0{\cdot}4)^x}{x!}; x = 0, 1, 2, \ldots$$

Using the recurrence formula (17·20), various probabilities can be easily calculated as shown in the following table.

No. of errors per page (X)	*Probability* $p(x)$	*Expected number of pages* $1000\,p(x)$
0	$p(0) = e^{-0{\cdot}4} = 0{\cdot}6703$	$670{\cdot}3 \simeq 670$
1	$p(1) = \frac{0{\cdot}4}{0+1}\,p(0) = 0{\cdot}26812$	$268{\cdot}12 \simeq 268$
2	$p(2) = \frac{0{\cdot}4}{1+1}\,p(1) = 0{\cdot}053624$	$53{\cdot}624 \simeq 54$
3	$p(3) = \frac{0{\cdot}4}{2+1}\,p(2) = 0{\cdot}0071298$	$7{\cdot}1298 \simeq 7$
4	$p(4) = \frac{0{\cdot}4}{3+1}\,p(3) = 0{\cdot}00071298$	$0{\cdot}71298 \simeq 1$

Example 7·44. *Fit a Poisson distribution to the following data which gives the number of doddens in a sample of clover seeds.*

No. of doddens: (x)	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Observed frequency: (f)	*56*	*156*	*132*	*92*	*37*	*22*	*4*	*0*	*1*

Solution.

$$\text{Mean} = \frac{1}{N}\sum fx = \frac{986}{500} = 1{\cdot}972$$

Taking the mean of the given distribution as the mean of the Poisson distribution we want to fit, we get $\lambda = 1{\cdot}972$,

and $$p(x) = \frac{e^{-\lambda}\cdot\lambda^x}{x!}; x = 0, 1, 2, \ldots, \infty$$

$$p(0) = e^{-\lambda} = e^{-1{\cdot}972}$$

$\therefore \log_{10} p(0) = -1{\cdot}972 \log_{10} e = -1{\cdot}972 \times 0{\cdot}43429$

$= -0{\cdot}856419 = \bar{1}{\cdot}143581$

$\therefore \quad p(0) = 0{\cdot}1392$

Using the recurrence formula (17·20) the various probabilities, *viz.*, $p(1), p(2),...$, can be easily calculated as shown in the following table :

x	$\frac{\lambda}{x+1}$	$p(x)$	*Expected frequency* $N.p(x)$
0	1·972	0·13920	69·6000
1	0·986	0·27455	137·2512
2	0·657	0·27006	135·3296
3	0·493	0·17793	88·9566
4	0·394	0·10964	43·8556
5	0·328	0·03459	17·2966
6	0·281	0·01137	5·6846
7	0·247	0·00320	1·6013
8	0·219	0·00078	0·3942

Since frequencies are always integers, therefore by converting them to nearest integers, we get

Observed frequency :	56	156	132	92	37	22	4	0	1
Expected frequency :	70	137	135	89	44	17	6	2	0

Remark. In rounding the figures to the nearest integer it has to be kept in mind that the total of the observed and the expected frequencies should be same.

EXERCISE 7 (b)

1. (*a*) Derrive Poisson distribution as a limiting form of a binomial distribution. **[Madras Univ. B. E., Dec. 1991]**

Hence find β_1 and β_2 of the distribution.

Give some examples of the occurrence of Poisson distribution in different fields.

(*b*) State and prove the reproductive property of the Poisson distribution. Show that the mean and variance of the Poisson distribution are equal.

Find the mode of the Poisson distribution with mean value 5.

(*c*) Prove that under certain conditions to be stated by you, the number of telephone calls on a trunkline in a given interval of time has a Poisson distribution. **[Calcutta Univ. B.Sc. (Maths Hons.), 1989]**

(*d*) Show that for a Poisson distribution, the coefficient of variation is the reciprocal of the standard deviation.

2. (*a*) If two independent variables X_1 and X_2 have Poisson distribution with means λ_1 and λ_2 respectively, then show that their sum $X_1 + X_2$ is a Poisson variate with mean $\lambda_1 + \lambda_2$.

Does the difference of two independent Poisson variates follow a Poisson distribution ? Give reasons. **[Sri Venketeswara Univ. B.Sc., 1991]**

(*b*) Prove that the sum of two independent Poisson variates is a Poisson variate. Is the result true for the difference also ? Give reasons.

[Delhi Univ. B.Sc. (Stat. Hons.) 1989]

(*c*) If $X_1, X_2, ..., X_k$ are independent random variables following the Poisson law with parameter $m_1, m_2, ..., m_k$ respectively, show that $\sum_{i=1}^{k} X_i$ follows the Poisson law with parameter $\sum_{i=1}^{k} m_i$ **[Madras Univ. B. E., 1993]**

3. (*a*) Prove the recurrence relation between the moments of Poisson distribution

$$\mu_{r+1} = \lambda \left(r \mu_{r-1} + \frac{d \mu_r}{d \lambda} \right), \text{ where } \mu_r = \sum_{j=0}^{\infty} \frac{e^{-\lambda} \lambda^j}{j!} (j - \lambda)^r$$

where μ_r is the rth moment about the mean λ. Hence obtain the skewness and kurtosis of Poisson distribution.

[Delhi Univ. B. Sc. (Stat. Hons.) 1989,' 86; Utkal Univ. B. Sc. 1993]

(*b*) Let X have a Poisson distribution with parameter $\lambda > 0$. If r is a non-negative integer and if $\mu_r' = E(X^r)$, prove that

$$\mu'_{r+1} = \lambda \left(\mu'_r + \frac{d \mu'_r}{d \lambda} \right)$$

[Madras Univ. B. Sc. Nov. 1988]

4. What do you understand by (*i*) cumulants, (*ii*) cumulative function. Obtain the cumulative function of a Poisson distribution with parameter λ. Hence or otherwise show that for a Poisson distribution with parameter λ, all the cumulants are λ.

5. For the Poisson distribution with parameter λ, show that the rth factorial moment $\mu'_{(r)}$ is given by $\mu'_{(r)} = \lambda^r$

Show further that $\mu_{(2)} = \lambda$, $\mu_{(3)} = -2\lambda$ and $\mu_{(4)} = 3\lambda(\lambda + 2)$

6. (*a*) If X and Y are independent *r.v. s.*' so that $X \sim P(\lambda)$ and $X + Y \sim P(\lambda + \mu)$, find the distribution of Y. **[Ans. $Y \sim P(\mu)$]**

(*b*) If $X \sim P(\lambda)$, find

(*i*) Karl Pearson's coefficient of skewness

(*ii*) Moment measure of skewness.

Is Poisson distribution positively skewed or negatively skewed ?

7. (*a*) It is known that the probability that an item produced by a certain machine will be defective is 0·01. By applying Poisson's approximation, show that the probability that random sample of 100 items selected at random from the total output will contain no more than one defective item is $2/e$.

(*b*) The probability of success in a trail is known to be 10^{-4}. It is possible to repeat the trial independently any desired number of times. Do you think that the number of successes in a series of trials, if the number of trials in the series increases indefinitely, will tend to follow a Poisson distribution ? Give your reasons.

(*c*) The probability of getting no misprint in a page of a book is e^{-1}. What is the probability that a page contains more than 2 misprints ? [State the assumptions you make in solving this problem.] **[Bombay Univ. B.Sc., 1989]**

8. In a certain factory turning out optical lenses, there is a small chance 1/500 for any lens to be defective. The lenses are supplied in a packet of 10. Use Poisson distribution to calculate the approximate number of packets containing no defective, one defective, two defective and three defective lenses in a consignment of 20,000 packets.

Ans. 19604, 392, 4 and 0 packets.

9. Red blood cell deficiency may be determined by examining a specimen of the blood under a microscope. Suppose a certain small fixed volume contains on the average 20 red cells for normal persons. Using Poisson distribution, obtain the probability that a specimen from a normal person will contain less than 15 red cells.

Ans. $\sum_{x=0}^{14} \{e^{-20}\,(20)^x/x\,!\}$

10. Assuming that the chance of a traffic accident in a day in a street of Delhi is 0·001, on how many days out of a trial of 1,000 days can we expect :

(*i*) no accident

(*ii*) more than three accidents, if there are 1,000 such streets in the whole city ?

11. Patients arrive randomly and independently at a doctor's surgery from 8·0 A.M. at an average rate of one in five minutes.The waiting room holds 2 persons. What is the probability that the room will be full when the doctor arrives at 9·0 A.M. (Estimate the probability to an accuracy of 5 per cent.)

Ans. 53·84 %

12. An office switchboard receives telephone calls at the rate of 3 calls per minute on an average. What is the probability of receiving (*i*) no calls in a one-minute interval, (*ii*) at the most 3 calls in a 5-minute interval ?

Ans. (*i*) 0·0323, (*ii*) 0

13. A hospital switchboard receives an average of 4 emergency calls in a 10-minute interval. What is the probability that (*i*) there are at the most 2

emergency calls in a 10-minute interval, (*ii*) there are exactly 3 emergency calls in a 10-minute interval ?

Ans. (*i*) 13^{-4}, (*ii*) $(32/3)\,e^{-4}$

14. (*a*) A distributor of bean seeds determines from extensive tests that 5% of large batch of seeds will not germinate. He sells the seeds in packets of 200 and guarantees 90% germination. Determine the probability that a particular packet will violate the guarantee.

Ans. $1 - \sum_{r=0}^{10} (e^{-10}\, 10^r / r!)$

(*b*) In an automatic telephone exchange the probability that any one call is wrongly connected is 0·001. What is the minimum number of independent calls required to ensure a probability of 0·90, that at least one call is wrongly connected?

15. (*a*) Fit a Poisson distribution to the following data with respect to the number of red blood corpuscles (x) per cell :

x:	0	1	2	3	4	5
Number of cells f:	142	156	69	27	5	1

(*b*) Data was collected over a period of 10 years, showing number of deaths from horse kicks in each of the 20 army corps. From the 200 corps-years, the distribution of deaths was as follows :

No. of deaths :	0	1	2	3	4
Frequency :	122	60	15	2	1

Graduate the data by Poisson distribution and calculate the theoretical frequencies.

Given	e^{-m}:	0·6703	0·6065	0·5488	0·4966
	m:	0·4	0·5	0·6	0·7

(*c*) Fit a Poisson distribution to the following data and calculate the expected frequencies :-

x:	0	1	2	3	4	5	6	7	8
f:	71	112	117	57	27	11	3	1	1

16. (*a*) If X is the number of occurrences of the Poisson variate with mean λ ; show that : $P(X \geq n) - P(X \geq n+1) = P(X = n)$

(*b*) Suppose that X has a Poisson distribution. If

$$P(X = 2) = \tfrac{2}{3}\, P(X = 1).$$

Evaluate (*i*) $P(X = 0)$ and (*ii*) $P(X = 3)$ [Ans. (*i*) 0·264.]

(*c*) If X has a Poisson distribution such that

$P(X = 1) = P(X = 2)$, find $P(X = 4)$. [Ans 0·09]

(*c*) If a Poisson variate X is such that

$$P(X = 1) = 2\,P(X = 2),$$

find $P\,(X = 0)$, mean and the variance.

(*d*) If for a Poisson variate X, $E\,(X^2) = 6$, what is $E\,(X)$?

(*e*) If X and Y are independent Poisson variates having means 1 and 3 respectively, find the variance of $3X + Y$.

17. Show that for a Poisson distribution

(*i*) $M\sigma\,\gamma_1\,\gamma_2 = 1$, (*ii*) $\beta_1^{1/2}\,(\beta_2 - 3)\,\mu_1'\sigma = 1$

18. Show that the function which generates the central moments of the Poisson distribution with parameter λ is

$$M(t) = \exp\{\lambda(e^t - 1 - t)\}$$

Show that it satisfies the equation

$$\frac{dM(t)}{dt} = \lambda t\,M(t) + \lambda\,\frac{dM(t)}{d\lambda}$$

19. (*a*) The random variable X has p.d.f.

$$f(x) = e^{-\theta}\,\frac{\theta^x}{x!};\ x = 0, 1, 2,\ldots$$

$$= 0, \text{ elsewhere}$$

Find the m.g.f. of $Y = 2X - 1$ and Var (Y).

(*b*) Identify the distribution with the following mgf's :

$$M_X(t) = (0{\cdot}3 + 0{\cdot}7\,e^t)^{10}$$

$$M_Y(t) = \exp\,[3(e^t - 1)]$$

Ans. $X \sim B\,(10, 0{\cdot}7)$, $Y \sim P(3)$.

20. If X has Poisson distribution with parameter λ, then

$$P\,[X \text{ is even}] = \frac{1}{2}\left[1 + e^{-2\lambda}\right]$$

[Delhi Univ. B. Sc. (Stat. Hons.) 1991]

21. (*a*) The m.g.f. of a r.v. is X is $\exp[4(e^t - 1)]$. Show that $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0{\cdot}931$

Hint. $X \sim P(\lambda = 4)$;

Required Probability. $= P(0 < X < 8) = P(1 \le X \le 7) = 0{\cdot}931$

(*b*) If $X \sim P(\lambda = 100)$, use Chebychev's inequality to determine a lower bound for $P\,(75 < X < 125)$ [**Ans.** 0·84]

22. If $X \sim P(m)$, show that $E\,|X - 1| = m - 1 + 2e^{-m}$

[Delhi Univ. B. Sc. (Maths. Hons.), 1983]

Hint. $E\,|X-1| = \sum_{x=0}^{\infty} |x-1|\,e^{-m}\,m^x/x! = e^{-m} + \sum_{x=2}^{\infty}\frac{(x-1)}{x!}\cdot e^{-m}\,m^x$

$$= e^{-m} + e^{-m}\cdot\sum_{x=2}^{\infty} m^x\left[\frac{1}{(x-1)!} - \frac{1}{x!}\right]$$

23. If $X \sim P(\lambda)$ and $Y|X = x \sim (B(x, p)$, then prove that $Y \sim P(\lambda p)$.

24. If the chances of 0, 1, 2, 3... events from one source are given by a Poisson distribution of mean m_1 and the chances of 0, 1, 2, 3,... events from another source by a Poisson distribution of mean m_2, show that the chances of 0, 1, 2, 3,... events from either source are given by

$$e^{-(m_1+m_2)}\left\{1, (m_1 + m_2), \frac{(m_1 + m_2)^2}{2!}, \ldots\right\}.$$

Show that the sum of any finite number of Poisson variates is itself a Poisson variate with mean equal to the sum of separate means.

25. X is a Poisson variate with mean λ.

Show that $E(X^2) = \lambda E(X + 1)$.

If $\lambda = 1$, show that $E|X - 1| = \frac{2}{e}$

26. Show that the mean deviation about mean for Poisson distribution

$$p(x) = \frac{e^{-m} m^x}{x!}; \; x = 0, 1, 2, \ldots$$

is $(2\mu) \cdot \frac{e^{-m} \cdot m^{\mu}}{\mu!}$

where μ is the greatest integer contained in $(m + 1)$.

[Delhi Univ. B. Sc. (Stat. Hons.), 1988,' 84]

27. Let X, Y be independent Poisson variates. The variance of $X + Y$ is 9 and

$$P(X = 3 | X + Y = 6) = 5/54$$

Find the mean of X. [Ans. $\frac{1}{2}(9 \pm 3\sqrt{3})$ *i.e.* 1·902 *or* 7·098]

28. If X is a Poisson variate with paramter m, show that

$$P(X < r) < \frac{m^r}{r!}; \; r = 0, 1, 2, \ldots$$

Deduce that $E(X) < e^m$. **[Delhi Univ. B.Sc. (Maths. Hons.), 1989]**

29. (*a*) The characteristic function of a variate X is

$$\varphi_X(t) = \left(\frac{1}{3} + \frac{2}{3}e^{it}\right)^6 \cdot \left[\exp\left\{-3(1 - e^{it})\right\}\right].$$

Recognise the variate.

[Burdwan Univ. B. Sc. (Maths. Hons.) 1989)

Hint. $X = U + V$, where $U \sim B\left(6, \frac{2}{3}\right)$ and $V \sim P(3)$ are independent *r.v.'s*

(*b*) Identify the variates X and Y where :

$$M_X(t) = (1/27)(1 + 2e^t)^3 \cdot \exp\left[3(e^t - 1)\right]$$

$$M_Y(t) = (1/32)\,(1 + e^t)^5 \cdot \exp\left[-2\,(1 - e^t)\right]$$

[Delhi Univ. B. Sc. (Stat. Hons.), 1987, 84]

Ans. $X = U + V$; $U \sim B\,(n = 3, p = 2/3)$ and $V \sim P\,(\lambda = 3)$ are independent.

$Y = U_1 + V_1$; $U_1 \sim B\,(n = 5, p = 1/2)$ and $V_1 \sim P\,(\lambda = 2)$ are independent.

30. If X and Y are correlated variates each having Poisson distribution, show that $X + Y$ cannot be a Poisson variate

[Delhi Univ. B. Sc. (Maths Hons.), 1988; Poona Univ. B.Sc., 1989]

Hint. Note that for Poisson variate mean and variance are equal. Let $X \sim P(\lambda)$, $Y \sim P(\mu)$; (X, Y) correlated.

$$\therefore\ E(X + Y) = E(X) + E(Y) = \lambda + \mu$$

$$\text{Var}\,(X + Y) = \text{Var}\,X + \text{Var}\,Y + 2\,\text{Cov}\,(X, Y)$$

$$= \lambda + \mu + 2\rho\sqrt{\lambda\mu}\ ,\ (\rho \neq 0)$$

Since $E(X + Y) \neq \text{Var}\,(X + Y)$; $X + Y$ cannot be a Poisson variate.

31. Let X, Y, Z be independent Poisson variates with parameters a, b and c respectively. Obtain :

(*i*) m.g.f. of $X + 2Y + 3Z$,

(*ii*) Conditional expectation of X given $X + Y + Z = n$

(Indian Civil Services, 1985)

Hint. $M_{X + 2Y + 3Z}(t) = M_X(t) \cdot M_Y(2t) \cdot M_Z(3t)$

$$= \exp\left[a\left(e^t - 1\right) + b\left(e^{2t} - 1\right) + c\left(e^{3t} - 1\right)\right]$$

$$P(X = x \mid X + Y + Z = n) = \frac{P(X = x \cap X + Y + Z = n)}{P(X + Y + Z = n)}$$

$$= \frac{P(X = x)\,P(Y + Z = n - x)}{P(X + Y + Z = n)} \quad (\because X, Y, Z \text{ are indep.})$$

$$= \frac{e^{-a} \cdot a^x}{x!} \times \frac{e^{-(b+c)} \cdot (b + c)^{n-x}}{(n - x)!}$$

$$\times \left[\frac{n!}{e^{-(a+b+c)} \cdot (a + b + c)^n}\right]$$

$$= \frac{n!}{x!\,(n - x)!}\left(\frac{a}{a + b + c}\right)^x \cdot \left(\frac{b + c}{a + b + c}\right)^n$$

$$\Rightarrow \quad X \mid (X + Y + Z = n) \sim B\left\{n,\ p = a/(a + b + c)\right\}$$

$$\Rightarrow \quad E\left[X \mid X + Y + Z = n\right] = np = \frac{na}{a + b + c}$$

32. The joint density of r.v.'s X and Y is :

$$f(x, y) = e^{-2} / \left[x!\,(y - x)!\right];\ y = 0, 1, 2, \ldots ;\ x = 0, 1, 2, \ldots, y.$$

Find the m.g.f. $M(t_1, t_2)$ of (X, Y) and correlation coefficient between X and Y. Show that the marginal distributions of X and Y are Poisson.

Hint. $M(t_1, t_2) = \sum_{y=0}^{\infty} \sum_{x=0}^{y} e^{t_1 x + t_2 y} \times \left[\frac{e^{-2}}{x!(y-x)!} \right]$

$$= e^{-2} \sum_{y=0}^{\infty} \left[\frac{e^{t_2 y}}{y!} \left\{ \sum_{x=0}^{y} {}^{y}C_x \cdot (e^{t_1})^x \right\} \right]$$

$$= e^{-2} \sum_{y=0}^{\infty} \left\{ \left[e^{t_2} \left(1 + e^{t_1} \right) \right]^y / y! \right\}$$

$$= e^{-2} \cdot \exp \left\{ e^{t_2} (1 + e^{t_1}) \right\}$$

$$M(t_1, 0) = \exp \left[2(e^{t_1} - 1) \right] \Rightarrow X \sim P(\lambda = 1)$$

$$M(0, t_2) = \exp \left[2(e^{t_2} - 1) \right] \Rightarrow Y \sim P(\mu = 2)$$

Observe $M(t_1, t_2) \neq M(t_1, 0) \times M(0, t_2) \Rightarrow X$ and Y are not independent.

$E(X) = 1$, $\text{Var}(X) = 1$; $E(Y) = 2 = \text{Var}\, Y$.

$$E(XY) = \left| \frac{\partial^2 M(t_1, t_2)}{\partial t_1 \, \partial t_2} \right|_{t_1 = t_2 = 0} = 3$$

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{3 - 1 \times 2}{1 \times \sqrt{2}} = 1/\sqrt{2}.$$

33. The joint p.g.f. of the r.v.'s X and Y is given by :

$P(s_1, s_2) = \exp \left[a(s_1 - 1) + b(s_2 - 1) + c(s_1 - 1)(s_2 - 1) \right]$, a,b,c, are all positive. Find $\rho(X, Y)$

Hint. $P_X(s_1) = P(s_1, 1) = \exp \left[a(s_1 - 1) \right] \Rightarrow X \sim P(a)$

$$P_Y(s_2) = P(1, s_2) = \exp \left[b(s_2 - 1) \right] \Rightarrow Y \sim P(b)$$

$$E(XY) = \left(\frac{\partial^2 P(s_1, s_2)}{\partial s_1 \, \partial s_2} \right)_{s_1 = s_2 = 1} = c + ab.$$

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{(c + ab) - ab}{\sqrt{a}\sqrt{b}} = \frac{c}{\sqrt{ab}}$$

34. An insurance company issues only two types of policy, household and motor. It has carried out an investigation into the experience of a group of policyholders who held one of each type of policy over a particular period and it has discovered that within that group and over that period the mean number of claims per household policy was 0·3 and the mean number of claims per motor policy was 0·8. Assume that the number of claims under each type of policy is independent of the number of claims under the other type of policy and that each can be represented by a Poisson distribution.

(*a*) If the number of claims per policyholder is the sum of the number of claims under each of his two policies, state with reasons how the number of claims per policyholder, within that group and over that period is distributed, and

(*b*) Calculate to the nearest whole number, the percentage of policyholders within that group and over that period who made more household claims than motor claims.

Hint. Household claim, $X \sim P(\cdot 3)$ and Motor claim, $Y \sim P(\cdot 8)$

$$\text{Required Probability} = P(X > Y) = \sum_{r=0}^{\infty}\left[\sum_{s=0}^{\infty} P(Y = r \cap X = r+s)\right]$$

$$= \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\left[P(Y=r)P(X=r+s)\right] = \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{e^{-\cdot 8}(\cdot 8)^r}{r!}\times\frac{e^{-\cdot 3}(\cdot 3)^{r+s}}{(r+s)!}$$

$$= e^{-\cdot 8}e^{-\cdot 3}\sum_{r=0}^{\infty}\left[\frac{(\cdot 8)^r}{r!}\sum_{s=0}^{\infty}\left\{\frac{(\cdot 3)^{r+s}}{(r+s)!}\right\}\right]$$

$$= \sum_{r=0}^{\infty}\left[\frac{e^{-\cdot 8}(\cdot 8)^r}{r!}e^{-\cdot 3}\left\{e^{\cdot 3} - \left(1 + \cdot 3 + \frac{(\cdot 3)^2}{2!} + \ldots + \frac{(\cdot 3)^{r-1}}{(r-1)!}\right)\right\}\right]$$

$$= 1 - e^{-\cdot 8}e^{-\cdot 3}\left[\left\{\frac{\cdot 8}{1} + \frac{(\cdot 8)^2}{2!}(1+\cdot 3) + \frac{(\cdot 8)^3}{3!}\left(1 + \cdot 3 + \frac{\cdot 09}{2}\right)\right.\right.$$

$$\left.\left. + \frac{(\cdot 8)^4}{4!}\left(1 + \cdot 3 + \frac{\cdot 09}{2} + \frac{\cdot 027}{3!}\right) + \ldots\right]\right.$$

35. (*i*) An event occurs instantaneously and is equally likely to occur at any instant. There is no limit on the number of occurrneces that may happen in any interval of time, but the expected number in a given time interval is T. Prove that the probability of the event occurring exactly r times in an interval of the same duration is $(T^r e^{-T})/r!$.

(*ii*) An insurance company which writes only fire and accident business defines a major claim as one which costs at least Rs. 50,000 for an accident claim or Rs. 100,000 for a fire claim. Any excess over these amounts is paid by reinsurers and hence every major claim is recorded at a cost of Rs. 50,000 or Rs. 100,000 respectively. The company divides the year into equal monthly accounting periods and a report is produced of the recorded cost of major claims. The expected number of major accident claims is 0·2 per month and of major fire claims 0·5 per month. Calculate the probability that in a particular month the recorded cost of major claims is Rs. 2,00,000 or more.

36. (*a*) The number of aeroplanes arriving at an airport in a 30 minute interval obeys the Poisson law with mean 25. Use Chebychev's inequality to find the least chance, that the number of planes to arrive within a given 30 minutes interval will be between 15 and 35. **[Sri Venketeswara U. B.Sc. 1992]**

(*b*) Suppose that the number of motor cars arriving in a certain parking lot in any 15 minutes period obeys a Poisson probability law with mean 80. Use Chebychev's inequality to determine a lower bound for the probability that the

CHAPTER EIGHT

Theoretical Continuous Distributions

8·1. Rectangular (or Uniform Distribution. A random variable X is said to have a continuous uniform distribution over an interval (a, b) if its probability density function is constant $= k$ (say), over the entire range of X, *i.e.*,

$$f(x) = \begin{cases} k, a < x < b \\ 0, \text{ otnerwise} \end{cases}$$

Since total probability is always unity, we have

$$\int_a^b f(x)\,dx = 1 \Rightarrow k\int_a^b dx = 1 \quad i.e., \quad k = \frac{1}{b-a}$$

$$\therefore \qquad f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, \text{ otherwise} \end{cases} \qquad ...(8·1)$$

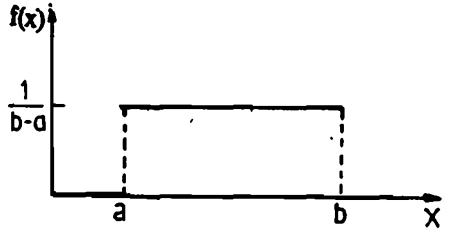
Remarks. 1. a and b, $(a < b)$ are the two parameters of the uniform distribution on (a, b).

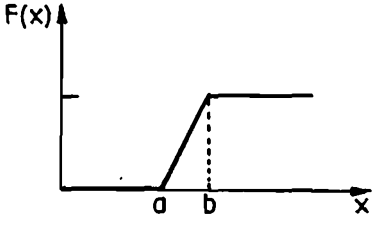
2. The distribution is also known as rectangular distribution, since the curve $y = f(x)$ describes a rectangle over the x-aixs and between the ordinates at $x = a$ and $x = b$.

3. The distribution function $F(x)$ is given by

$$F(x) = \begin{cases} 0, \text{ if } -\infty < x < a \\ \frac{x-a}{b-a}, \; a \le x \le b \\ 1, \quad b < x < \infty \end{cases} \qquad ...(8·1\,a)$$

Since $F(x)$ is not continuous at $x = a$ and $x = b$, it is not differentiable at these points. Thus $\frac{d}{dx} F(x) = f(x) = \frac{1}{b-a} \neq 0$, exists everywhere except at the points $x = a$ and $x = b$. and consequently p.d.f. $f(x)$ is given by (8·1).





4. The graphs of uniform p.d.f. $f(x)$ and the corresponding distribution function $F(x)$ are given on page 8·1 :

5. For a rectangular or uniform variate X in $(-a, a)$, p.d.f. is given by

$$f(x) = \begin{cases} \frac{1}{2a}, -a < x < a \\ 0, \text{ otherwise.} \end{cases}$$

8·1·1. Moments of Rectangular Distribution.

$$\mu'_r = \int_a^b x^r f(x)\,dx = \frac{1}{(b-a)} \int_a^b x^r\,dx = \frac{1}{(b-a)} \left[\frac{b^{r+1} - a^{r+1}}{r+1} \right] \qquad ...(8·2)$$

In particular

$$\text{Mean} = \mu_1' = \frac{1}{(b-a)} \left[\frac{b^2 - a^2}{2} \right] = \frac{b+a}{2}$$

and $$\mu_2' = \frac{1}{(b-a)} \left[\frac{b^3 - a^3}{3} \right] = \frac{1}{3}(b^2 + ab + a^2)$$

$$\therefore \qquad \mu_2 = \mu_2' - \mu_1'^2 = \frac{1}{3}(b^2 + ab + a^2) - \left(\frac{b+a}{2} \right)^2 = \frac{1}{12}(b-a)^2$$

8·1·2. Moment Generating Function is given by

$$M_X(t) = \int_a^b e^{tx} f(x)\,dx = \frac{e^{bt} - e^{at}}{t(b-a)}$$

8·1·3. Characteristic Function is given by

$$\varphi_X(t) = \int_a^b e^{itx} f(x)\,dx = \frac{e^{ibt} - e^{iat}}{it(b-a)}$$

8·1·4. Mean Deviation about Mean, η is given by

$$\eta = E\,|X - \text{Mean}| = \int_a^b |x - \text{Mean}|\, f(x)\,dx$$

$$= \frac{1}{(b-a)} \int_a^b \left| x - \frac{a+b}{2} \right| dx$$

$$= \frac{1}{(b-a)} \int_{-(b-a)/2}^{(b-a)/2} |t|\,dt \qquad \left[t = x - \frac{a+b}{2} \right]$$

$$= \frac{1}{(b-a)} \cdot 2 \int_0^{(b-a)/2} t\,dt = \frac{b-a}{4}$$

Example 8·1 *If X is uniformly distributed with mean 1 and variance 4/3, find $P(X<0)$.* **[Delhi Univ. B.A. (Hons. Spl. Course-Statistics), 1989]**

Solution. Let $X \sim U(a, b)$, so that $p(x) = \frac{1}{b-a}$; $a < x < b$. We are given:

$$\text{Mean} = \frac{b+a}{2} = 1 \quad \Rightarrow \quad b + a = 2$$

$$\text{Var}(X) = \frac{1}{12}(b-a)^2 = \frac{4}{3} \quad \Rightarrow \quad (b-a)^2 = 16 \quad \Rightarrow \quad b - a = \pm 4$$

Solving, we get : $a = -1$ and $b = 3$; $(a < b)$.

$$\therefore \qquad p(x) = \frac{1}{4} \; ; \; -1 < x < 3$$

$$P(X<0) = \int_{-1}^{0} p(x)\,dx = \frac{1}{4} \mid x \mid_{-1}^{0} = \frac{1}{4}$$

Example 8·2. *Subway trains on a certain line run every half hour between mid-night and six in the morning. What is the probability that a man entering the station at a random time during this period will have to wait at least twenty minutes?*

Solution. Let the *r.v.* X denote the waiting time (in minutes) for the next train. Under the assumption that a man arrives at the station at random, X is distributed uniformly on (0, 30), with *p.d.f.*,

$$f(x) = \begin{cases} \frac{1}{30}, & 0 < x < 30 \\ 0, & \text{otherwise} \end{cases}$$

The probability that he has to wait at least 20 minutes is

$$P(X \geq 20) = \int_{20}^{30} f(x)\,dx = \frac{1}{30}\int_{20}^{30} 1\,dx = \frac{1}{30}(30-20) = \frac{1}{3}$$

Example 8·3. *If X has a uniform distribution in [0, 1], find the distribution (p.d.f.) of $-2 \log X$. Identify the distribution also.*

[Delhi Univ. B.Sc. (Stat. Hons.), 1989, '86]

Solution. Let $Y = -2 \log X$. Then the distribution function G of Y is

$$G_Y(y) = P(Y \leq y) = P(-2 \log X \leq y)$$

$$= P(\log X \geq -y/2) = P(X \geq e^{-y/2}) = 1 - P(X \leq e^{-y/2})$$

$$= 1 - \int_0^{e^{-y/2}} f(x)\,dx = 1 - \int_0^{e^{-y/2}} 1 \cdot dx = 1 - e^{-y/2}$$

$$g_Y(y) = \frac{d}{dy} G(y) = \frac{1}{2} e^{-y/2}, \; 0 < y < \infty \qquad \ldots(*)$$

[$\because$ as X ranges in (0, 1), $Y = -2 \log X$ ranges from 0 to ∞]

Remark. This example illustrates that if $X \sim U[0, 1]$, then $Y = -2 \log X$, has an exponential distribution with parameter $\theta = \frac{1}{2}$. [*c.f.* § 8·6] or $Y = -2 \log X$ has chi-square distribution with $n = 2$ degrees of freedom [*c.f.* Chapter 13, § 13·2].

Example 8·4. *Show that for the rectangular distribution :*

$$f(x) = \frac{1}{2a}, -a < x < a$$

the m.g.f. about origin is $\frac{1}{at}(\sinh at)$. *Also show that moments of even order are given by* $\mu_{2n} = \frac{a^{2n}}{(2n+1)}$

Solution. M.G.F. about origin is given by

$$M_X(t) = E(e^{tx}) = \int_{-a}^{a} e^{tx} f(x)\, dx = \frac{1}{2a}\int_{-a}^{a} e^{tx}\, dx$$

$$= \frac{1}{2a}\left| \frac{e^{tx}}{t} \right|_{-a}^{a} = \frac{1}{2at}(e^{at} - e^{-at}) = \frac{\sinh at}{at}$$

$$= \frac{1}{at}\left[at + \frac{(at)^3}{3!} + \frac{(at)^5}{5!} + \ldots \right] = 1 + \frac{a^2 t^2}{3!} + \frac{a^4 t^4}{5!} + \ldots$$

Since there are no terms with odd powers of t in $M(t)$, all moments of odd order about origin vanish, *i.e.*,

$$\mu'_{2n+1} \text{ (about origin)} = 0$$

In particular μ_1' (about origin) $= 0$, *i.e.*, mean $= 0$

Thus μ_r' (about origin) $= \mu_r$ (since mean is origin)

Hence $\mu_{2n+1} = 0 ; n = 0, 1, 2, \ldots$

i.e., all moments of odd order about mean vanish. The moments of even order are given by

$$\mu_{2n} = \text{coefficient of } \frac{t^{2n}}{(2n)!} \text{ in } M(t) = \frac{a^{2n}}{(2n+1)}$$

Example 8·5. *If X_1 and X_2 are independent rectangular variates on [0, 1], find the distributions of*

(*i*) X_1/X_2, (*ii*) $X_1 X_2$, (*iii*) $X_1 + X_2$, and (*iv*) $X_1 - X_2$

Solution. We are given

$$f_{X_1}(x_1) = f_{X_2}(x_2) = 1 ; 0 < x_1 < 1, 0 < x_2 < 1$$

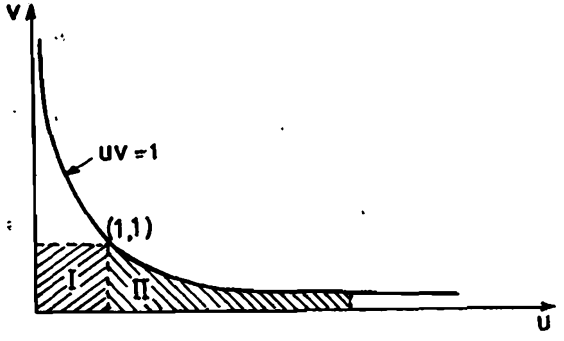
Since X_1 and X_2 are independent, their joint p.d.f. is

$$f(x_1, x_2) = f(x_1) f(x_2) = 1$$

(*i*) Let us transform to

$$u = \frac{x_1}{x_2}, v = x_2 \quad i.e., \quad x_1 = uv, x_2 = v$$

$$J = \frac{\partial (x_1, x_2)}{\partial (u, v)} = \begin{vmatrix} v & 0 \\ u & 1 \end{vmatrix} = v$$



$x_1 = 0$ maps to $u = 0, v = 0$

$x_1 = 1$ maps to $uv = 1$ (Rectangular hyperbola)

$x_2 = 0$ maps to $v = 0$ and $x_2 = 1$ maps to $v = 1$.

The joint p.d.f. of U and V becomes

$$g(u, v) = f(x_1, x_2) |J| = v \; ; 0 < u < \infty, 0 < v < \infty$$

To obtain the marginal distribution of U, we have to integrate out v.

In region (I),

$$g_1(u) = \int_0^1 v \, dv = \left| \frac{v^2}{2} \right|_0^1 = \frac{1}{2}, 0 \le u \le 1$$

In region (II),

$$g_1(u) = \int_0^{1/u} v \, dv = \left| \frac{v^2}{2} \right|_0^{1/u} = \frac{1}{2u^2}, 1 < u < \infty$$

Hence the distribution of $U = \frac{X_1}{X_2}$ is given by

$$g(u) = \frac{1}{2}, 0 \le u \le 1$$

$$= \frac{1}{2u^2}, 1 \le u < \infty$$

(ii) Let $u = x_1 x_2$, $v = x_1$, *i.e.*, $x_1 = v, x_2 = \frac{u}{v}$

$$J = \begin{vmatrix} 0 & 1 \\ \frac{1}{v} & -\frac{u}{v^2} \end{vmatrix} = -\frac{1}{v}$$

$x_1 = 0$ maps to $v = 0, x_1 = 1$ maps to $v = 1$

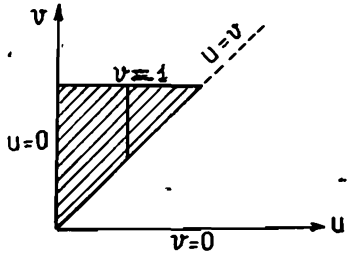
$x_2 = 0$ maps to $u = 0$, and $x_2 = 1$ maps to $u = v$

Moreover, $v = \frac{u}{x_2} \Rightarrow v \geq u$ (since $0 < x_2 < 1$),

The joint p.d.f. of U and V is

$$g(u, v) = f(x_1, x_2) \mid J \mid = \frac{1}{v} ; 0 < u \leq 1, 0 < v < 1$$

$$g(u) = \int_u^1 \frac{1}{v} dv = [\log v]_u^1 = -\log u, 0 < u < 1$$



(*iii*) and (*iv*). Let $u = x_1 + x_2$, $v = x_1 - x_2$

$$i.e., x_1 = \frac{u+v}{2} \quad \begin{cases} x_1 = 0 \Rightarrow u + v = 0 \\ i.e., v = -u \\ x_2 = 0 \Rightarrow u - v = 0 \end{cases}$$

$$x_2 = \frac{u-v}{2} \quad \begin{cases} i.e., v = u \\ x_1 = 1 \Rightarrow u + v = 2 \\ x_2 = 1 \Rightarrow u - v = 2 \end{cases}$$

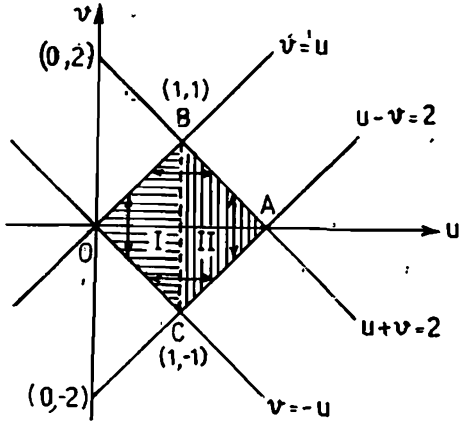
$$\text{and } J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

$\therefore\ g(u, v) = f(x_1, x_2) \mid J \mid = \frac{1}{2}, 0 < u < 2, -1 < v < 1$

In region (*I*), (see figure below)

$$g_1(u) = \int_{-u}^{u} \frac{1}{2} dv = \frac{1}{2} [v]_{-u}^{u} = u$$

and in region (*II*),



$$g_2(u) = \int_{u-2}^{2-u} \frac{1}{2}\, dv = \frac{1}{2} \mid v \mid_{u-2}^{2-u} = 2 - u$$

$$\therefore \quad g(u) = \begin{cases} u, & 0 < u < 1 \\ 2 - u, & 1 < u < 2 \end{cases}$$

For the distribution of V, we split the region as: OAB and OAC

In region OAB:

$$h_1(v) = \int_{v}^{2-v} \frac{1}{2}\, du = \frac{1}{2}[2 - v - v] = 1 - v, \; 0 < v < 1$$

In region OAC:

$$h_2(v) = \int_{-v}^{2+v} \frac{1}{2}\, du = \frac{1}{2}[2(1 + v)] = 1 + v, \; -1 < v < 0$$

Hence the distribution of $V = X_1 - X_2$ is given by

$$h(v) = \begin{cases} 1 - v, & 0 < v < 1 \\ 1 + v, & -1 < v < 0 \end{cases}$$

Example 8·6. *If X is a random variable with a continuous distribution function F, then $F(X)$ has a uniform distribution on [0, 1].*

[Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1987,' 85]

Solution. Since F is a distribution function, it is non-decreasing. Let $Y = F(X)$, then the distribution function G of Y is given by

$$G_Y(y) = P(Y \le y) = P[F(X) \le y] = P[X \le F^{-1}(y)],$$

the inverse exists, since F is non-decreasing and given to be continuous.

$$\therefore \quad G_Y(y) = F[F^{-1}(y)],$$

since F is the distribution function of X.

$$\therefore \quad G_Y(y) = y$$

Therefore the p.d.f. of $Y = F(X)$ is given by:

$$g_Y(y) = \frac{d}{dy}[G_Y(y)] = 1$$

Since F is a d.f., Y takes the values in the range [0, 1].

Hence $g_Y(y) = 1, \; 0 \le y \le 1$

$\Rightarrow$ Y is a uniform variate on [0, 1].

Remark. Suppose X is a random variable with p.d.f.,

$$f_X(x) = \begin{cases} e^{-x}, & x \ge 0 \\ 0, & \text{otherwise} \end{cases}$$

then

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - e^{-x}, & \text{if } x \ge 0 \end{cases}$$

Then by above result $F(X) = 1 - e^{-X}$ is uniformly distributed on [0, 1].

Example 8·7. *If X and Y are independent rectangular variates for the range $-a$ to a each, then show that the sum $X + Y = U$, has the probability density*

$$\varphi(u) = \frac{2a + u}{4a^2}, \quad -2a \leq u \leq 0$$

$$\varphi(u) = \frac{2a - u}{4a^2}, \quad 0 \leq u \leq 2a$$

Solution. Since X and Y are independent rectangular variates, each in the interval $(-a, a)$, we have

$$f_1(x) = \begin{cases} \frac{1}{2a}, & -a < x < a \\ 0, & \text{elsewhere} \end{cases}$$

and

$$f_2(y) = \begin{cases} \frac{1}{2a}, & -a < y < a \\ 0, & \text{elsewhere} \end{cases}$$

Hence by compound probability theorem, the joint probability differential of X and Y is given by

$$dP(x, y) = f_1(x) f_2(y)\, dx\, dy = \frac{1}{4a^2} dx\, dy, \quad -a < (x, y) < a$$

Let us define new variables U and V as follows :

$$u = x + y, \qquad v = x - y$$

$$\Rightarrow \qquad x = \frac{u + v}{2} \quad \text{and} \quad y = \frac{u - v}{2}$$

Jacobian of the transformation J is given by

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

Thus the probability differential of U and V becomes

$$dG(u, v) = \frac{1}{4a^2} |J|\, du\, dv = \frac{1}{8a^2} du\, dv \qquad \ldots(*)$$

Integrating w.r.to. v over specified range, we can find the distribution of U.

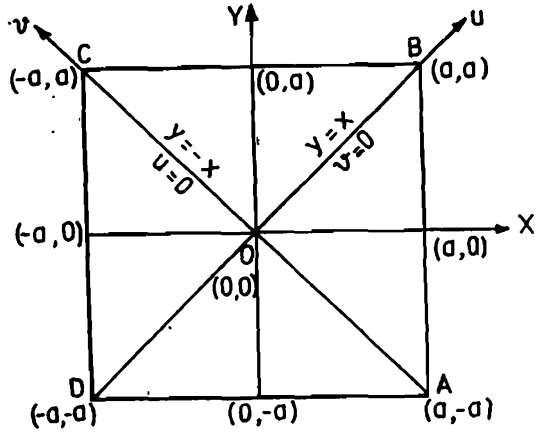
Let us consider the region to the left of $v-$ axis, *i.e.*, to the left of the line AC. In this region, the values of v are bounded by the lines $x = -a$ and $y = -a$.

For fixed values of u,

$$x = -a \quad \Rightarrow \quad \frac{u + v}{2} = -a \quad \Rightarrow \quad v = -(u + 2a)$$

and

$$y = -a \quad \Rightarrow \quad \frac{u - v}{2} = -a \quad \Rightarrow \quad v = (u + 2a)$$



Thus integrating (*) w.r.to. v between the limits $-(u+2a)$ and $(u+2a)$, the distribution of U becomes

$$g_1(u)\,du = \int_{-(u+2a)}^{u+2a} \frac{1}{8a^2}\,du\,dv = \frac{1}{8a^2}\,|\,v\,|_{-(u+2a)}^{u+2a}\,du = \frac{u+2a}{4a^2}\,du$$

In the region to the left of v-axis, *i.e.*, below the line AC, u varies from the points $(x=-a,\ y=-a)$ to the point $(x=0, y=0)$ and since $u=x+y$, in this region u lies between $(-a-a)$ and $(0+0)$, *i.e.*, between $-2a$ to 0.

$$\therefore \qquad g_1(u)\,du = \frac{u+2a}{4a^2},\ -2a \le u \le 0$$

In the region to the right of v–axis, *i.e.*, above the line AC, the values of v are bounded by the lines $x=a$ and $y=a$ and for fixed values of u,

$$x=a \quad \Rightarrow \quad \frac{u+v}{2}=a \quad \Rightarrow \quad v=2a-u$$

$$y=a \quad \Rightarrow \quad \frac{u-v}{2}=a \quad \Rightarrow \quad v=-(2a-u)$$

In this region u varies from the point $(x=0, y=0)$ to the point $(x=a, y=a)$, i.e., $u=x+y$ varies from 0 to $2a$. Thus integrating (*) w.r.to. v between the limits $-(2a-u)$ to $(2a-u)$, we get the distribution of U as

$$g_1(u)\,du = \int_{-(2a-u)}^{2a-u} \frac{1}{8a^2}\,du\,dv = \frac{1}{8a^2}\,|\,v\,|_{-(2a-u)}^{2a-u}\,du$$

$$= \frac{2a-u}{4a^2}\,du,\ 0 \le u \le 2a$$

For an alternative and simpler solution, see Remark 5 to § 8·1·5, (Triangular Distribution).

Example. 8·8. *On the x-axis* $(n+1)$ *points are taken independently between the origin and* $x=1$, *all positions being equally likely. Show that probability*

that the $(k+1)$ th of these points, counted from the origin, lies in the interval $x-\frac{1}{2}dx$ to $x+\frac{1}{2}dx$ is

$$\binom{n}{k}(n+1)\,x^k\,(1-x)^{n-k}\,dx$$

Verify that integral of this expression from $x=0$ to $x=1$ is unity.

Solution. Here X is given to be a random variable uniformly distributed on $[0, 1]$.

$$\therefore \qquad f_X(x) = 1,\ 0 \le x \le 1$$

$$\text{Now} \qquad P(0 < X < x) = \int_0^x f(x)\,dx = \int_0^x 1 \cdot dx = x \qquad \ldots(1)$$

$$\therefore \qquad P(X > x) = 1 - P(X \le x) = 1 - x \qquad \ldots(2)$$

$$\text{Also } P\left(x - \frac{dx}{2} < X < x + \frac{dx}{2}\right) = \int_{x-\frac{dx}{2}}^{x+\frac{dx}{2}} f(x)\,dx = dx \qquad \ldots(3)$$

Required probability 'p' is given by

$p = P\{$ out of $(n+1)$ points, k points lie in the closed interval $\left[0, x - \frac{dx}{2}\right]$

and out of the remaining $(n+1-k)$ points, $(n-k)$ points lie in

$$\left[x + \frac{dx}{2}, 1\right] \text{ and one point lies in } \left(x - \frac{dx}{2}, x + \frac{dx}{2}\right)\Big\}$$

$$= \left[\binom{n+1}{k}x^k\right] \times \left[\binom{n+1-k}{n-k}(1-x)^{n-k}\right] \times dx,$$

on using (1), (2) and (3) respectively.

$$\therefore \qquad p = \frac{(n+1)!}{k!\,(n+1-k)!} \cdot x^k \cdot \frac{(n+1-k)!}{(n-k)!} \cdot (1-x)^{n-k}\,dx$$

$$= \binom{n}{k}(n+1)\,x^k\,(1-x)^{n-k}\,dx$$

To prove that the area of this expression from $x=0$ to $x=1$ is unity, use Beta-integral

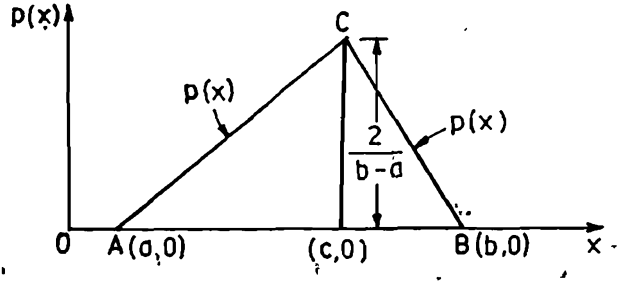
$$\int_0^1 x^{m-1}(1-x)^{n-1}\,dx = B(m, n) = \frac{\Gamma m\ \Gamma n}{\Gamma(m+n)};\ m > 0, n > 0.$$

8·1·5. Triangular Distribution. A random variable X is said to have a triangular distribution in the interval (a, b), if its p.d.f. is given by:

$$f(x) = \begin{cases} 2(x-a)/\{(b-a)(c-a)\} & ;\ a < x \le c \\ 2(b-x)/\{(b-a)(b-c)\} & ;\ c < x < b \end{cases} \qquad \ldots(8\cdot2a)$$

Remarks. 1. We write $X \sim$ Trg. (a, b), with peak at $x = c$. The graph of the p.d.f. is shown in the diagram on page 8·11.

2. The distribution is so called because the graph of its p.d.f. is a triangle with peak at $x = c$.



3. The m.g.f. of Trg (a, b) variate, with peak at $x = c$ is given by:

$$M_X(t) = \int_a^b e^{tx} f(x)\,dx = \left(\int_a^c + \int_c^b\right) e^{tx} f(x)\,dx .$$

$$= \frac{2}{(b-a)(c-a)} \int_a^c e^{tx}(x-a)\,dx + \frac{2}{(b-a)(b-c)} \int_c^b e^{tx}.(b-x)\,dx$$

$$= \frac{2}{t^2}\left\{\frac{e^{at}}{(a-b)(a-c)} + \frac{e^{ct}}{(c-a)(c-b)} + \frac{e^{bt}}{(b-a)(b-c)}\right\}; a < b < c$$

(On integration by parts) ...(8·2*b*)

4. In particular, taking $a = 0$, $c = 1$ and $b = 2$, in (8·2*a*), the p.d.f. of the Trg (0, 2) variate with peak at $x = 1$ is given by:

$$f(x) = \begin{cases} x; & 0 \le x \le 1 \\ 2 - x; & 1 \le x \le 2 \\ 0, & \text{otherwise} \end{cases} \quad ...(8·2\,c)$$

and its m.g.f. is $M_X(t) = (e^t - 1)^2/t^2$, ...(8·2*d*)

which is left as an exercise to the reader.

5. In particular, replacing a by $-2a$, b by $2a$ and c by 0, the p.d.f. of triangular distribution on the interval $(-2a, 2a)$ with peak at $x = 0$ is given by:

$$f(x) = \begin{cases} (2a + x)/4a^2; & -2a < x < 0 \\ (2a - x)/4a^2; & 0 < x < 2a \end{cases} \quad ...(8·2e)$$

The m.g.f. of (8·2*e*) is given by :

$$M_X(t) = \int_{-2a}^{2a} e^{tx} f(x)\,dx$$

$$= \frac{1}{4a^2}\left[\int_{-2a}^{0} e^{tx}.(2a+x)\,dx + \int_0^{2a} e^{tx}(2a-x)\,dx\right]$$

$$= \frac{1}{4a^2}\left[e^{tx}\left\{ \frac{2a+x}{t} - \frac{1}{t^2} \right\} \right]_{-2a}^{0} + \frac{1}{4a^2}\left[e^{tx}\left\{ \frac{2a-x}{t} + \frac{1}{t^2} \right\} \right]_{0}^{2a}$$

[On integrating by parts]

$$= \frac{1}{4a^2}\left[-\frac{2}{t^2} + \frac{1}{t^2}\left\{ e^{2at} + e^{-2at} \right\} \right]$$

$$= \frac{1}{4a^2 t^2}\left\{ e^{2at} + e^{-2at} - 2 \right\} = \left[\frac{e^{at} - e^{-at}}{2at} \right]^2 \qquad ...(8\cdot2f)$$

Aliter. We may obtain (8·2 *f*) directly from (8·2 *b*) on replacing a by $-2a$, *b* by $2a$ and *c* by 0.

Example 8·9. *If X and Y are i.i.d. U* [$-a, a$] *variates, find the p.d.f. of* $Z = X + Y$ *and identify the distribution.*

Solution. Since *X* and *Y* are i.i.d. *U* [$-a, a$], we have : [*c.f.* § 8·1·2.],

$$M_X(t) = M_Y(t) = (e^{at} - e^{-at})/(2at) \qquad ...(*)$$

$$M_{X+Y}(t) = M_X(t)\, M_Y(t) = \left[\frac{e^{at} - e^{-at}}{2at} \right]^2, \qquad ...(**)$$

since *X* and *Y* are independent.

But (**) is the m.g.f. of Trg $(-2a, 2a)$ variate with peak at $x = 0$

[*c.f.* Remark 5, equation (8·2*f*)]

Hence by uniqueness theorem of m.g.f., $Z = X + Y \sim$ Trg $(-2a, 2a)$ with p.d.f. as given in (8·2 *e*), Remark 5.

Aliter $M_{X+Y}(t) = \dfrac{1}{4a^2 t^2}\left[e^{2at} - 2 + e^{-2at} \right]$ [From (**)]

$$= \frac{2}{t^2}\left[\frac{e^{-2at}}{(-2a-0)(-2a-2a)} + \frac{e^{0.t.}}{(0+2a)(0-2a)} + \frac{e^{2at}}{(2a-0)(2a+2a)} \right]$$

which is of the form (8·2 *b*), [*c.f.* Remark 3], with *a* replaced by $-2a$ and *b* replaced by $2a$ and *c* by 0. Hence $X + Y \sim$ Trg $(-2a, 2a)$ with p.d.f. $p(x)$ given in (8·2*e*).

Remarks 1. The distribution of $X + Y$ has also been obtained in Example 8·7.

2. Similarly we can find the distribution of $X - Y$.

$$M_{X-Y}(t) = M_X(t)\,.\,M_Y(-t) = \left[\frac{e^{at} - e^{-at}}{2at} \right]^2 \qquad \text{[From (*)]}$$

$\Rightarrow X - Y \sim$ Trg $(-2a, 2a)$, with peak at $x = 0$.

EXERCISE 8 (a)

1. The bus company *A* schedules a north bound bus every 30 minutes at a certain bus-stop. A man comes to the stop at a random time. Let the random variable *X* count the number of minutes he has to wait for the next bus. Assume *X* has a

uniform distribution over the interval (0, 30). This is how we interpret the statement that he enters the station at the random time].

(*i*) For each $k = 5, 10, 15, 20, 30$ compute the probability that he has to wait at least k minutes for the next bus.

(*ii*) A competitor, the bus company B is allowed to schedule a north bound bus every 30 minutes at the same station but at least 5 minutes must elapse between the arrivals of the competitive buses. Assume the passengers come at the bus stop at random times and always board the first bus that arrives. Show that the company B can arrange its schedule so that it receives five times as many passengers as that of its competitor.

2. (*a*) A random variable X has a uniform distribution over $(-3, 3)$, compute

(*i*) $P(X = 2)$, $P(X < 2)$, $P(|X| < 2)$ and $P(|X - 2| < 2)$

(*ii*) Find k for which $P(X > k) = 1/3$. **[Gorakhpur Univ. B.Sc. 1992]**

(*b*) Suppose that X is uniformly distributed over $(-\alpha, +\alpha)$, where $\alpha > 0$. Determine α so that

(*i*) $P(X > 1) = 1/3$, (*ii*) $P(X < 1/2) = 0{\cdot}3$ and

(*ii*) $P(|X| < 1) = P(|X| > 1)$.

Ans. (*i*) $\alpha = 3$, (*ii*) $\alpha = 5/6$, (*iii*) $\alpha = 2$.

(*c*) Calculate the coefficient of variation for the rectangular distribution in $(0, b)$ given that the probability law of the distribution is

$$P(X \le t) = \frac{t}{b}$$

(*d*) If X is uniformly distributed over $[1, 2]$, find z so that

$$P(X > z + \mu_x) = \frac{1}{4} \qquad \text{(Ans. } Z = \frac{1}{4}\text{)}.$$

3 (*a*). If a random variable X has the density function $f(x)$, prove that

$$Y = \int_{-\infty}^{x} f(x)\,dx$$

has a rectangular distribution over (0, 1). If

$$f(x) = \tfrac{1}{2}(x - 1), \quad 1 \le x \le 3$$
$$= 0, \text{ otherwise}$$

determine what interval for Y will correspond to the interval

$$1{\cdot}1 \le X \le 2{\cdot}9.$$

Ans. $y = F(x) = (x - 1)^2/4$; $1 \le x \le 3$; $0{\cdot}0025 \le y \le 0{\cdot}9025$

(*b*). Show that whatever be the distribution function $F(x)$ of a r.v. X,

$$P[a \le F(X) \le b] = b - a, \quad 0 \le (a, b) \le 1.$$

[Delhi Univ. B.Sc. (Stat. Hons), 1986]

Hint. $Y = F(X) \sim U[0, 1]$.

4. (*a*) For the rectantular distribution,

$$f(x) = \frac{1}{2a}, -a \le x \le a,$$
$$= 0, \text{ otherwise},$$

show that the moments of odd order are zero, and $\mu_{2r} = a^{2r}/(2r+1)$.

[Madurai Kamraj Univ. B.Sc., 1992]

(*b*) A distribution is given by

$$f(x)\,dx = \frac{1}{2a}\,dx, -a \le x \le a$$

Find the first four central moments and obtain β_1 and β_2.

[Delhi Univ B.Sc. Oct., 1992; Madras Univ. B.Sc., 1991]

(*c*) For a rectangular distribution

$$dP = k\,dx,\ 1 \le x \le 2,$$

show that Arithmetic mean > Geometric mean > Harmonic mean.

[Vikram Univ. B.Sc. 1993)

(*d*) If the random variable X follows the rectangular distribution with p.d.f.,

$$f(x) = 1/\theta,\ 0 \le x \le \theta,$$

derive the first four moments and the skewness and kurtosis conffi cients of the distribution.

(*e*) Let X and Y be independent variates which are uniformly distributed over the unit interval (0,1). Find the distribution function and the p.d.f. of random variable $Z = X + Y$. Is Z a uniformly distributed variable ? Give reasons.

[Delhi Univ. B.Sc. (Maths. Hons.), 1986]

5. Let X_1 and X_2 be independent random variables unifromly distributed over the interval (0, 1). Find

(*i*) $P(X_1 + X_2 < 0{\cdot}5)$, (*ii*) $P(X_1 - X_2 < 0{\cdot}5)$,

(*iii*) $P(X_1^2 + X_2^2 < 0{\cdot}5)$, (*iv*) $P(e^{-X_1} < 0{\cdot}5)$, and (*v*) $P(\cos \pi X_2 < 0{\cdot}5)$.

Ans. (*i*) 0·125, (*ii*) 0·875. (*iii*) 0·393, (*iv*) $1 - \log 2$, and (*v*) 2/3.

6. A random variable X is uniformly distributed over (0, 1), find the probability density functions of

(*i*) $Y = X^2 + 1$, and (*ii*) $Z = 1/(X+1)$.

7. (*a*) If the random variable X is uniformly distributed over $(0, \frac{1}{2}\pi)$, compute the expectation of the function $\sin X$. Also find the distribution of $Y = \sin X$, and show that the mean of this distribution is the same as the above expectation.

Ans. $2/\pi$, $f_Y(y) = 2/(\pi\sqrt{1-y^2})$, $0 < y < 1$.

(*b*) If $X \sim U[-\pi/2, \pi/2]$ distributed, find the p.d.f. of $Y = \tan X$.

[Delhi Univ. B.A. Hons. (Spl. Course-Statistics), 1989]

8. (*a*) Show that for the rectangular distribution :

$$dF = dx,\ 0 \le x < 1$$

μ'_1 (about origin) = 1/2, variance = 1/12 and mean deviation about mean = 1/4. **[Madras Univ B.Sc. Sept. 1991; Delhi U. B.Sc. Sept. 1992]**

(*b*) Find the characteristic function of the random variable $Y = \log F(X)$ where $F(X)$ is the distribution function of a random variable X. Evaluate the rth moment of Y.

9. If $X \sim U[0, 1]$, find the distribution of $Y = 1/X$. Find $E(1/X)$, if it exists.

Ans. $g_Y(y) = 1/y^2$; $1 \le y < \infty$; $E(Y) = E(1/X)$ does not exist.

10. Let X be uniformly distributed on $[-1, 1]$. Find the distribution function and hence the p.d.f. of $Y = X^2$. **[Delhi Univ. B.Sc. (Maths. Hons.), 1988]**

11. Let $f_X(x) = 6x(1-x)$; $0 \le x \le 1$. Find y as a function of x such that Y has p.d.f.

$$g(y) = 3(1 - \sqrt{y}); \quad 0 \le y \le 1$$

[Delhi Univ. B.A. Hons. (Spl. Course-Statistics), 1988]

Hint. $$F(x) = \int_0^x f(x)\,dx = 3x^2 - 2x^3 \sim U[0, 1]$$

$$G(y) = \int_0^y g(y)\,dy = 3y - 2y^{3/2} \sim U[0, 1]$$

Setting $F(x) = G(y)$, we get $y = x^2$.

12. The variates a and b are independently and uniformly distributed in the intervals [0,6] and [0, 9] respectively. Find the probability that $x^2 - ax + b = 0$ has two real roots.

Ans. $$P(b \le a^2/4) = \int_{a=0}^{1} \int_{b=0}^{a^2/4} \frac{1}{6 \times 9}\,da\,db = 1/3.$$

13. Find the probability that the roots of the equation $x^2 + 2bx + c = 0$ should be real, given that $b \sim U[-\alpha, \alpha]$ and $c \sim U[-\beta, \beta]$ are independent.

Ans. Probability $= P(b^2 \ge c) = 1 - P(b^2 \le c) = 1 - P(|b| \le \sqrt{c})$

$$= \int_{-\beta}^{\beta} \left(\int_{-\sqrt{c}}^{\sqrt{c}} \left(\frac{1}{2\alpha}\right)\left(\frac{1}{2\beta}\right) db \right) dc$$

14. If a, b, c are randomly chosen between 0 and 1, find the probability that the quadratic equation $ax^2 + bx + c = 0$ has real roots.

Ans. Probability $= P(b^2 \ge 4ac) = 1 - \int_0^1 \int_0^1 \int_0^{\sqrt{4ac}} 1\,da\,dc\,db = \frac{8}{9}$

15. (*a*) Suppose X has a rectangular distribution on $(-1, 1)$. Compute $P\left[\frac{|X - E(X)|}{\sigma_X} \ge 2\right]$ and compare it with the upper bound given by Chebyshev's inequality.

(*b*) Compare the upper bound of the probability,

$$P\left\{|X - E(X)| \geq 2\sqrt{V(X)}\right\},$$

obtained from Chebyshev's inequality with exact probability if X is uniformly distributed over $(-1, 3)$.

Ans. (*b*) Probability $\leq 1/4$, Exact Probability $= 0$

16. Two independent variates are each uniformly distributed within the range $-a$ to $+a$. Show that their sum X has a probability density given by

$$f(x) = \frac{2a+x}{4a^2}, \quad -2a \leq x \leq 0$$

$$= \frac{2a-x}{4a^2}, \quad 0 \leq x \leq 2a$$

Verify that the m.g.f. calculated from the value of $f(x)$ is equal to

$$\left(\frac{1}{at}\sinh at\right)^2$$

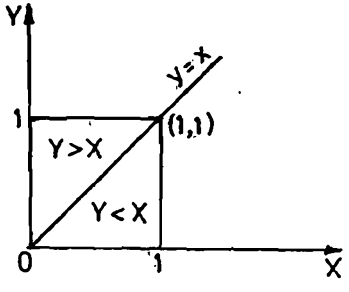
17. The random variables X and Y are independent and both have the uniform distribution on $[0, 1]$. Let $Z = |X - Y|$. Prove that. for real θ,

$$\varphi(Z, \theta) = 2\left[1 + i\theta - e^{i\theta}\right]/2.$$

Hence deduce the general expression for $E(Z^n)$.

Hint. $\varphi(\theta; |X - Y|) = \int_0^1 \int_0^1 e^{i\theta|x-y|} f(x, y)\, dx\, dy$

$$= 2\int_0^1 \left(\int_0^x e^{i\theta(x-y)}\, dy\right) dx$$



Ans. $2/[(n+1)(n+2)]$

18. If X and Y are independently and uniformly distributed random variables in the interval $(0, 1)$, show that the distribution of $X + Y$ is given by the density function

$$f(z) = \begin{cases} z & 0 \leq z < 1 \\ 2 - z & 1 \leq z \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

[**Hint.** See Triangular distribution]

19. Ship A makes radio signals to the base and the probability of the interval between consecutive signals is uniformly distributed between 4 hours and 24 hours and is zero outside this range. Ship B makes radio signals to the base and the probability of the interval between consecutive signals is uniformly distributed between 10 hours and 15 hours and is zero outside this range.

(*i*) Ship A has just signalled. What is the probability that it will make two further signals in the next 12 hours ?

(*ii*) **Ships** A and B have just signalled at the same time. What is the probability that Ship A will make at least two further signals before ship B next signals? **[Institute of Actuaries (London), April 1978]**

20. If $X \sim U[0, 1]$, prove that for $b < c$ fixed, $Y = (c - b)X + b$ is uniform on $[b, c]$.

8·2. Normal Distribution. The normal distribution was first discovered in 1733 by English mathematician De-Moivre, who obtained this continuous distribution as a limiting case of the binomial distribution and applied it to problems arising in the game of chance. It was also known to Laplace, no later than 1774 but through a historical error it was credited to Gauss, who first made reference to it in the beginning of 19th century (1809), as the distribution of errors in Astronomy. Gauss used the normal curve to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies. Throughout the eighteenth and nineteeth centuries, various efforts were made to establish the normal model as the underlying law ruling all continuous random variables. Thus, the name "*normal*" . These efforts, however, failed because of false premises. The normal model has, nevertheless, become the most important probability model in statistical analysis.

Definition. *A random variable X is said to ahve a normal distribution with parameters* μ *(called "mean") and* σ^2 *(called "variance") if its density function is given by the probability law :*

$$f(x\,;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^2\right]$$

or $$f(x\,;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}$$

$$-\infty < x < \infty,\ -\infty < \mu < \infty,\ \sigma > 0 \quad ...(8·3)$$

Remarks. 1. A random variable X with mean μ and variance σ^2 and following the normal law (8·3) is expressed by $X \sim N(\mu, \sigma^2)$

2. If $X \sim N(\mu, \sigma^2)$, then $Z = \dfrac{X-\mu}{\sigma}$, is a standard normal variate with

$$E(Z) = 0 \text{ and } \text{Var}(Z) = 1$$

and we write $Z \sim N(0,1)$.

3. The p.d.f. of standard normal variate Z is given by

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

and the corresponding distribution function, denoted by $\Phi(z)$ is given by

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \varphi(u)\, du$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du$$

We shall prove below two important results on the distribution function $\Phi(.)$ of standard normal variate.

Result 1. $\Phi(-z) = 1 - \Phi(z)$

Proof. $\Phi(-z) = P(Z \leq -z) = P(Z \geq z)$ (By symmetry)

$$= 1 - P(Z \leq z)$$

$$= 1 - \Phi(z)$$

Result 2. $P(a \leq X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$, where $X \sim N(\mu, \sigma^2)$

Proof. $P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right)$; $\left(Z = \frac{X-\mu}{\sigma}\right)$.

$$= P\left(Z \leq \frac{b-\mu}{\sigma}\right) - P\left(Z \leq \frac{a-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

4. The graph of $f(x)$ is a famous *'bell-shaped'* curve. The top of the bell is directly above the mean μ. For large values of σ, the curve tends to flatten out and for small values of σ, it has a sharp peak.

8·2·1. Normal Distribution as a Limiting form of Binomial Distribution. Normal distribution is another limiting form of the binomial distribution under the following conditions :

(*i*) n, the number of trials is indefinitely large, *i.e.*, $n \to \infty$ and

(*ii*) neither p nor q is very small.

The probability function of the binomial distribution with parameters n and p is given by

$$p(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!\,(n-x)!}\, p^x q^{n-x}; \; x = 0, 1, 2, \ldots, n \qquad \ldots(*)$$

Let us now consider the standard binomial variate :

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - np}{\sqrt{npq}}; \; X = 0, 1, 2, \ldots, n \qquad \ldots(**)$$

When $X = 0, Z = \frac{-np}{\sqrt{npq}} = -\sqrt{np/q}$

and when $X = n$, $Z = \dfrac{n - np}{\sqrt{npq}} = \sqrt{nq/p}$

Thus in the limit as $n \to \infty$, Z takes the values from $-\infty$ to ∞. Hence the distribution of X will be a continuous distribution over the range $-\infty$ to ∞.

We want the limiting form of (*) under the above two conditions. Using Stirling's approximation to $r!$ for large r, *viz.*,

$$\lim_{r \to \infty} r! \simeq \sqrt{2\pi}\, e^{-r}\, r^{r+(1/2)},$$

we have in the limit as $n \to \infty$ and consequently $x \to \infty$,

$$\lim p(x) = \lim \left[\frac{\sqrt{2\pi}\, e^{-n}\, n^{n+\frac{1}{2}}\, p^x\, q^{n-x}}{\sqrt{2\pi}\, e^{-x}\, x^{x+\frac{1}{2}}\, \sqrt{2\pi}\, e^{-(n-x)}\, (n-x)^{n-x+\frac{1}{2}}} \right]$$

$$= \lim \left[\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{npq}} \cdot \frac{(np)^{x+\frac{1}{2}}\, (nq)^{n-x+\frac{1}{2}}}{x^{x+\frac{1}{2}}\, (n-x)^{n-x+\frac{1}{2}}} \right]$$

$$= \lim \left[\frac{1}{\sqrt{2\pi}\, \sqrt{npq}} \left(\frac{np}{x} \right)^{x+\frac{1}{2}} \left(\frac{nq}{n-x} \right)^{n-x+\frac{1}{2}} \right] \quad \ldots(***)$$

From (**), we have

$$X = np + Z\sqrt{npq} \quad \Rightarrow \quad \frac{X}{np} = 1 + Z\sqrt{q/(np)}$$

Also

$$n - X = n - np - Z\sqrt{npq} = nq - Z\sqrt{npq}$$

$$\therefore \quad \frac{n - X}{nq} = 1 - Z\sqrt{p/(nq)} \quad . \text{ Also } dz = \frac{1}{\sqrt{npq}}\, dx$$

Hence the probability differential of the distribution of Z, in the limit is given from (***) by

$$d\,G(z) = g(z)\, dz = \lim_{n \to \infty} \left[\frac{1}{\sqrt{2\pi}} \times \frac{1}{N} \right] dz \qquad \ldots(8{\cdot}4)$$

where $N = \left[\dfrac{x}{np} \right]^{x+\frac{1}{2}} \left[\dfrac{n-x}{nq} \right]^{n-x+\frac{1}{2}}$

$$\log N = (x + \tfrac{1}{2}) \log (x/np) + (n - x + \tfrac{1}{2}) \log \{ (n-x)/nq \},$$

$$= (np + z\sqrt{npq} + \tfrac{1}{2}) \log [\, 1 + z\sqrt{(q/np)}\,]$$

$$+ (nq - z\sqrt{npq} + \tfrac{1}{2}) \log [\, 1 - z\sqrt{(p/nq)}\,]$$

$$= (np + z\sqrt{npq} + \tfrac{1}{2}) [\, z\sqrt{(q/np)} - \tfrac{1}{2} z^2 (q/np) + \tfrac{1}{3} z^3 (q/np)^{3/2} - \ldots\,]$$

$$+ (nq - z\sqrt{npq} + \tfrac{1}{2}) \left[- z\sqrt{(p/nq)} - \tfrac{1}{2} z^2 (p/nq) - \tfrac{1}{3} z^3 (p/nq)^{3/2} - \ldots \right]$$

$$= \left[\left\{ z\sqrt{npq} - \tfrac{1}{2} qz^2 + \tfrac{1}{3} z^3 \frac{q^{3/2}}{\sqrt{np}} + z^2 q - \tfrac{1}{2} z^3 \frac{q^{3/2}}{\sqrt{np}} + \tfrac{1}{2} z\sqrt{q/np} - \tfrac{1}{4} z^2 \frac{q}{np} + \ldots \right) + \left(- z\sqrt{npq} - \tfrac{1}{2} z^2 p - \tfrac{1}{3} z^3 \frac{p^{3/2}}{\sqrt{nq}} + z^2 p + \tfrac{1}{2} z^3 \frac{p^{3/2}}{\sqrt{np}} - \tfrac{1}{2} z\sqrt{p/nq} - \tfrac{1}{4} z^2 \frac{p}{np} + \ldots \right) \right\} \right]$$

i.e.,

$$\log N = \left[-\frac{1}{2} z^2 (p+q) + z^2 (p+q) + \frac{z}{2\sqrt{n}} \left\{ \sqrt{q/p} + \sqrt{p/q} \right\} + 0\left\{ n^{-1/2} \right\} \right]$$

$$= \frac{z^2}{2} + 0\,(n^{-1/2}) \to \frac{z^2}{2} \text{ as } n \to \infty$$

$$\therefore \lim_{n \to \infty} \log N = \frac{z^2}{2} \quad \Rightarrow \quad \lim_{n \to \infty} N = e^{z^2/2}$$

Substituting in (8·4), we get

$$d\,G(z) = g\,(z)\,dz = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,dz, \; -\infty < z < \infty \qquad \ldots(8\cdot4\,a)$$

Hence the probability function of Z is

$$g\,(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \; -\infty < z < \infty \qquad \ldots(8\cdot4\,b)$$

This is the probability density function of the *normal distribution* with mean 0 and unit variance.

If X is normal variate with mean μ and s.d. σ then $Z = (X - \mu)/\sigma$ is standard normal variate. Jacobian of transformation is $1/\sigma$. Hence substituting in {8·4 (*b*)}, the p.d.f. of a normal variate X with $E(X) = \mu$, $\text{Var}(X) = \sigma^2$ is given by

$$f_X(x) = \begin{cases} \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \; -\infty < x < \infty \\ 0, \text{ otherwise} \end{cases}$$

Remark. Normal distribution can also be obtained as a limiting case of Poisson Distribution with the parameter $\lambda \to \infty$.

8·2·2. Chief Characteristics of the Normal Distribution and Normal Probability Curve. The normal probability curve with mean μ and standard deviation σ is given by the equation

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

and has the following properties :

(*i*) The curve is bell shaped and symmetrical about the line $x = \mu$.

(*ii*) Mean, median and mode of the distribution coincide.

(*iii*) As x increases numerically, $f(x)$ decreases rapidly, the maximum probability occurring at the point $x = \mu$, and given by $[p(x)]_{max} = \frac{1}{\sigma\sqrt{2\pi}}$.

(*iv*) $\beta_1 = 0$ and $\beta_2 = 3$.

(*v*) $\mu_{2r+1} = 0$, $(r = 0, 1, 2, ...)$,
and $\mu_{2r} = 1.3.5 ... (2r-1)\sigma^{2r}$, $(r = 0, 1, 2, ...)$.

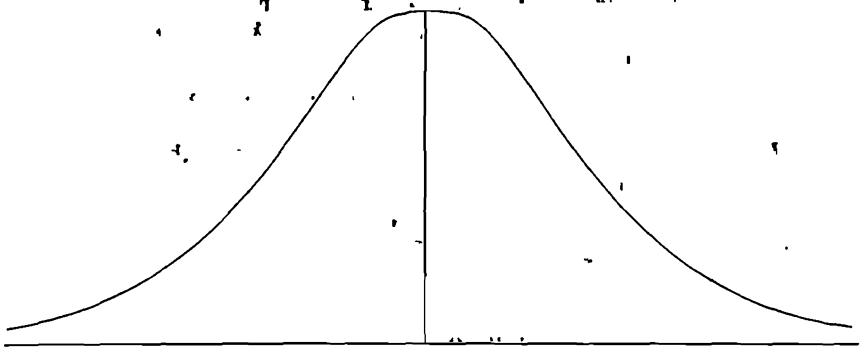
(*vi*) Since $f(x)$ being the probability, can never be negative, no portion of the curve lies below the x-axis.

(*vii*) Linear combination of independent normal variates is also a normal variate.

(*viii*) x-axis is an asymptote to the curve.

(*ix*) The points of inflexion of the curve are given by

$$\left[x = \mu \pm \sigma, f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2}\right]$$



(Normal Probability Cuve)

(*x*) Mean deviation about mean is $\sqrt{2/\pi}\,\sigma \approx \frac{4}{5}\sigma$ (approx.)

Q.D. $= \frac{Q_3 - Q_1}{2} \simeq \frac{2}{3}\sigma$

We have (approximately)

$$\textbf{Q.D. : M.D. : S.D.} :: \frac{2}{3}\sigma : \frac{4}{5}\sigma : \sigma :: \frac{2}{3} : \frac{4}{5} : 1$$

$\Rightarrow$ *Q. D. : M.D.: S.D.* :: 10 : 12 : 15

(*xi*) *Area Property*

$$P(\mu - \sigma < X < \mu + \sigma) = 0{\cdot}6826$$
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0{\cdot}9544$$
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0{\cdot}9973$$

The following table gives the area under the normal probability curve for some important values of standard normal variate Z.

Distances from the mean ordinates in terms of $\pm\sigma$	*Area under the curve*
$Z = \pm 0.745$	$50\% = 0.50$
$Z = \pm 1.00$	$68.26\% = 0.6826$
$Z = \pm 1.96$	$95\% = 0.95$
$Z = \pm 2.0$	$95.44\% = 0.9544$
$Z = \pm 2.58$	$99\% = 0.99$
$Z = \pm 3.0$	$99.73\% = 0.9973$

(*xii*) If X and Y are independent standard normal variates, then it can be easily proved that $U = X + Y$ and $V = X - Y$ are independently distributed, $U \sim N(0, 2)$ and $V \sim N(0, 2)$.

We state (without proof) the converse of this result which is due to D. Bernstein.

Bernstein's Theorem. If X and Y are independent and identically distributed random variables with finite varaince and if $U = X + Y$ and $V = X - Y$ are independent, then all r.v.'s X, Y, U and V are normally distributed.

(*xiii*) We state below another result which characterises the normal distribution.

If $X_1, X_2, \ldots, X_n$ are *i.i.d. r.v.'s* with finite variance, then the common distribution is normal if and only if :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

or $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2$

are independent. **[For 'If part', see Theorem 13.5]**

In the following sequences we shall establish some of these properties.

8.2.3. Mode of Normal Distribution. Mode is the value of x for which $f(x)$ is maximum, *i.e.*, mode is the solution of

$$f'(x) = 0 \text{ and } f''(x) < 0$$

For normal distribution with mean μ and standard deviation σ,

$$\log f(x) = c - \frac{1}{2\sigma^2}(x - \mu)^2,$$

where $c = \log(1/\sqrt{2\pi}\,\sigma)$, is a constant.

Differentiating w.r.t. x, we get

$$\frac{1}{f(x)} \cdot f'(x) = -\frac{1}{\sigma^2}(x - \mu) \Rightarrow f'(x) = -\frac{1}{\sigma^2}(x - \mu) f(x)$$

and $$f''(x) = -\frac{1}{\sigma^2}\left[1 \cdot f(x) + (x-\mu) f'(x)\right] = -\frac{f(x)}{\sigma^2}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right] \quad ...(8·6)$$

Now $f'(x) = 0 \Rightarrow x - \mu = 0$ *i.e.*, $x = \mu$

At the point $x = \mu$, we have from (8·6)

$$f''(x) = -\frac{1}{\sigma^2}[f(x)]_{x=\mu} = -\frac{1}{\sigma^2} \cdot \frac{1}{\sigma\sqrt{2\pi}} < 0$$

Hence $x = \mu$, is the mode of the normal distribution.

8·2·4. Median of Normal Distribution. If M is the median of the normal distribution, we have

$$\int_{-\infty}^{M} f(x)\,dx = \frac{1}{2} \Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{M} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx = \frac{1}{2}$$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx + \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} \exp\left\{-(x-\mu)^2/(2\sigma^2)\right\} dx = \frac{1}{2} \quad ...(8·7)$$

But $$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} \exp\left(-z^2/2\right) dz = \frac{1}{2}$$

$\therefore$ From (8·7), we get

$$\frac{1}{2} + \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx = \frac{1}{2}$$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \int_{\mu}^{M} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx = 0 \Rightarrow \mu = M$$

Hence for the normal distribution, Mean = Median.

Remark. From § 8·2·3 and § 8·2·4, we find that for the normal distribution mean, median and mode coincide. Hence the distribution is *symmetrical.*

8·2·5. M.G.F. of Normal Distribution. The m.g.f. (about origin) is given by

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{t(\mu + \sigma z)\right\} \exp\left(-z^2/2\right) dz, \qquad \left[z = \frac{x-\mu}{\sigma}\right]$$

$$= e^{\mu t} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\tfrac{1}{2}(z^2 - 2t\sigma z)\right\} dz$$

$$= e^{\mu t} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\tfrac{1}{2}\left\{(z-\sigma t)^2 - \sigma^2 t^2\right\}\right] dz$$

$$= e^{\mu t + t^2\sigma^2/2} \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\tfrac{1}{2}(z-\sigma t)^2\right\} dz$$

$$= e^{\mu t + t^2\sigma^2/2} \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(-u^2/2)\, du$$

Hence $$M_X(t) = e^{\mu t + t^2\sigma^2/2} \qquad ...(8·8)$$

Remark. *M.G.F. of Standard Normal Variate.* If $X \sim N(\mu, \sigma^2)$, then standard normal variate is given by

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

$$\text{Now } M_Z(t) = e^{-\mu t/\sigma}\ M_X(t/\sigma) = \exp(-\mu t/\sigma) \cdot \exp\left(\frac{\mu t}{\sigma} + \frac{t^2}{\sigma^2}\cdot\frac{\sigma^2}{2}\right)$$

$$= \exp(t^2/2) \qquad ...(8·8\,a)$$

8·2·6. Cumulant Generating Function (c.g.f.) of Normal Distribution. The c.g.f. of normal distribution is given by

$$K_X(t) = \log_e M_X(t) = \log_e (e^{\mu t + t^2\sigma^2/2}) = \mu t + \frac{t^2\sigma^2}{2}$$

$\therefore$ Mean $= \kappa_1 =$ Coefficient of t in $K_X(t) = \mu$

Variance $= \kappa_2 =$ Coefficient of $\frac{t^2}{2!}$ in $K_X(t) = \sigma^2$

and $\kappa_r =$ Coefficient of $\frac{t^r}{r!}$ in $K_X(t) = 0$; $r = 3, 4\ldots$

Thus $\mu_3 = \kappa_3 = 0$ and $\mu_4 = \kappa_4 + 3\kappa_2^2 = 3\sigma^4$

Hence $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$ and $\beta_2 = \frac{\mu_4}{\mu_2^2} = 3$ ...(8·9)

8·2·7. Moments of Normal Distribution. Odd order moments about mean are given by

$$\mu_{2n+1} = \int_{-\infty}^{\infty} (x-\mu)^{2n+1} f(x)\, dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^{2n+1} \exp\left\{-(x-\mu)^2/2\sigma^2\right\} dx$$

$\therefore$ $$\mu_{2n+1} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^{2n+1} \exp(-z^2/2)\, dz \qquad \left[z = \frac{x-\mu}{\sigma}\right]$$

$$= \frac{\sigma^{2n+1}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n+1} \exp(-z^2/2)\, dz = 0, \qquad ...(8·10)$$

since the integrand $z^{2n+1} e^{-z^2/2}$ is an odd function of z.

Even order moments about mean are given by

$$\mu_{2n} = \int_{-\infty}^{\infty} (x-\mu)^{2n} f(x)\, dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z)^{2n} \exp(-z^2/2)\, dz$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2n} \exp(-z^2/2)\, dz$$

$$= \frac{\sigma^{2n}}{\sqrt{2\pi}} \cdot 2 \int_{0}^{\infty} z^{2n} \exp(-z^2/2)\, dz$$

(since integrand is an even function of z)

$$\therefore \quad \mu_{2n} = \frac{2\sigma^{2n}}{\sqrt{2\pi}} \int_0^{\infty} (2t)^n e^{-t} \frac{dt}{\sqrt{2t}} \qquad \left[\frac{z^2}{2} = t\right]$$

$$= \frac{2^n \cdot \sigma^{2n}}{\sqrt{\pi}} \int_0^{\infty} e^{-t} t^{(n+\frac{1}{2})-1}\, dt$$

$$\Rightarrow \quad \mu_{2n} = \frac{2^n \sigma^{2n}}{\sqrt{\pi}} \cdot \Gamma(n+\tfrac{1}{2})$$

Changing n to $(n-1)$, we get

$$\mu_{2n-2} = \frac{2^{n-1} \cdot \sigma^{2n-2}}{\sqrt{\pi}} \Gamma(n-\tfrac{1}{2})$$

$$\therefore \quad \frac{\mu_{2n}}{\mu_{2n-2}} = 2\sigma^2 \cdot \frac{\Gamma(n+\frac{1}{2})}{\Gamma(n-\frac{1}{2})} = 2\sigma^2 (n-\tfrac{1}{2}) \quad [\because \Gamma(r) = (r-1)\Gamma(r-1)]$$

$$\Rightarrow \quad \mu_{2n} = \sigma^2 (2n-1)\, \mu_{2n-2} \qquad ...(8·11)$$

which gives the *recurrence relation* for the moments of normal distribution.

From (8·11), we have

$$\mu_{2n} = [(2n-1)\sigma^2][(2n-3)\sigma^2]\, \mu_{2n-4}$$

$$= [(2n-1)\sigma^2][2n-3)\sigma^2][(2n-5)\sigma^2]\, \mu_{2n-6}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$= [(2n-1)\sigma^2][(2n-3)\sigma^2][(2n-5)\sigma^2] \ldots (3\sigma^2)(1\sigma^2) \cdot \mu_0$$

$$= 1.3.5 \ldots (2n-1)\sigma^{2n} \qquad ...(8·12)$$

From (8·10) and (8·12) we conclude that *for the normal distribution all odd order moments about mean vanish and the even order moments about mean are given by (8·12).*

Aliter. The above result can also be obtained quite conveniently as follows:

The m.g.f. (about mean) is given by

$$E\left[e^{t(X-\mu)}\right] = e^{-\mu t} E\left(e^{tX}\right) = e^{-\mu t} M_X(t)$$

where $M_X(t)$ is the m.g.f. (about origin).

$$\therefore \quad \text{m.g.f. (about mean)} = e^{-\mu t} e^{\mu t + t^2\sigma^2/2} = e^{t^2\sigma^2/2}$$

$$= \left[1 + (t^2\sigma^2/2) + \frac{(t^2\sigma^2/2)^2}{2!} + \frac{(t^2\sigma^2/2)^3}{3!} + \ldots + \frac{(t^2\sigma^2/2)^n}{n!} + \ldots\right] \quad \ldots(8\cdot13)$$

The co efficient of $\frac{t^r}{r!}$ in (8·13) gives μ_r, the rth moment about mean. Since there is no term with odd powers of t in (8·13), all moments of odd order about mean vanish.

i.e., $\mu_{2n+1} = 0\,;\ n = 0, 1, 2, \ldots$

and $\mu_{2n} =$ **Coefficient of** $\frac{t^{2n}}{(2n)!}$ in (8·13) $= \frac{\sigma^{2n} \times (2n)!}{2^n\, n!}$

$$= \frac{\sigma^{2n}}{2^n\, n!}\left[2n(2n-1)(2n-2)(2n-3)\ldots 5.4.3.2.1\right]$$

$$= \frac{\sigma^{2n}}{2^n . n!}\left[1.3.5\ldots(2n-1)\right]\left[2.4.6\ \ldots\ (2n-2).2n\right]$$

$$= \frac{\sigma^{2n}}{2^n . n!}\left[1.3.5\ \ldots\ (2n-1)\right]2^n\left[1.2.3\ldots n\right]$$

$$= 1.3.5\ldots(2n-1)\,\sigma^{2n}$$

Remark. In particular, we have from (8·10) and (8·12),

$$\mu_3 = 0 \text{ and } \mu_2 = 1\,.\,\sigma^2,\ \mu_4 = 1{\cdot}3\ \sigma^4$$

Hence $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$ and $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3,$

the results which have already been obtained in (8·9).

8·2·8. A linear combination of independent normal variates is also a normal variate. Let X_i, $(i = 1, 2, \ldots, n)$ be n independent normal variates with mean μ_i and variance σ_i^2 respectively. Then

$$M_{X_i}(t) = \exp\left\{\mu_i t + (t^2\sigma_i^2/2)\right\} \quad \ldots(8\cdot14)$$

The m.g.f. of their linear combination $\sum_{i=1}^{n} a_i X_i$, where $a_1, a_2, \ldots, a_n$ are constants, is given by

$$M_{\Sigma a_i X_i}(t) = M_{a_1 X_1 + a_2 X_2 + \ldots + a_n X_n}(t)$$

$$= M_{a_1 X_1}(t) \cdot M_{a_2 X_2}(t) \ldots M_{a_n X_n}(t)$$

($\because X_i$'s are independent)

$$= M_{X_1}(a_1 t) \cdot M_{X_2}(a_2 t) \ldots M_{X_n}(a_n t) \quad \ldots(8\cdot15)$$

[$\because M_{cX}(t) = M_X(ct)$]

From (8·14), we have

$$M_{X_i}(a_i t) = e^{\mu_i a_i t + t^2 a_i^2 \sigma_i^2/2}$$

$\therefore$ (8·15), gives

$$M_{\sum a_i X_i}(t) = [\, e^{\mu_1 a_1 t + t^2 a_1^2 \sigma_1^2/2} \times e^{\mu_2 a_2 t + t^2 a_2^2 \sigma_2^2/2} \times \ldots \times e^{\mu_n a_n t + t^2 a_n^2 \sigma_n^2/2} \,]$$

$$= \exp\left[\left(\sum_{i=1}^{n} a_i \mu_i \right) t + t^2 \left(\sum_{i=1}^{n} a_i^2 \sigma_i^2 \right) \Big/ 2 \right],$$

which is the m.g.f. of a normal variate with mean $\sum_{i=1}^{n} a_i \mu_i$ and variance $\sum_{i=1}^{n} a_i^2 \sigma_i^2$. Hence by uniqueness theorem of m.g.f.,

$$\sum_{i=1}^{n} a_i X_i \sim N\left[\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right]. \quad \ldots(8\cdot15\,a)$$

Remarks 1. If we take $a_1 = a_2 = 1$, $a_3 = a_4 = \ldots = 0$, then

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

If we take $a_1 = 1$, $a_2 = -1$, $a_3 = a_4 = \ldots = 0$, then

$$X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

Thus we see that the sum as well as the difference of two independent normal variates is also a normal variate. This result provides a sharp contrast to the Poisson distribution, in which case though the sum of two independent Poisson variates is a Poisson variate, the difference is not a Poisson variate.

2. If we take

$a_1 = a_2 = \ldots = a_n = 1$, then we get $\quad \ldots(8\cdot15\,b)$

$$\sum_{i=1}^{n} X_i \sim N\left[\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2 \right]$$

i.e., the sum of independent normal variates is also a normal variate, which establishes the *additive property* of the normal distribution.

3. If X_i; $i = 1, 2, \ldots, n$ are identically and independently distributed as $N(\mu, \sigma^2)$ and if we take $a_1 = a_2 = \ldots = a_n = 1/n$,

then $$\frac{1}{n} \sum_{i=1}^{n} X_i \sim N\left\{ \frac{1}{n} \sum_{i=1}^{n} \mu, \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 \right\}$$

$$\Rightarrow \qquad \bar{X} \sim N(\mu, \sigma^2/n), \text{ where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This leads to the following important conclusion :

If X_i, $(i = 1, 2, ..., n)$, are identically and independently distributed normal variates with mean μ and variance σ^2, then their mean $\bar{X}$ is also $N(\mu, \sigma^2/n)$.

8·2·9. Points of Inflexion of Normal Curve. At the point of inflexion of the normal curve, we should have

$$f''(x) = 0, \text{ and } f'''(x) \neq 0$$

For normal curve, we have from (8·6)

$$f''(x) = -\frac{f(x)}{\sigma^2}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right],$$

$$\therefore \qquad f''(x) = 0 \Rightarrow 1 - \frac{(x-\mu)^2}{\sigma^2} = 0 \Rightarrow x = \mu \pm \sigma$$

It can be easily verified that at the points $x = \mu \pm \sigma, f'''(x) \neq 0$.

Hence the points of inflexion of the normal curve are given by $x = \mu \pm \sigma$ and $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2}$ *i.e.*, they are equi-distant (at a distance σ) from the mean.

8·2·10. Mean Deviation from the Mean for Normal Distribution.

$$\text{M.D. (about mean)} = \int_{-\infty}^{\infty} |x - \mu| f(x)\, dx$$

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

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |z| e^{-z^2/2} dz \qquad \left[\frac{x-\mu}{\sigma} = z\right]$$

$$= \frac{2\sigma}{\sqrt{2\pi}} \cdot \int_{0}^{\infty} |z| e^{-z^2/2} dz,$$

since the integrand $|z| e^{-z^2/2}$ is an even function of z.

Since in $[0, \infty]$, $|z| = z$, we have

$$\text{M.D. (about mean)} = \sqrt{2/\pi}\ \sigma \int_{0}^{\infty} z e^{-z^2/2} dz$$

$$= \sqrt{2/\pi}\ \sigma \int_{0}^{\infty} e^{-t} dt \qquad \left[\frac{z^2}{2} = t\right]$$

$$= \sqrt{2/\pi}\ \sigma \left| \frac{e^{-t}}{-1} \right|_0^\infty$$

$$= \sqrt{2/\pi}\ \sigma$$

$$= \frac{4}{5}\sigma \text{ (approx.)}$$

8·2·11. Area Property (Normal Probability Integral). If $X \sim N(\mu, \sigma^2)$, then the probability that random value of X will lie between $X = \mu$ and $X = x_1$ is given by

$$P(\mu < X < x_1) = \int_\mu^{x_1} f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_\mu^{x_1} e^{-(x-\mu)^2/(2\sigma^2)}\,dx$$

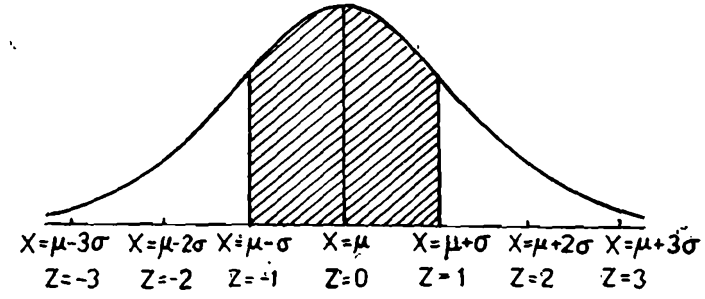
Put $\frac{X-\mu}{\sigma} = Z$, *i.e.*, $X - \mu = \sigma Z$

When $X = \mu$, $Z = 0$ and when $X = x_1$, $Z = \frac{x_1 - \mu}{\sigma} = z_1$, (say).

$$\therefore\ P(\mu < X < x_1) = P(0 < Z < z_1) = \frac{1}{\sqrt{2\pi}} \int_0^{z_1} e^{-z^2/2}\,dz = \int_0^{z_1} \varphi(z)\,dz$$

where $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, is the probability function of standard normal variate.

The definite integral $\int_0^{z_1} \varphi(z)\,dz$ is known as *normal probability integral* and gives the area under standard normal curve between the ordinates at $Z = 0$ and $Z = z_1$. These areas have been tabulated for different values of z_1, at intervals of 0·01 [c.f. Appendix, Table IV].



In particular, the probability that a random value of X lies in the interval $(\mu - \sigma, \mu + \sigma)$ is given by

$$P(\mu - \sigma < X < \mu + \sigma) = \int_{\mu-\sigma}^{\mu+\sigma} f(x)\,dx$$

$$\Rightarrow \quad P(-1 < Z < 1) = \int_{-1}^{1} \varphi(z)\, dz \qquad \left[z = \frac{x-\mu}{\sigma} \right]$$

$$= 2\int_{0}^{1} \varphi(z)\, dz \qquad \text{(By symmetry)}$$

$$= 2 \times 0{\cdot}3413 = 0{\cdot}6826 \qquad \text{(From tables)} \quad ...(8{\cdot}17)$$

Similarly

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = P(-2 < Z < 2) = \int_{-2}^{2} \varphi(z)\, dz$$

$$= 2\int_{0}^{2} \varphi(z)\, dz = 2 \times 0{\cdot}4772 = 0{\cdot}9544 \quad ...(8{\cdot}18)$$

and

$$P(\sigma - 3\sigma < X < \mu + 3\sigma) = P(-3 < Z < 3) = \int_{-3}^{3} \varphi(z)\, dz$$

$$= 2\int_{0}^{3} \varphi(z)\, dz = 2 \times 0{\cdot}49865 = 0{\cdot}9973 \quad ...(8{\cdot}19)$$

Thus the probability that a normal variate X lies outside the range $\mu \pm 3\sigma$ is given by

$$P(|X - \mu| > 3\sigma) = P(|Z| > 3) = 1 - P(-3 \le Z \le 3) = 0{\cdot}0027$$

Thus in all probability, we should expect a normal variate to lie within the range $\mu \pm 3\sigma$, though theoretically, it may range from $-\infty$ to ∞.

Remarks. 1. The total area under normal probability curve is unity, *i.e.*,

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{-\infty}^{\infty} \varphi(z)\, dz = 1$$

2. Since in the normal probability tables, we are given the areas under standard normal curve, in numerical problems we shall deal with the standard normal variate Z rather than the variable X itself.

3. If we want to find area under normal curve, we will somehow or other try to convert the given area to the form $P(0 < Z < z_1)$, since the areas have been given in this form in the tables.

8·2·12. Error Function. If $X \sim N(0, \sigma^2)$, then

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

If we take $h^2 = \dfrac{1}{2\sigma^2}$ then $f(x) = \dfrac{h}{\sqrt{\pi}} e^{-h^2 x^2}$

The probability that a random value of the variate lies in the range $\pm x$ is given by

$$P = \int_{-\lambda}^{x} f(x)\,dx = \frac{h}{\sqrt{\pi}} \int_{-\lambda}^{\lambda} e^{-h^2x^2}\,dx$$

$$= \frac{2h}{\sqrt{\pi}} \int_{0}^{x} e^{-h^2x^2}\,dx = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-h^2x^2}\,(h\,dx) \qquad ...(*)$$

Taking

$$\psi(y) = \frac{2}{\sqrt{\pi}} \int_{0}^{y} e^{-y^2}\,dy, \text{ (*) may be re-written as}$$

$$P = \psi(hx) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-h^2x^2}\,(h\,dx) \qquad ...(**)$$

The function $\psi(y)$, known as the *error function*, is of fundamental importance in the *theory of errors* in Astronomy.

8·2·3. Importance of Normal Distribution. Normal distribution plays a very important role in statistical theory because of the following reasons :

(*i*) Most of the distributions occurring in practice, *e.g.*, Binomial, Poisson, Hypergeometric distributions, etc., can be approximated by normal distribution. Moreover, many of the sampling distributions, *e.g.*, Student's '*t*', Snedecor's F, Chi-square distributions, etc., tend to normality for large samples.

(*ii*) Even if a variable is not normally distributed, it can sometimes be brought to normal form by simple transformation of variable. For example, if the distribution of X is skewed, the distribution of $\sqrt{X}$ might come out to be normal [c.f. Variate Transformations at the end of this Chapter].

(*iii*) If $X \sim N(\mu, \sigma^2)$, then

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0{\cdot}9973$$

$$\Rightarrow \quad P(-3 < Z < 3) = 0{\cdot}9973$$

$$\Rightarrow \quad P(|Z| < 3) = 0{\cdot}9973$$

$$\Rightarrow \quad P(|Z| > 3) = 0{\cdot}0027$$

This property of the normal distribution forms the basis of entire *Large Sample* theory.

(*iv*) Many of the distributions of sample statistic (*e.g.*, the distributions of sample mean, sample variance, etc.) tend to normality for large samples and as such they can best be studied with the help of the normal curves.

(*v*) The entire theory of small sample tests, *viz.*, t, F, χ^2 tests-etc., is based on the fundamental assumption that the parent populations from which the samples have been drawn follow normal distribution.

(*vi*) Theory of normal curves can be applied to the graduation of the curves which are not normal.

(*vii*) Normal distribution finds large applications in Statistical Quality Control in industry for setting control limits.

The following quotation due to Lipman rightly reveals the popularity and importance of normal distribution.

"Every body believes in the law of errors (the normal curve), the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is experimental fact."

W. J. Youden of the National Bureau of Standards describes the importance of the Normal distribution artistically in the following words :

THE NORMAL
LAW OF ERRORS
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALISATIONS OF NATURAL
PHILOSOPHY. IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES,
IN THE PHYSICAL AND SOCIAL SCIENCES
AND IN MEDICINE, AGRICULTURE AND
ENGINEERING. IT IS AN INDISPENSABLE TOOL FOR
THE ANALYSIS AND THE INTERPRETATION OF THE
BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT.

The above presentation, strikingly enough, gives the shape of the normal probability curve.

8·2 14. Fitting of Normal Distribution. In order to fit normal distribution to the given data we first calculate the mean μ, (say), and standard deviation σ, (say), from the given data. Then the normal curve fitted to the given data is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(x-\mu)^2/2\sigma^2\right], \quad -\infty < x < \infty$$

To calculate the expected normal frequencies we first find the standard normal variates corresponding to the lower limits of each of the class intervals, *i.e.*, we compute $z_i = \frac{x_i' - \mu}{\sigma}$, where x_i' is the lower limit of the ith class interval. Then the areas under the normal curve to the left of the ordinate at $z = z_i$, say, $\varphi(z_i)$ are computed from the tables. Finally, the areas for the successive class intervals are obtained by subtraction, *viz.*, $\varphi(z_{i+1}) - \varphi(z_i)$, $(i = 1, 2, \ldots)$ and on multiplying these areas by N, we get the expected normal frequencies.

Example 8·10. *Obtain the equation of the normal curve that may be fitted to the following data :*

Class.	60–65	65–70	70–75	75–80	80–85	85–90	90–95	95–100
Frequency.	3	21	150	335	326	135	26	4

Also obtain the expected normal frequencies

Solution. For the given data, we have

$N = 1000$, $\mu = 79.945$ and $\sigma = 5.545$

Hence the equation of the normal curve fitted to the given data is

$$f(x) = \frac{1000}{\sqrt{2\pi} \times 5.545} \exp\left\{-\frac{1}{2}\left(\frac{x - 79.945}{5.545}\right)^2\right\}$$

Theoretical normal frequencies can be obtained as follows :

class	Lower class boundry (X')	$Z = \frac{X' - \mu}{\sigma}$	$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-z^2/2} dz$	$\Delta\phi(z) = \psi_{z+1} - \psi_z$	Expected frequency $N\Delta\psi(z)$
Below 60	$-\infty$	$-\infty$	0	0.000112	0.12 ≅ 0
60–65	60	−3.663	0.000112	0.002914	2.914 ≅ 3
65–70	65	−2.745	0.003026	0.031044	31.044 ≅ 31
70–75	70	−1.826	0.034070	0.147870	147.870 ≅ 148
75–80	75	−0.908	0.181940	0.322050	322.050 ≅ 322
80–85	80	0.010	0.503990	0.319300	319.300 ≅ 319
85–90	85	0.928	0.823290	0.144072	144.072 ≅ 144
90–95	90	1.487	0.967362	0.029792	29.792 ≅ 30
95–100	95	2.675	0.997154	0.002733	2.733 ≅ 3
100 and over	100	3.683	0.999887		
Total					1000

Example 8.11. *For a certain normal distribution, the first moment about 10 is 40 and the fourth moment about 50 is 48. What is the arithmetic mean and standard deviation of the distribution ?*

[Delhi Univ. B.Sc. (Hons. Subs.), 1987; Allahabad Univ. B.Sc. 1990]

Solution. We know that if μ_1' is the first moment about the point $X = A$, then arithmetic mean is given by:

$$\text{Mean} = A + \mu_1'$$

We are given

μ_1' (about the point $X = 10$) $= 40 \Rightarrow$ Mean $= 10 + 40 = 50$

Also we are given

μ_4' (about the point $X = 50$) $= 48$, *i.e.*, $\mu_4 = 48$ $(\because$ Mean $= 50)$

But for a normal distribution with standard deviation σ,

$$\mu_4 = 3\sigma^4 \Rightarrow 3\sigma^4 = 48 \quad i.e., \quad \sigma = 2$$

Example 8.12. *X is normally distributed and the mean of X is 12 and S.D. is 4. (a) Find out the probability of the following :*

(i) $X \geq 20$, *(ii)* $X \leq 20$, and *(iii)* $0 \leq X \leq 12$

(b) Find x', when $P(X > x') = 0.24$.

(c) Find x_0' and x_1', when $P(x_0' < X < x_1') = 0.50$ and $P(X > x_1') = 0.25$

Solution. *(a)* We have $\mu = 12$, $\sigma = 4$, *i.e.*, $X \sim N(12, 16)$.

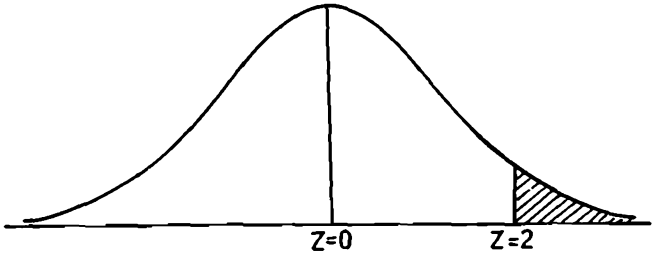
(*i*) $P(X \geq 20) = ?$

When $X = 20$, $\quad Z = \dfrac{20 - 12}{4} = 2$

$\therefore \quad P(X \geq 20) = P(Z \geq 2) = 0{\cdot}5 - P(0 \leq Z \leq 2) = 0{\cdot}5 - 0{\cdot}4772 = 0{\cdot}0228$

(*ii*) $P(X \leq 20) = 1 - P(X \geq 20)$ $\qquad$ ($\because$ Total probability = 1)

$= 1 - 0{\cdot}0228 = 0{\cdot}9772$



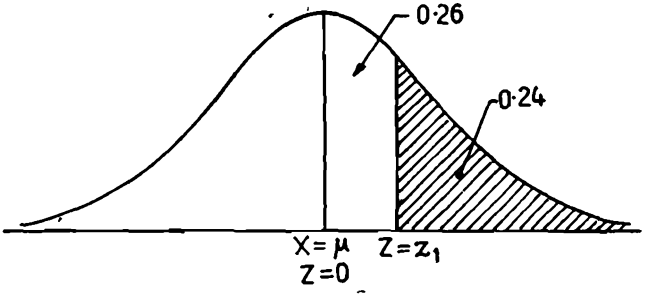
(*iii*) $P(0 \leq X \leq 12) = P(-3 \leq Z \leq 0)$ $\qquad \left(Z = \dfrac{X - 12}{4}\right)$

$= P(0 \leq Z \leq 3) = 0{\cdot}49865$ $\qquad$ (From symmetry)

(*b*) When $X = x'$, $Z = \dfrac{x' - 12}{4} = z_1$ (say)

then, we are given

$P(X > x') = 0{\cdot}24 \quad \Rightarrow \quad P(Z > z_1) = 0{\cdot}24$, *i.e.*, $P(0 < Z < z_1) = 0{\cdot}26$



$\therefore$ From normal tables,

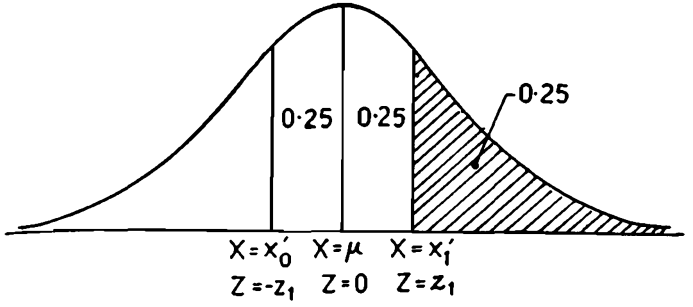
$z_1 = 0{\cdot}71$ (approx.)

Hence $\quad \dfrac{x_1' - 12}{4} = 0{\cdot}71 \quad \Rightarrow \quad x_1' = 12 + 4 \times 0{\cdot}71 = 14{\cdot}84$

(*c*) We are given

$P(x_0' < X \leq x_1') = 0{\cdot}50$ and $P(X > x_1') = 0{\cdot}25$ $\qquad$ ...(*)

From (*), obviously the points x_0' and x_1' are located as shown in the figure.



When $X = x_1'$, $\quad Z = \dfrac{x_1' - 12}{4} = z_1$ (say)

and when $X = x_0'$, $\quad Z = \dfrac{x_0' - 12}{4} = -z_1$ (It is obvious from the figure)

We have

$$P(Z > z_1) = 0{\cdot}25 \quad \Rightarrow \quad P(0 < Z < z_1) = 0{\cdot}25$$

$\therefore \quad z_1 = 0{\cdot}67$ (From tables)

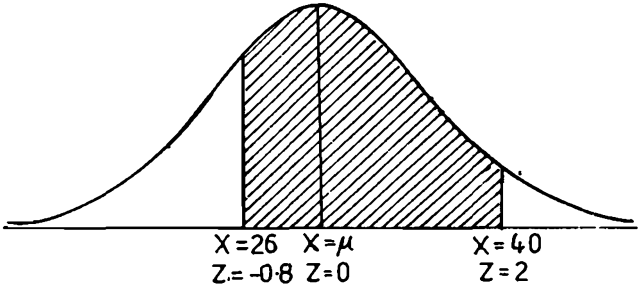
Hence $\dfrac{x_1' - 12}{4} = 0{\cdot}67 \quad \Rightarrow \quad x_1' = 12 + 4 \times 0\,67 = 14{\cdot}68$

and $\dfrac{x_0' - 12}{4} = -0{\cdot}67 \quad \Rightarrow \quad x_0' = 12 - 4 \times 0{\cdot}67 = 9{\cdot}32$

Example 8·13. *X is a normal variate with mean 30 and S.D. 5. Find the probabilities that*

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

Solution. Here $\mu = 30$ and $\sigma = 5$.

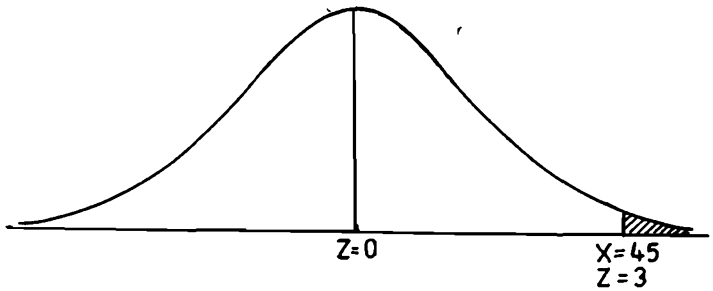


(*i*) When $X = 26$, $Z = \dfrac{X - \mu}{\sigma} = \dfrac{26 - 30}{5} = -0{\cdot}8$

and when $X = 40,\ Z = \dfrac{40 - 30}{5} = 2$

$$\therefore \quad P(26 \le X \le 40) = P(-0{\cdot}8 \le Z \le 2)$$
$$= P(-0{\cdot}8 \le Z \le 0) + P(0 \le Z \le 2)$$
$$= P(-0{\cdot}8 \le Z \le 0) + 0{\cdot}4772 \qquad \text{(From tables)}$$
$$= P(0 \le Z \le 0{\cdot}8) + 0{\cdot}4772 \qquad \text{(From symmetry)}$$
$$= 0{\cdot}2881 + 0{\cdot}4772 = 0{\cdot}7653$$

$P(X \ge 45) = ?$



When $X = 45, \qquad Z = \dfrac{45 - 30}{5} = 3$

$$\therefore \quad P(X \ge 45) = P(Z \ge 3) = 0{\cdot}5 - P(0 \le Z \le 3)$$
$$= 0{\cdot}5 - 0{\cdot}49865 = 0{\cdot}00135$$

(iii) $P(|X - 30| \le 5) = P(25 \le X \le 35) = P(-1 \le Z \le 1)$
$$= 2P(0 \le Z \le 1) = 2 \times 0{\cdot}3413 = 0{\cdot}6826$$

$$\therefore \quad P(|X - 30| > 5) = 1 - P(|X - 30| \le 5)$$
$$= 1 - 0{\cdot}6826 = 0{\cdot}3174$$

Example 8·14. *The mean yield for one-acre plot is 662 kilos with a s.d. 32 kilos. Assuming normal distribution, how many one-acre plots in a batch of 1,000 plots would you expect to have yield (i) over 700 kilos, (ii) below 650 kilos, and (iii) what is the lowest yield of the best 100 plots?*

Solution. If the r.v. X denotes the yield (in kilos) for one-acre plot, then we are given that $X \sim N(\mu, \sigma^2)$, where $\mu = 662$ and $\sigma = 32$.

(i) The probability that a plot has a yield over 700 kilos is given by

$$P(X > 700) = P(Z > 1{\cdot}19); \qquad Z = \frac{X - 662}{32}$$
$$= 0{\cdot}5 - P(0 \le Z \le 1{\cdot}19)$$
$$= 0{\cdot}5 - 0{\cdot}3830$$
$$= 0{\cdot}1170$$

Hence in a batch of 1,000 plots, the expected number of plots with yield over 700 kilos is $1{,}000 \times 0{\cdot}117 = 117$.

(ii) Required number of plots with yield below 650 kilos is given by

$$1000 \times P(X < 650) = 1000 \times P(Z < -0{\cdot}38) \qquad \left[Z = \frac{650 - 662}{32} \right]$$
$$= 1000 \times P(Z > 0{\cdot}38) \quad \text{(By symmetry)}$$
$$= 1000 \times [\, 0{\cdot}5 - P(0 \le Z \le 0{\cdot}38)\,]$$
$$= 1000 \times [\, 0{\cdot}5 - 0{\cdot}1480\,] = 1000 \times 0{\cdot}352$$
$$= 352$$

(*iii*) The lowest yield, say, x_1 of the best 100 plots is given by

$$P(X > x_1) = \frac{100}{1000} = 0{\cdot}1$$

When $\quad X = x_1,\ Z = \dfrac{x_1 - \mu}{\sigma} = \dfrac{x_1 - 662}{32} = z_1$ (say) $\qquad \ldots(*)$

such that $\quad P(Z > z_1) = 0{\cdot}1 \quad \Rightarrow \quad P(0 \le Z \le z_1) = 0{\cdot}4$

$\Rightarrow \quad z_1 = 1{\cdot}28$ (approx.) [From Normal Probability Tables]

Substituting in (*), we get

$$x_1 = 662 + 32z = 662 + 32 \times 1{\cdot}28$$
$$= 662 + 40{\cdot}96 = 702{\cdot}96$$

Hence the best 100 plots have yield over 702·96 kilos.

Example 8·15. *There are six hundred Economics students in the post-graduate classes of a university, and the probability for any student to need a copy of a particular book from the university library on any day is* 0·05. *How many copies of the book should be kept in the university library so that the probability may be greater than 0·90 that none of the students needing a copy from the library has to come back disappointed? (Use normal approximation to the binomial distribution).* **[Delhi Univ. M.A. (Eco.), 1989]**

Solution. We are given :

$$n = 600, p = 0{\cdot}05, \mu = np = 600 \times 0{\cdot}05 = 30$$
$$\sigma^2 = npq = 600 \times 0{\cdot}05 \times 0{\cdot}95 = 28{\cdot}5 \Rightarrow \sigma = \sqrt{28{\cdot}5} = 5{\cdot}3$$

We want x_1 such that

$$P(X < x_1) > 0{\cdot}90$$
$$\Rightarrow \quad P(Z < z_1) > 0{\cdot}90 \qquad \left[z_1 = \frac{x_1 - 30}{5{\cdot}3} \right]$$
$$\Rightarrow \quad P(0 < Z < z_1) > 0{\cdot}40$$
$$\Rightarrow \quad z_1 > 1{\cdot}28 \qquad \text{[From Normal Probability Tables]}$$
$$\Rightarrow \quad \frac{x_1 - 30}{5{\cdot}3} > 1{\cdot}28 \quad \Rightarrow \quad x_1 > 30 + 5{\cdot}3 \times 1{\cdot}28$$
$$\Rightarrow \quad x_1 > 30 + 6{\cdot}784 \quad \Rightarrow \quad x_1 > 36{\cdot}784 \simeq 37$$

Hence the university library should keep at least 37 copies of the book.

Example 8·16. *The marks obtained by a number of students for a certain subject are assumed to be approximately normally distributed with mean value 65*

and with a standard deviation of 5. If 3 students are taken at random from this set what is the probability that exactly 2 of them will have marks over 70?

Solution. Let the r.v. X denote the marks obtained by the given set of students in the given subject. Then we are given that $X \simeq N(\mu, \sigma^2)$ where $\mu = 65$ and $\sigma = 5$. The probability 'p' that a randomly selected student from the given set gets marks over 70 is given by

$$p = P(X > 70)$$

When $X = 70$, $Z = \dfrac{X - \mu}{\sigma} = \dfrac{70 - 65}{5} = 1$.

$$\therefore \quad p = P(X > 70) = P(Z > 1)$$
$$= 0{\cdot}5 - P(0 \le Z \le 1)$$
$$= 0{\cdot}5 - 0{\cdot}3413 = 0{\cdot}1587 \quad \text{[From Normal probability tables]}$$

Since this probability is same for each student of the set, the required probability that 'out of 3 students selected at random from the set, exactly 2 will have marks over 70, is given by the binomial probability law:

$${}^3C_2\, p^2 \,.\, (1-p) = 3 \times (0{\cdot}1587)^2 \times (0{\cdot}8413) = 0{\cdot}06357$$

Example 8·17. (*a*) *If $\log_{10} X$ is normally distributed with meant 4 and variance 4, find the probability of*

$$1{\cdot}202 < X < 83180000$$

(Given $\log_{10}$ 1202 = 3·08, $\log_{10}$ 8318 = 3·92).

(b) $\log_{10} X$ is normally distributed with mean 7 and variance 3, $\log_{10} Y$ is normally distributed with mean 3 and variance unity. If the distributions of X and Y are independent, find the probability of $1{\cdot}202 < (X/Y) < 83180000$.

[Given $\log_{10}$ (1202) = 3·08, $\log_{10}$ (8318) = 3·92]

Solution. (*a*) Since $\log X$ is a non-decreasing function of X, we have

$$P(1{\cdot}202 < X < 83180000) = P(\log_{10} 1{\cdot}202 < \log_{10} X < \log_{10} 83180000)$$
$$= P(0{\cdot}08 < \log_{10} X < 7{\cdot}92)$$
$$= P(0{\cdot}08 < Y < 7{\cdot}92)$$

where $Y = \log_{10} X \sim N(4, 4)$ (given).

When $Y = 0{\cdot}08$, $Z = \dfrac{0{\cdot}08 - 4}{2} = -1{\cdot}96$

and when $Y = 7{\cdot}92$, $Z = \dfrac{7{\cdot}92 - 4}{2} = 1{\cdot}96$

$\therefore$ Required probability $= P(0{\cdot}08 < Y < 7{\cdot}92)$
$= P(-1{\cdot}96 < Z < 1{\cdot}96) = 2\,P(0 < Z < 1{\cdot}96)$ (By symmetry)
$= 2 \times 0{\cdot}4750 = 0{\cdot}9500$

(*b*) $P[\, 1{\cdot}202 < (X/Y) < 83180000 \,]$
$= P[\, \log_{10} 1{\cdot}202 < \log_{10} (X/Y) < \log_{10} 83180000 \,]$
$= (0{\cdot}08 < U < 7{\cdot}92)$

where $U = \log_{10} (X/Y) = \log_{10} X - \log_{10} Y$

Since $\log_{10} X \sim N(7, 3)$ and $\log_{10} Y \sim N(3, 1)$, are independent,

$$\log_{10} X - \log_{10} Y \sim N(7-3, 3+1) \qquad \text{(c.f. Remark 1, § 8·2·8)}$$

$$\Rightarrow \quad U = (\log_{10} X - \log_{10} Y) \sim N(4, 4)$$

$\therefore$ Required probability is given by

$$p = P(0{\cdot}08 < U < 7{\cdot}92), \text{ where } U \sim N(4, 4)$$

$$= 0{\cdot}95 \qquad \text{[See part (a)]}$$

Example 8·18. *Two independent random variates X and Y are both normally distributed with means 1 and 2 and standard deviations 3 and 4 respectively. If Z = X − Y, write the probability density function of Z. Also state the median, s.d. and mean of the distribution of Z. Find Prob.* $\{Z + 1 \leq 0\}$.

Solution. Since $X \sim N(1, 9)$ and $Y \sim N(2, 16)$ are independent, $Z = Y - Y \sim N(1-2, 9+16)$, *i.e.*, $Z = X - Y \sim N(-1, 25)$. Hence p.d.f. of Z is

$$p(z) = \frac{1}{5\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}\left(\frac{z+1}{5}\right)^2\right]; -\infty < z < \infty.$$

For the distribution of Z,

$$\text{Median} = \text{Mean} = -1 \quad \text{and} \quad \text{s.d.} = \sqrt{25} = 5$$

$$P(Z + 1 \leq 0) = P(Z \leq -1)$$

$$= P(U \leq 0); \qquad \left[U = \frac{Z+1}{5} \sim N(0, 1)\right]$$

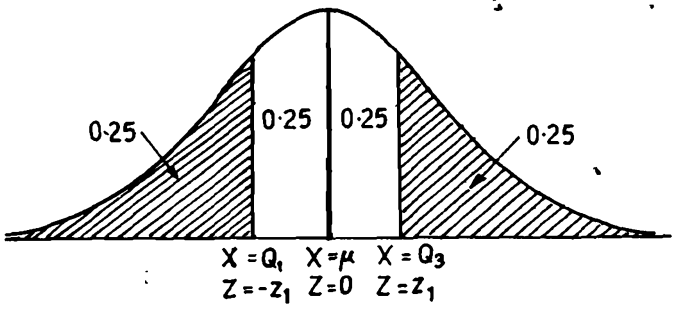
$$= 0{\cdot}5$$

Example 8·19. *Prove that for the normal distribution, the quartile deviation, the mean deviation and standard deviation are approximately 10 : 12 : 15.* **[Dibrugarh Univ. B.Sc. 1993]**

Solution. Let X be a $N(\mu, \sigma^2)$. If Q_1 and Q_3 are the first and third quartiles respectively, then by definition

$$P(X < Q_1) = 0{\cdot}25 \quad \text{and} \quad P(X > Q_3) = 0{\cdot}25$$

The points Q_1 and Q_3 are located as shown in the figure given below.



When $X = Q_3,\ Z = \dfrac{Q_3 - \mu}{\sigma} = z_1$, (say),

and when $X = Q_1,\ Z = \dfrac{Q_1 - \mu}{\sigma} = -z_1$ (This is obvious from the figure)

Subtracting, we have

$$\frac{Q_3 - Q_1}{\sigma} = 2z_1$$

The quartile deviation is given by

$$Q.D. = \frac{Q_3 - Q_1}{2} = \sigma z_1$$

From the figure, obviously, we have

$P(0 < Z < z_1) = 0{\cdot}25 \Rightarrow z_1 = 0{\cdot}67$ (approx.) (From Normal Tables)

$\therefore \quad Q.D. = \sigma z_1 = 0{\cdot}67\,\sigma = \frac{2}{3}\sigma$

For normal distribution mean deviation about mean (*c.f.* § 8·2·10) is given by

$$\text{M.D.} = \sqrt{2/\pi}\,\sigma = \frac{4}{5}\sigma$$

Hence Q.D. : M.D. : S.D. : : $\frac{2}{3}\sigma : \frac{4}{5}\sigma : \sigma :: \frac{2}{3} : \frac{4}{5} : 1 : : 10 : 12 : 15$

Example 8·20 (*a*). *In a distribution exactly normal, 7% of the items are under 35 and 89% are under 63. What are the mean and standard deviation of the distribution ?* **[Kerala Univ. B.Sc., May 1991]**

(*b*) *Of a large group of men, 5% are under 60 inches in height and 40% are between 60 and 65 inches. Assuming a normal distribution, find the mean height and standard deviation.* **[Nagpur Univ. B.Sc., 1992]**

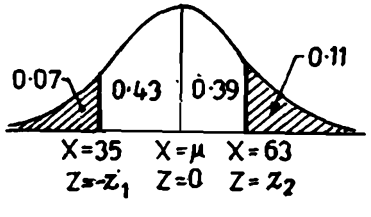
Solution. If $X \sim N(\mu, \sigma^2)$, then we are given

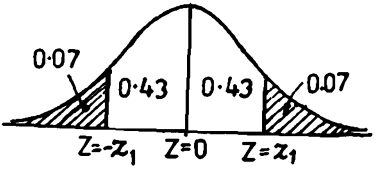
$$P(X < 63) = 0{\cdot}89 \quad \Rightarrow \quad P(X > 63) = 0{\cdot}11 \text{ and } P(X < 35) = 0{\cdot}07$$

The points $X = 63$ and $X = 35$ are located as shown in Fig. (*i*) below.

Since the value $X = 35$ is located to the left of the ordinate at $X = \mu$, the corresponding value of Z is negative.

When $X = 35,\ Z = \dfrac{35 - \mu}{\sigma} = -z_1$, (say),





and when $X = 63$, $Z = \dfrac{63 - \mu}{\sigma} = z_2$, (say),

Thus we have, as is obvious from figures (*i*) and (*ii*)

$$P(0 < Z < z_2) = 0.39 \text{ and } P(0 < Z < z_1) = 0.43$$

Hence from normal tables, we have

$$z_2 = 1.23 \text{ and } z_1 = 1.48$$

$$\therefore \quad \frac{63 - \mu}{\sigma} = 1.23 \text{ and } \frac{35 - \mu}{\sigma} = -1.48$$

Subtracting, we get

$$\frac{28}{\sigma} = 2.71 \Rightarrow \sigma = \frac{28}{2.71} = 10.33$$

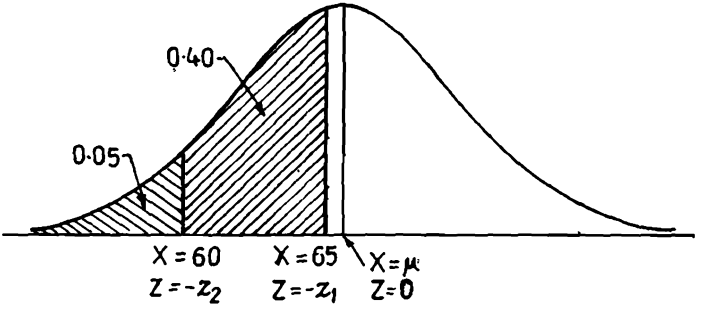
$$\therefore \quad \mu = 35 + 1.48 \times 10.33 = 35 + 15.3 = 50.3$$

(*b*) We are given

$$P(X < 60) = 0.05 \text{ and } P(60 < X < 65) = 0.40$$

i.e., $P(X < 65) = 0.45$

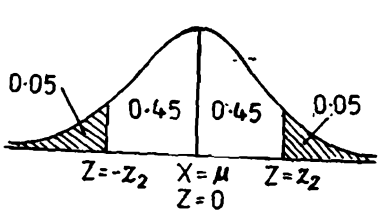
Since the total area to the left of the ordinate at $X = \mu$ is 0.5, both the points $X = 60$ and $X = 65$ are located to the left of $X = \mu$ and consequently the corresponding values of Z are negative.

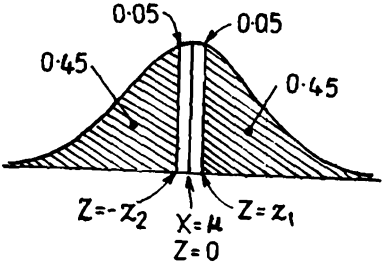


Let $X \sim N(\mu, \sigma^2)$.

When $X = 65$, $\quad Z = \dfrac{65 - \mu}{\sigma} = -z_1$ (say),

and when $X = 60$, $\quad Z = \dfrac{60 - \mu}{\sigma} = -z_2$ (say).





Thus we have

$$P(0 < Z < z_2) = 0{\cdot}45 \text{ and } P(0 < Z < z_1) = 0{\cdot}05$$

$\therefore$ $z_2 = 1{\cdot}645$ and $z_1 = 0{\cdot}13$ (approx.) (From Normal Tables)

Hence $\dfrac{60-\mu}{\sigma} = -1{\cdot}645$ ...(*) ; and $\dfrac{65-\mu}{\sigma} = -0{\cdot}13$ ...(**)

Dividing, we get $\dfrac{60-\mu}{65-\mu} = \dfrac{1{\cdot}645}{0{\cdot}13} \Rightarrow \mu = \dfrac{19825}{303} = 65{\cdot}42$

$\therefore$ From (*), we have $\sigma = \dfrac{60-65{\cdot}42}{-1{\cdot}645} = 3{\cdot}29$

Remarks. If we substitute the value of μ in (**), we get $\sigma = 3{\cdot}23$ which is only an approximate value since the value of $z_1 = 0{\cdot}13$, seen from the table, is not exact but only approximate. On the other hand, the value of $z_2 = 1{\cdot}645$ is exact and hence use of (*) for estimating σ gives better approximation.

Example 8·21 *If the skulls are classified as A, B and C according as the length-breadth index is under 75, between 75 and 80, or over 80, find approximately (assuming that the distribution is normal) the mean and standard deviation of a series in which A are 58%, B are 38% and C are 4%, being given that if*

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_0^t \exp(-x^2/2)\, dx,$$

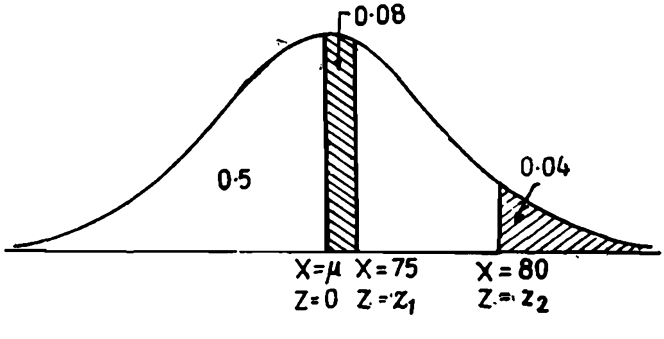
then $f(0{\cdot}20) = 0{\cdot}08$ and $f(1{\cdot}75) = 0{\cdot}46$

[Delhi Univ. B.Sc., 1989; Burdwan Univ. B.Sc., 1990]

Solution. Let the length-breadth index be denoted by the variable X, then we are given

$$P(X < 75) = 0{\cdot}58 \text{ and } P(X > 80) = 0{\cdot}04 \qquad ...(1)$$

Since $P(X < 75)$ represents the total area to the left of the ordinate at the point $X = 75$ and $P(X > 80)$ represents the total area to the right of the ordinate at the point $X = 80$, it is obvious from (1) that the points $X = 75$ and $X = 80$ are located at the positions shown in the figure below.



Now $\frac{1}{\sqrt{2\pi}} \int_0^t \exp(-x^2/2)\,dx$ represents the area under standard normal curve between the ordinates at $Z = 0$ and $Z = t$, Z being a $N(0, 1)$ variate.

$$\therefore \quad f(t) = \frac{1}{\sqrt{2\pi}} \int_0^t \exp(-x^2/2)\,dx = P(0 < Z < t)$$

Hence $f(0{\cdot}20) = P(0 < Z < 0{\cdot}20) = 0{\cdot}08$ ...(2)

and $f(1{\cdot}75) = P(0 < Z < 1{\cdot}75) = 0{\cdot}46$

Let μ and σ be the mean and standard deviation of the distribution. Then $X \sim N(\mu, \sigma^2)$.

When $X = 75$, $Z = \frac{75 - \mu}{\sigma} = z_1$ (say),

and when $X = 80$, $Z = \frac{80 - \mu}{\sigma} = z_2$ (say).

Thus from the figure, it is obvious that

$$P(X < 75) = 0{\cdot}58 \Rightarrow P(0 < Z < z_1) = 0{\cdot}08$$

$\therefore$ Using (2), we have

$$z_1 = \frac{75 - \mu}{\sigma} = 0{\cdot}20 \qquad ...(3)$$

Also $P(X > 80) = 0{\cdot}04 \Rightarrow P(0 < Z < z_2) = 0{\cdot}46$

$\therefore$ From (2), we get

$$z_2 = \frac{80 - \mu}{\sigma} = 1{\cdot}75 \qquad ...(4)$$

Solving the equations (3) and (4), we get

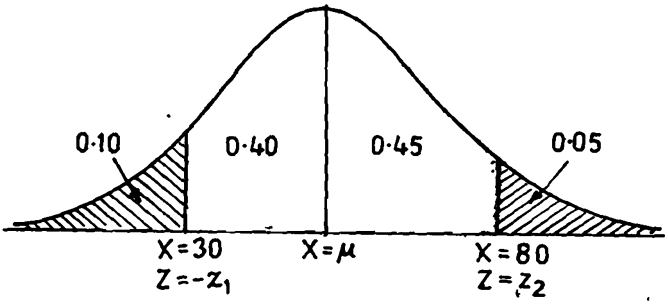
$$\mu = 74{\cdot}4 \text{ (approx.) and } \sigma = 3{\cdot}2 \text{ (approx.)}$$

Example 8·22. *In an examination it is laid down that a student passes if he secures 30 per cent or more marks. He is placed in the first, second or third division according as he secures 60% or more marks, between 45% and 60% marks and marks between 30% and 45% respectively. He gets distinction in case he secures 80% or more marks. It is noticed from the result that 10% of the students failed in the examination, whereas 5% of them obtained distinction. Calculate the percentage of students placed in the second division. (Assume normal distribution of marks.)* **[Aligarh Univ. B.Sc., 1991]**

Solution. Let the variable X denote the marks (out of 100) in the examination and let $X \sim N(\mu, \sigma^2)$. Then we are given

$$P(X < 30) = 0{\cdot}10 \text{ and } P(X \geq 80) = 0{\cdot}05$$

Thus from the figure on next page, we have



When $X = 30$, $\quad Z = \dfrac{30-\mu}{\sigma} = -z_1$ (say),

and when $X = 80$, $\quad Z = \dfrac{80-\mu}{\sigma} = z_2$ (say).

$\therefore \quad P(0 < Z < z_2) = 0{\cdot}5 - 0{\cdot}05 = 0{\cdot}45$

and $\quad P(0 < Z < z_1) = P(-z_1 < Z < 0)$ (By symmetry)

$= 0{\cdot}50 - 0{\cdot}10 = 0{\cdot}40$

$\therefore$ From normal tables, we get

$z_1 = 1{\cdot}28$ and $z_2 = 1{\cdot}64$

Hence $\quad \dfrac{30-\mu}{\sigma} = -1{\cdot}28$

$\Rightarrow \quad \dfrac{\mu-30}{\sigma} = 1{\cdot}28$ and $\dfrac{80-\mu}{\sigma} = 1{\cdot}64$

Adding, we get

$$\frac{50}{\sigma} = 2{\cdot}92 \Rightarrow \sigma = \frac{50}{2{\cdot}92} = 17{\cdot}12$$

$\therefore \quad \mu = 30 + 1{\cdot}28 \times 17{\cdot}12 = 30 + 21{\cdot}9136 = 51{\cdot}9136 \simeq 52$

The probability 'p' that a candidate is placed in the second division is equal to the probability that his score lies between 45 and 60, *i.e.*,

$$p = P(45 < X < 60) = P(-0{\cdot}41 < Z < 0{\cdot}47) \qquad \left[Z = \frac{X-52}{17{\cdot}12} \right]$$

$= P(-0{\cdot}41 < Z < 0) + P(0 < Z < 0{\cdot}47)$

$= P(0 < Z < 0{\cdot}41) + P(0 < Z < 0{\cdot}47)$ (By symmetry)

$= 0{\cdot}1591 + 0{\cdot}1808 = 0{\cdot}3399 = 0{\cdot}34$ (approx.)

Therefore, 34% candiates got second division in the examination.

Example 8·23. *The local authorities in a certain city instal 10,000 electric lamps in the streets of the city. If these lamps have an average life of 1,000 burning hours with a standard deviation of 200 hours, assuming normality, what number of lamps might be expected to fail (i) in the first 800 burning hours? (ii) between 800 and 1,200 burning hours ? After what period of burning hours would you expect that (a) 10% of the lamps would fail? (b) 10% of the lamps would be still burning?*

[In a normal curve, the area between the ordinates corresponding to $\frac{X-\bar{X}}{\sigma}=0$ *and* $\frac{X-\bar{X}}{\sigma}=1$ *is* 0·34134 *and 80% of the area lies between the ordinates corresponding to* $\frac{X-\bar{X}}{\sigma}=\pm 1{\cdot}28$ *].*

Solution. If the variable X denotes the life of a bulb in burning hours, then we are given that $X \sim N(\mu, \sigma^2)$, where $\mu = 1{,}000$ and $\sigma = 200$.

(*i*) The probability 'p' that bulb fails in the first 800 burning hours is given by

$$p = P(X < 800) = P(Z < -1) = P(Z > 1) \qquad \left[Z = \frac{800-1000}{200}\right]$$

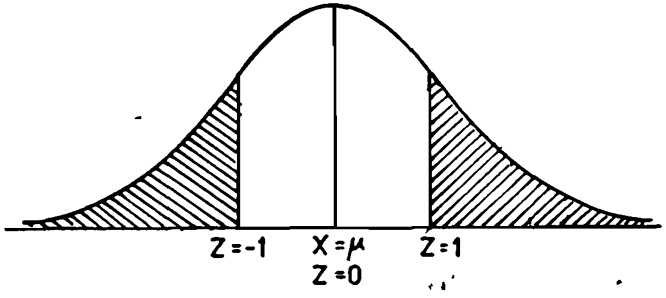
$$= 0{\cdot}5 - P(0 < Z < 1) = 0{\cdot}5 - 0{\cdot}3413 = 0{\cdot}1587$$

Therefore out of 10,000 bulbs, the number of bulbs which fail in the first 800 hours is

$$10{,}000 \times 0{\cdot}1587 = 1587$$

(*ii*) Required probability $= P(800 < X < 1200) = P(-1 < Z < 1)$

$$= 2P(0 < Z < 1) = 2 \times 0{\cdot}3413 = 0{\cdot}6826$$



Hence the expected number of blubs with life between 800 and 1,200 hours of burning life is: $10{,}000 \times 0{\cdot}6826 = 6826$

(*a*) Let 10% of the bulbs fail after x_1 hours ot ourning life. Then we have to find x_1 such that $P(X < x_1) = 0{\cdot}10$

When $X = x_1$, $\quad Z = \frac{x_1 - 1000}{200} = -z_1$ (say).

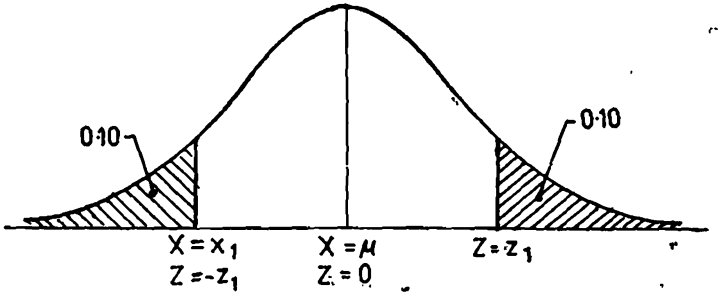
$\therefore \quad P(Z < -z_1) = 0{\cdot}10 \Rightarrow P(Z > z_1) = 0{\cdot}10$

$\Rightarrow \quad P(0 < Z < z_1) = 0{\cdot}40 \qquad ...(1)$

We are given that

$P(-1{\cdot}28 < Z < 1{\cdot}28) = 0{\cdot}80 \Rightarrow 2P(0 < Z < 1{\cdot}28) = 0{\cdot}80$

$\Rightarrow \quad P(0 < Z < 1{\cdot}28) = 0{\cdot}40 \qquad ...(2)$



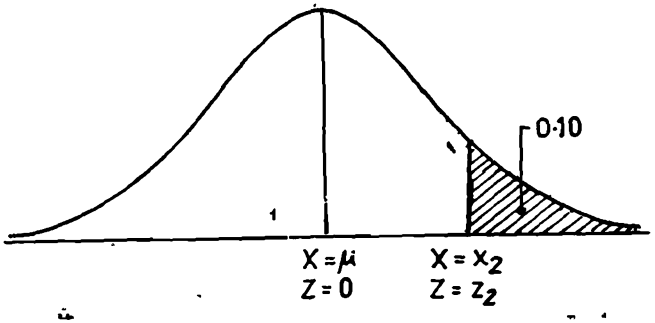
$\therefore$ From (1) and (2), we get

$$z_1 = 1{\cdot}28$$

Hence $\dfrac{x_1 - 1000}{200} = -1{\cdot}28 \Rightarrow x_1 = 1000 - 256 = 744$.

Thus after 744 hours of burning life, 10% of the blubs will fail.

(*b*) Let 10% of the blubs be still burning after, (say), x_2 hours of burning life. Then we have



$$P(X > x_2) = 0{\cdot}10 \Rightarrow P(Z > z_2) = 0{\cdot}10 \qquad \left[z_2 = \frac{x_2 - 1000}{200}\right]$$

$$\Rightarrow \qquad P(0 < Z < z_2) = 0{\cdot}40$$

$$\therefore \qquad z_2 = 1{\cdot}28 \qquad \text{[From (2)]}$$

i.e., $\dfrac{x_2 - 1000}{200} = 1{\cdot}28 \Rightarrow x_2 = 1000 + 256 = 1256$

Hence after 1256 hours of burning life, 10% of the blubs will be still burning.

Example 8·24. *Let* $X \sim N(\mu, \sigma^2)$. *If* $\sigma^2 = \mu^2, (\mu > 0)$, *express* $P(X < -\mu \mid X < \mu)$ *in terms of cumulative distribution function of* $N(0, 1)$.

[Delhi Univ. B.Sc. (Maths. Hons.) 1988; (Stat. Hons.). 1993]

Solution.

$$P(X < -\mu \mid X < \mu) = \frac{P(X < -\mu \cap X < \mu)}{P(X < \mu)} = \frac{P(X < -\mu)}{P(X < \mu)}; \qquad (\because \mu > 0)$$

$$= \frac{P(Z<-2)}{P(Z<0)} \qquad \left(Z=\frac{X-\mu}{\sigma}=\frac{X-\mu}{\mu}\right)$$

$$= \frac{P(Z>2)}{(1/2)}; \qquad \text{(By symmetry)}$$

$$= 2[1-P(Z\le 2)] = 2[1-\Phi(2)]$$

where $\Phi(.)$ is the distribution function of standard normal variate.

Example 8.25 *Can X and $-X$ have the same distribution ? If so, when ?* **[Delhi Univ. B.A., (Spl. Course Statistics), 1989]**

Solution. Yes; X and $-X$ can have the same distribution provided the p.d.f. $f(x)$ of X is symmetric about origin *i.e.*, if $f(-x)=f(x)$.

For example, X and $-X$ have the same distribution if :

(*i*) $X \sim N(0, 1)$

(*ii*) X has standard cauchy distribution [c.f. § 8.9]

$$f(x)=\frac{1}{\pi}\cdot\frac{1}{(1+x^2)}; \quad -\infty<x<\infty$$

(*iii*) X has standard Laplace distribution [c.f. § 8.7]

$$p(x)=\tfrac{1}{2}e^{-|x|}; \quad -\infty<x<\infty.$$

and so on. Obviously X and $Y=-X$ are not identical.

Remark. This example illustrates that if the r.v.'s. X and Y are identical, they have the same distributions. However if X and Y have the same distribution, it does not imply that they are identical.

Example 8.26. *If X, Y are independent normal variates with means 6, 7 and variances 9, 16 respectively, determine λ such that*

$$P(2X+Y\le\lambda)=P(4X-3Y\ge 4\lambda)$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1988; B.Sc., 1987]

Solution. Since X and Y are independent, by § 8.2.8 [c.f. equation (8.15*a*)], we have

$$U=2X+Y \sim N(2\times 6+7,\ 4\times 9+16),\ \textit{i.e.},\ U\sim N(19, 52)$$

$$V=4X-3Y \sim N(4\times 6-3\times 7,\ 16\times 9+9\times 16),\ \textit{i.e.},\ V\sim N(3, 288)$$

and $$P(2X+Y\le\lambda)=P(U\le\lambda)=P\left(Z\le\frac{\lambda-19}{\sqrt{52}}\right), \text{ where } Z\sim N(0, 1)$$

and $$P(4X-3Y\ge 4\lambda)=P(V\ge 4\lambda)=P\left(Z\ge\frac{4\lambda-3}{12\sqrt{2}}\right), \text{ where } Z\sim N(0, 1)$$

Now $$P(2X+Y\le\lambda)=P\{(4X-3Y)\ge 4\lambda\}$$

$$\Rightarrow \quad P\left(Z\le\frac{\lambda-19}{\sqrt{52}}\right)=P\left(Z\ge\frac{4\lambda-3}{12\sqrt{2}}\right)$$

$$\Rightarrow \quad \frac{\lambda-19}{\sqrt{52}}=-\left(\frac{4\lambda-3}{12\sqrt{2}}\right)$$

[Since $P(Z\le a)=P(Z\ge b) \Rightarrow a=-b$, because normal probability curve is symmetric about $Z=0$].

$$\Rightarrow \quad \frac{\lambda - 19}{\sqrt{13}} = \frac{3 - 4\lambda}{6\sqrt{2}}$$

$$\Rightarrow \quad (6\sqrt{2} + 4\sqrt{13})\,\lambda = 114\sqrt{2} + 3\sqrt{13}$$

$$\Rightarrow \quad \lambda = \frac{114\sqrt{2} + 3\sqrt{13}}{6\sqrt{2} + 4\sqrt{13}}$$

Example 8·27. *If X and Y are independent normal variates possessing a common mean μ such that*

$$P(2X + 4Y \le 10) + P(3X + Y \le 9) = 1$$

$$P(2X - 4Y \le 6) + P(Y - 3X \ge 1) = 1,$$

determine the values of μ and the ratio of the variances of X and Y.

Solution. Let $\text{Var}(X_1) = \sigma_1^2$ and $\text{Var}(Y) = \sigma_2^2$

Since $E(X) = E(Y) = \mu$, (Given) and X and Y are independent by § 8·2·8 [c.f. equation (8·15*a*)], we have

$2X + 4Y \sim N(2\mu + 4\mu, 4\sigma_1^2 + 16\sigma_2^2)$, *i.e.*, $N(6\mu, 4\sigma_1^2 + 16\sigma_2^2)$

$3X + Y \sim N(3\mu + \mu, 9\sigma_1^2 + \sigma_2^2)$, *i.e.*, $N(4\mu, 9\sigma_1^2 + \sigma_2^2)$

$2X - 4Y \sim N(2\mu - 4\mu, 4\sigma_1^2 + 16\sigma_2^2)$, *i.e.*, $N(-2\mu, 4\sigma_1^2 + 16\sigma_2^2)$

$Y - 3X \sim N(\mu - 3\mu, \sigma_2^2 + 9\sigma_1^2)$, *i.e.*, $N(-2\mu, 9\sigma_1^2 + \sigma_2^2)$

Let us further write :

$$4\sigma_1^2 + 16\sigma_2^2 = \alpha^2 \quad \text{and} \quad 9\sigma_1^2 + \sigma_2^2 = \beta^2 \qquad \ldots(1)$$

If Z denotes the Standard Normal Variate, *i.e*, if $Z \sim N(0, 1)$, we get

$$P(2X + 4Y \le 10) + P(3X + Y \le 9) = 1$$

$$\Rightarrow \quad P\left(Z \le \frac{10 - 6\mu}{\alpha}\right) + P\left(Z \le \frac{9 - 4\mu}{\beta}\right) = 1$$

$$\Rightarrow \quad P\left(Z \le \frac{10 - 6\mu}{\alpha}\right) = 1 - P\left(Z \le \frac{9 - 4\mu}{\beta}\right) = P\left(Z \ge \frac{9 - 4\mu}{\beta}\right)$$

$$\Rightarrow \quad \frac{10 - 6\mu}{\alpha} = -\left(\frac{9 - 4\mu}{\beta}\right), \qquad \ldots(2)$$

(Since normal distribution is symmetric about $Z = 0$).

Similarly

$$P(2X - 4Y \le 6) + P(Y - 3X \ge 1) = 1$$

$$\Rightarrow \quad P\left(Z \le \frac{6 + 2\mu}{\alpha}\right) + P\left(Z \ge \frac{1 + 2\mu}{\beta}\right) = 1$$

$$\Rightarrow \quad P\left(Z \le \frac{6 + 2\mu}{\alpha}\right) = 1 - P\left(Z \ge \frac{1 + 2\mu}{\beta}\right) = P\left(Z \le \frac{1 + 2\mu}{\beta}\right)$$

$$\Rightarrow \quad \frac{6 + 2\mu}{\alpha} = \frac{1 + 2\mu}{\beta} \qquad \ldots(3)$$

Solving (2) and (3), we get

$$\frac{\alpha}{\beta} = \frac{6 + 2\mu}{1 + 2\mu} = \frac{10 - 6\mu}{4\mu - 9} \qquad \ldots(4)$$

$$\Rightarrow \quad (6+2\mu)(4\mu-9) = (10-6\mu)(1+2\mu)$$

$$\Rightarrow \quad 5\mu^2 - 2\mu - 16 = 0 \qquad \text{(On simplification)}$$

$$\Rightarrow \quad \mu = \frac{2 \pm \sqrt{4+320}}{10} = \frac{2 \pm 18}{10}$$

$$\Rightarrow \quad \mu = 2 \text{ or } -1{\cdot}6$$

Substituting $\mu = 2$ in (4), we get.

$$\frac{\alpha}{\beta} = \frac{10}{5} = 2, \quad i.e., \quad 4 = \frac{\alpha^2}{\beta^2}$$

From (1), we get

$$4 = \frac{4\sigma_1^2 + 16\sigma_2^2}{9\sigma_1^2 + \sigma_2^2} = \frac{4+16\lambda}{9+\lambda} \qquad \left[\text{Taking } \lambda = \frac{\sigma_2^2}{\sigma_1^2}\right]$$

$$\Rightarrow \quad 4(9+\lambda) = 4 + 16\lambda \;\Rightarrow\; \lambda = \frac{32}{12} = \frac{8}{3}$$

Again putting $\mu = -1{\cdot}6$ in (4), we get

$$\left(\frac{14}{11}\right)^2 = \frac{\alpha^2}{\beta^2} = \frac{4+16\lambda}{9+\lambda} \;\Rightarrow\; \lambda = \frac{1280}{1740} = \frac{64}{87}$$

Example 8·28. *If two normal universes A and B have the same total frequency but the standard deviation of universe A is k times that of the universe B, show that maximum frequency of universe A is 1/k times that of universe B.*

Solution. Let N be the same total frequency for each of the two universes A and B. If σ is the standard deviation of universe B, then the standard deviation of universe A is $k\sigma$. Let μ_1 and μ_2 be the means of the universes A and B respectively.

The frequency function of universe A is given by

$$f_A(x) = \frac{N}{k\sigma\sqrt{2\pi}} \exp\left\{-(x-\mu_1)^2/2k^2\sigma^2\right\}$$

and the frequency function of universe B is given by

$$f_B(x) = \frac{N}{\sigma\sqrt{2\pi}} \exp\left\{-(x-\mu_2)^2/2\sigma^2\right\}$$

Since, for a normal distribution, the maximum frequency occurs at the point x = mean, we have

$$[f_A(x)]_{\max} = \text{Maximum frequency of universe } A$$

$$= \left[f_A(x)\right]_{x=\mu_1}$$

$$= \left[\frac{N}{k\sigma\sqrt{2\pi}} \exp\left\{-(x-\mu_1)^2/2k^2\sigma^2\right\}\right]_{x=\mu_1} = \frac{N}{k\sigma\sqrt{2\pi}}$$

Similarly

$$[f_B(x)]_{\max} = [f_B(x)]_{x=\mu_2}$$

$$= \left[\frac{N}{\sigma\sqrt{2\pi}} \exp\left\{ -(x-\mu_2)^2/2\sigma^2 \right\} \right]_{x=\mu_1} = \frac{N}{\sigma\sqrt{2\pi}}$$

$$\therefore \quad \frac{|f_A(x)|_{\max}}{|f_B(x)|_{\max}} = \frac{1}{k}$$

EXERCISE 8 (b)

1. "If the Poisson and the Normal distributions are limiting cases of Binomial distribution, then there must be a limiting relation between the Poisson and the Normal distributions." Investigate the relation.

2. (*a*) Derive the mathematical form and properties of normal distribution Discuss the importance of normal distribution in Statistics.

(*b*) Mention the chief characteristics of Normal distribution and Normal probability curve. **[Delhi Univ. B.Sc. (Stat Hons.), 1989]**

3. (*a*) Explain, under what conditions and how the binomial distribution can be approximated to the normal distribution.

(*b*) For a normal distribution with mean 'μ' and standard deviation σ, show that the mean deviation from the mean 'μ' is equal to $\sigma\sqrt{(2/\pi)}$. What will be the mean deviation from median ?

(*c*) The distribution of a variable X is given by the law:

$$f(x) = \text{Constant} \exp\left[-\frac{1}{2}\left(\frac{x-100}{5} \right)^2 \right], -\infty \leq x \leq \infty$$

Write down the value of :

(*i*) the constant,
(*ii*) the mean,
(*iii*) the median,
(*iv*) the mode,
(*v*) standard deviation,
(*vi*) the mean deviation,
(*vii*) the quartile deviation of the distribution.

(Gujarat Univ. B.Sc. April 1978)

Ans. (*i*) $\frac{1}{5\sqrt{2\pi}}$, (*ii*) 100, (*iii*) 100, (*iv*) 100, (*v*) 5 (*vi*) $\sqrt{(2/\pi)} \times 5 \simeq 4$, (*vii*) $\frac{1}{2} \times 5 = 3{\cdot}33$ **(approx .)** .

(*d*) Define Normal probability distribution. If the mean of a Normal population is μ and its variance σ^2, what are its (*i*) mode, (*ii*) Median, (*iii*) β_1 and β_2 ?

(*e*) For a normal distribution $N(\mu, \sigma^2)$:

(*i*) Show that the mean, the median and the mode coincide.
(*ii*) Find the recurrence relation between μ_{2n} and μ_{2n-2}.
(*iii*) State and prove additive property of normal variates.
(*iv*) Obtain the points of inflexion for the normal distribution $N(\mu, \sigma^2)$.
(*v*) Obtain mean deviation about mean.

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

(*f*) Show that any linear combination of n independent normal variates is also a normal variate. **[Delhi Univ. B.Sc. (Stat. Hons.), 1989]**

(*g*) Show that for the normal curve :

(*i*) The maximum occurs at the mean of the distribution, and

(*ii*) the points of inflexion lie at a distance of $\pm \sigma$ from the mean, where σ is the standard deviation. **[Delhi Univ. M.A. (Eco.), 1987]**

(*h*) Describe the steps involved in fitting a normal distribution to the given data and computing the expected frequencies.

(*i*) Explain how the normal probability integral

$$\int_0^{z_1} \varphi(z)\, dz ,$$

is used in computing normal probabilities.

4. Write a note on the salient features of a normal distribution. $N(\mu, \sigma^2)$ denotes the normal distribution of each of the random variables $X_1, X_2, X_3, ..., X_n$, where μ is the mean and σ^2 the variance. Prove the following :

(*i*) If $X_1, X_2, ..., X_n$ are independent, then $X_1 + X_2 + ... + X_n$ has the distribution $N(n\mu, n\sigma^2)$.

(*ii*) kX, where k is a constant has the distribution $N(k\mu, k^2\sigma^2)$.

(*iii*) $X + a$, where a is a constant has the distribution $N(\mu + a, \sigma^2)$

(*iv*) In (*i*) if $\bar{X} = \dfrac{X_1 + X_2 + ... + X_n}{n}$ then

$$\sqrt{n}\,\frac{(\bar{X} - \mu)}{\sigma} \text{ has the distribution } N(0, 1) .$$

5. (*a*) Show that for a normal distribution with mean μ and variance σ^2, the central moments satisfy the relation

$$\mu_{2n} = (2n - 1)\,\mu_{2n-2}\,\sigma^2 ; \mu_{2n+1} = 0$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1987]

Hence show that $\mu_{2n} = \dfrac{(2n)!}{n!}\left(\frac{1}{2}\sigma^2\right)^n$ and $\mu_{2n+1} = 0 ; n = 1, 2, ...$

[Delhi Univ. B.Sc. (Stat Hons.) 1985]

(*b*) State the mathematical equation of a normal curve. Discuss its chief features.

(*c*) Find the moment generating function of the normal distribution (m, σ^2), and deduce that

$$\mu_{2n+1} = 0,$$
$$\mu_{2n} = 1 \cdot 3 \cdot 5 ... (2n - 1)\,\sigma^{2n},$$

where μ_n denotes the nth central moment.

[Delhi Univ. B.Sc. (Stat. Hons.) 1990,' 82]

(*d*) Show that all central moments of a normal distribution can be expressed in terms of the standard deviation and obtain the expression in the general case.

[Aligarh Univ. B.Sc. 1992]

(*e*) The normal table gives the values of the integral:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}t^2\right) dt$$

for different values of x.

Explain how to use this table to obtain the proportion of observations of a normal variate with mean μ and S.D. σ, which lie above a given value 'a',

(*i*) where $a > \mu$, (*ii*) where $a < \mu$.

6. (*a*) If X_1 and X_2 and two independent normal variates with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively, show that the variables U and V where $U = X_1 + X_2$ and $V = X_1 - X_2$, are independent normal variates. Find the means and variances of U and V.

(*b*) If X_1 and X_2 are independent standard normal variates obtain the p.d.f. of $(X_1 - X_2)/\sqrt{2}$.

Ans. $U = (X_1 - X_2)/\sqrt{2} \sim N(0, 1)$

(*c*) Suppose $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 1)$ are independent r.v.'s.

(*i*) Find the joint distribution of $(X_1 + X_2)/\sqrt{2}$ and $(X_1 - X_2)/\sqrt{2}$.

(*ii*) Argue that $2X_1X_2$ and $X_2^2 - X_1^2$ have the same distribution.

Ans. (*i*) $U = (X_1 + X_2)/\sqrt{2}$ and $V = (X_1 - X_2)/\sqrt{2}$ are independent $N(0, 1)$ variates

(*ii*) **Hint.** $X_2^2 - X_1^2 = 2\left[\frac{X_2 + X_1}{\sqrt{2}}\right]\left[\frac{X_2 - X_1}{\sqrt{2}}\right] = 2(UV)$

$= 2 \times$ **[Product of two independent SNV 's]**

$2X_1X_2 = 2 \times$ **[Product of two independent SNV 's]**

Hence the result.

7. (*a*) Let X be normally distributed with mean 8 and s.d. 4. Find

(*i*) $P(5 \le X \le 10)$, (*ii*) $P(10 \le X \le 15)$, (*ii*) $P(X \ge 15)$, (*iv*) $P(X \le 5)$.

Ans. (*i*) 0·4649 (*ii*) 0·2684 (*iii*) 0·0401 (*vi*) 0·2266.

(*b*) The standard deviation of a certain group of 1,000 high school grades was 11% and the mean grade 78%. Assuming the distribution to be normal, find

(*i*) How many grades were above 90%?

(*ii*) What was the highest grade of the lowest 10?

(*iii*) What was the interquartile range ?

(*iv*) Within what limits did the middle 90% lie?

Ans. (*i*) 138, (*ii*) 52, (*iii*) $Q_1 = 70{\cdot}575$; $Q_3 = 85{\cdot}425$, and (*iv*) 60 *%to* 96·2%

(*c*) If X is normally distributed with mean 2 and variance 1, find

$P(|X - 2| < 1)$. **Ans.** 0·6826 [or $\Phi(1) - \Phi(-1)$]

(*d*) If $X \sim N(\mu = 2, \sigma^2 = 2)$, find $P(|X - 1| \leq 2)$ in terms of distribution function of standard normal variate.

Ans. Probability $= P(-1 \leq X \leq 3) = \Phi(1/\sqrt{2}) - \Phi(-3/\sqrt{2})$

(*e*) If $X \sim N(30, 5^2)$ and $Y \sim N(15, 10^2)$, show that

$$P(26 \leq X \leq 40) = P(7 \leq Y \leq 35).$$

Hint. Each Probability $= P(-0{\cdot}8 \leq Z \leq 2)$ where $Z \sim N(0, 1)$

(*f*) If $X \sim N(30, 5^2)$, find the probabilities of

(*i*) $26 \leq X \leq 40$, (*ii*) $|X - 30| > 5$, (*iii*) $X \geq 42$, (*iv*) $X \leq 28$

[Bihar P.C.S., 1988]

Ans. (*i*) 0·7653, (*ii*) 0·3174, (*iii*) 0·0082, (*iv*) 0·3446

8. (*a*) In a normal population with mean 15·00 and standard deviation 3·5, it is known that 647 observations exceed 16·25. What is the total number of observations in the population? **(Sri Venkateswara Univ. B.Sc. April 1990)**

Hint. Let $X \sim N(\mu, \sigma^2)$ where $\mu = 15$ and $\sigma = 3{\cdot}5$.

If N is the total number of observations in the population, then we have to find N such that

$$N \times P(X > 16{\cdot}25) = 647$$

(*b*) Assume the mean heights of soldiers to be 68·22 inches with a variance of 10·8 (in.)^2. How many soldiers in a regiment of 1,000 would you expect to be over 6 feet tall? (Given that the area under the standard normal curve between $X = 0$ and $X = 0{\cdot}35$ is 0·1368 and between $X = 0$ and $X = 1{\cdot}15$ is 0·3746).

Ans. 125 **[Osmania Univ. M.A., 1992]**

9. (*a*) If 100 true coins are thrown, how would you obtain an approximation for the probability of getting (*i*) 55 heads, (*ii*) 55 or more heads, using Tables of Area of normal probability function.

(*b*) Prove that Binomial distribution in certain cases becomes normal.

A six faced dice is thrown 720 times. Explain how an approximate value of the probability of the following events can be found out easily. (Finding out the numerical values of these probabilities is not necessary):

(*i*) 'six' comes for more than 130 times

(*ii*) chance of 'six' lies between 100 and 140.

10. (*a*) The number (X) of items of a certain kind demanded by customers follows the Poisson law with parameter 9. What stock of this item should a retailer keep in order to have a probability of 0·99 of meeting all demands made on him? Use normal approximation to the Poisson law.

(*b*) Show that the probability that the number of heads in 400 throws of a fair coin lies between 180 and 220 is $\simeq 2F(2) - 1$, where $F(x)$ denotes the standard normal distribution function.

11. In an intelligence test administered to 1,000 children, the average score is 42 and standard deviation 24.

(*i*) Find the number of children exceeding the score 60, and

(*ii*) Find the number of children with score lying between 20 and 40. (Assume the normal distribution.) **Ans.** (*i*) 227 (*iii*) 289

12. The mean I.Q. (intelligence quotient) of a large number of children of age 14 was 100 and the standard deviation 16. Assuming that the distribution was normal, find

(*i*) What % of the children had I.Q. under 80?

(*ii*) Between what limits the I.Q.'s of the middle 40% of the children lay?

(*iii*) What % of the children had I.Q.'s within the range $\mu \pm 1.96\,\sigma$?

Ans. (*i*) 10·56%, (*ii*) 91·6, 108·4, (*iii*) 0·95

13. (*a*) In a university examination of a particular year, 60% of the students failed when mean of the marks was 50% and s.d. 5%. University decided to relax the conditions of passing by lowering the pass marks, to show its result 70%. Find the minimum marks for a student to pass, supposing the marks to be normally distributed and no change in the performance of students takes place.

Ans. 47·375.

(*b*) The width of a slot on a forging is normally distributed with mean 0·900 inch and standard deviation 0·004 inch. The specifications are $0{\cdot}900 \pm 0{\cdot}005$ inch. What percentage of forgings will be defective ?

Hint. Let X denote the width (in inches) of the slot. We want

$$100 \times P(X \text{ lies outside specification limits})$$
$$= 100\,[\,1 - P(X \text{ lies within specification limits})\,]$$
$$= 100\,[\,1 - P(0{\cdot}895 < X < 0{\cdot}905)\,]\,.$$

14. (*a*) The monthly incomes of a group of 10,000 persons were found to be normally distributed with mean Rs. 750 and s.d. Rs. 50. Show that of this group, about 95% had income exceeding Rs. 668 and only 5% had income exceeding Rs 832. What was the lowest income among the richest 100?

Ans. Rs. 866·3 .

(*b*) Given that X is normally distributed with mean 10 and

$$P(X > 12) = 0{\cdot}1587,$$

what is the probability that X will fall in the interval (9, 11)?

Take $\Phi(1) = 0{\cdot}8413$ and $\Phi\left(-\frac{1}{2}\right) = 0{\cdot}3085$

where $$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-u^2/2)\, du$$

Ans. 0·3830

(*c*) A normal distribution has mean 25 and variance 25. Find

(*i*) the limits which include the middle 50% of the area under the curve, and

(*ii*) the values of x corresponding to the points of inflexion of the curve.

Ans. (*i*) Limits which include the middle 50% of the area under the curve are:

$Q_1 = \mu - 0{\cdot}6745\,\sigma = 21{\cdot}7275$; $Q_3 = \mu + 0{\cdot}6745\,\sigma = 38{\cdot}2725$

(*ii*) (30, 20)

15. (*a*) In a distribution exactly normal 7% of the items are under 35 and 89% are under 63. What are the mean and standard deviation of the distribution?

Ans. $\mu = 50{\cdot}3$, $\sigma = 10{\cdot}33$.

(*b*) In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and variance of the distribution.

Given that area between mean-ordinates and ordinate at any σ distance from mean,

$$Z = \frac{X-\mu}{\sigma} \quad : \quad 0{\cdot}496 \quad 1{\cdot}405$$

$$\text{Area} \quad : \quad 0{\cdot}19 \quad 0{\cdot}42$$

[Delhi Univ. B.Sc., 1987; Madras Univ. B.Sc., 1990]

Ans. $\mu = 50, \sigma = 10$

16. (*a*) A minimum height is to be prescribed for eligibility to government services such that 60% of the young men will have a fair chance of coming up to that standard. The heights of youngmen are normally distributed with mean 60·6" and s.d. 2·55". Determine the minimum specification.

Ans. 59·9".

Hint. We want x_1 s.t. $P(X > x_1) = 0{\cdot}6$

When $X = x_1$, $Z = \dfrac{x_1 - 60{\cdot}6}{2{\cdot}55} = -z_1$, (say) . ...(*)

[Note the negative sign, which is obvious from the diagram]

Obviously $P(0 < Z < z_1) = 0{\cdot}10 \Rightarrow z_1 = 0{\cdot}254$

Substituting in (*), we get

$$x_1 = 60{\cdot}6 - 2{\cdot}55 \times 0{\cdot}254 = 60{\cdot}6 - 0{\cdot}65 = 59{\cdot}95"$$

(*b*) The height measurements of 600 adult males are arranged in ascending order and it is observed that 180th and 450th entries are 64·2" and 67·8" respectively. Assuming that the sample of heights is drawn from a normal population, estimate the mean and s.d. of the distribution.

Ans. 67·78", 3"

17. (*a*) Marks secured by students in sections I and II of a class are independently normally distributed with means 50 and 60 respectively and variances 10 and 6 respectively. What is the probability that a randomly chosen student from section II scores more marks than a randomly chosen student from section I? What percentage of students are expected to secure first division (*i.e.*, 60 marks or more) in section I? Write down your results in terms of the standard normal distribution fucntion.

Hint. $X \sim N(50, 10)$, $Y \sim N(60, 6)$ are independent r.v.'s. $U = Y - X \sim N(10, 16)$. We want $P(Y > X) = P(U > 0)$.

(*b*) In an examination, the mean and standard deviation (s.d.) of marks in Mathematics and Chemistry are given below

	Mean	s.d.
Maths.	45	10
Chem.	50	15

Assuming the marks in the two subjects to be independent normal variates, obtain the probability that a student scores total marks lying between 100 and 130. [Full marks in each subject are 100]. Given that

$$F(0{\cdot}28) = 0{\cdot}1103,\ F(1{\cdot}94) = 0{\cdot}4738,$$

where $$F(z) = \frac{1}{\sqrt{2\pi}} \int_0^z \exp\left(-\tfrac{1}{2}x^2\right) dx .$$

[Bhagalpur Univ. B.Sc., 1990]

18. (*a*) One thousand candidates in an examination were grouped into three classes I, II, III in descending order of merits. The numbers in the first two classes were 50 and 350 respectively. The highest and the lowest marks in class II were 60 and 50 respectively. Assuming the distribution to be normal, prove that the average mark is approximately 48·2 and standard deviation, approximately 7·1. The following data may be used:

The area A is measured from the mean zero to any ordinate X.

$\frac{X}{\sigma}$	A	$\frac{X}{\sigma}$	A
0·2	0·079	1·5	0·433
0·3	0·118	1·6	0·445
0·4	0·155	1·7	0·455

(*b*) In an examination marks obtained by the students in Mathematics, Physics and Chemistry are distributed normally about the means 50, 52 and 48 with S.D. 15, 12, 16 respectively. Find the probability of securing total marks of

(*i*) 180 or above, (*ii*) 90 or below.

$$\left[\frac{1}{\sqrt{2\pi}} \int_{1{\cdot}2}^{\infty} \exp\left(-z^2/2\right) dz = 0{\cdot}1942,\ \frac{1}{\sqrt{2\pi}} \int_{2{\cdot}4}^{\infty} \exp\left(-z^2/2\right) dz = 0{\cdot}0224\right]$$

Ans. 0·1942, 0·0224 **[Agra Univ. B.Sc., 1988]**

19. In a certain examination the percentage of passes and distinctions were 46 and 9 respectively. Estimate the average marks obtained by the candidates, the minimum pass and distinction marks being 40 and 75 respectively. (Assume the distribution of marks to be normal.) (**Ans.** $\mu = 36{\cdot}4,\ \sigma = 28{\cdot}2$)

Also determine what would have been the minimum qualifying marks for admission to a re-examination of the failed candidates, had it been desired that the best 25% of them should be given another opportunity of being examined.

Ans. 29.

20. The local authorities in a certain city installed 2,000 electric lamps in a street of the city. If the lamps have an average life of 1,000 burning hours with a S.D. of 200 hours,

(*i*) What number of the lamps might be expected to fail in the first 700 burning hours,

(*ii*) After what periods of burning hours would we expect that

(*a*) 10% of the lamps would have failed, and

(*b*) 10% of the lamps would be still burning?

Assume that lives of the lamps are normally distributed.

You are given that $F(1{\cdot}50) = 0{\cdot}933$, $F(1{\cdot}28) = {\cdot}900$,

where $$F(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz$$

Ans. (*i*) 134, (*ii*) (*a*) 744, (*b*) 1256. **[Allahabad Univ. B.Sc., 1987]**

21. (*a*) The quartiles of a normal distribution are 8 and 14 respectively. Estimate the mean and standard deviation.

Ans. $\mu = 11$, $\sigma = 4{\cdot}4$.

(*b*) The third decile and the upper quartile of a normal distribution are 56 and 63 respectively. Find the mean and varaince of the distribution.

Ans. $\mu = 59{\cdot}1$, $\sigma = 5{\cdot}8$.

22. (*a*) 5,000 variates are normally distributed with mean 50 and probable error (semi-interquartile range) 13·49. Without using tables, find the values of the quartiles, median, mode standard deviation and mean deviation. Find also the value of the variate for which cumulative frequency is 1250.

[Meerut Univ. B.Sc., 1989]

Ans. $Q_1 = 36{\cdot}51$ $Q_3 = 63{\cdot}49$, $\sigma = 20$, M.D. $= 16$, $x_1 = 36{\cdot}51$.

(*b*) The following table gives frequencies of occurrence of a varaible X between certain limits :

Variable X	Frequency
Less than 40	30
40 or more but less than 50	33
50 and more	37

The distribution is exactly normal. Find the distribution and also obtain the frequency between $X = 50$ and $X = 60$. **[Kurukshetra Univ. M.A. (Eco.), 1990]**

Ans. Hint. $50 - \mu = 0{\cdot}33\,\sigma$; $40 - \mu = -0{\cdot}52\,\sigma$

$\mu = 46{\cdot}12$, $\sigma = 11{\cdot}76$

$N.P(50 < X < 60) = 100 \times 0{\cdot}2517 \simeq 25$

23. (*a*) Suppose that a doorway being constructed is to be used by a class of people whose heights are normally distributed with mean 70" and standard deviation 3". How long may the doorway be without causing more than 25% of the

people to bump their heads? If the height of the doorway is fixed at 76", how many persons out of 5,000 are expected to bump their heads?

[For a normal distributionthe quartile deviation is 0·6745 times standard deviation. For a standard normal distribution $Z = \frac{X - \bar{X}}{\sigma}$, the area under the curve between $Z = 0$ and $Z = 2$ is 0 4762.]

(*b*) A normal population has a coefficient of variation 2% and 8% of the population lies above 120. Find the mean and S.D.

Ans. $\mu = 122, \sigma = 2{\cdot}44$

24. Steel rods are manufactured to be 3 inches in diameter but they are acceptable if they are inside the limits 2·99 inches and 3·01 inches. It is observed that 5% are rejected as oversize and 5% are rejected as undersize. Assuming that the diameters are normally distributed, find the standard deviation of the distribution. Hence calculate, what would be the proportion of rejects if the permissible limits were widened to 2·985 inches and 3·015 inches.

[**Hint.** Let X denote the diameter of the rods in inches and let $X \sim N(\mu, \sigma^2)$. Then we are given

$$P(X > 3{\cdot}01) = 0{\cdot}05 \text{ and } P(X < 2{\cdot}99) = 0{\cdot}05$$

$$\Rightarrow \quad \frac{3{\cdot}01 - \mu}{\sigma} = 1{\cdot}65 \quad \text{and} \quad \frac{2{\cdot}99 - \mu}{\sigma} = -1{\cdot}65$$

Solving we get $\mu = 3$ and $\sigma = \frac{1}{165}$

The probability that a random value of X lies within the rejection limits is

$$P(2{\cdot}985 < X < 3{\cdot}015) = P(-2{\cdot}475 < Z < 2{\cdot}475) = 2 \times P(0 < Z < 2{\cdot}475)$$
$$= 2 \times 0{\cdot}4933 = 0{\cdot}9866$$

Hence the probability that X lies outside the rejection limits is

$$1 - 0{\cdot}9866 = 0{\cdot}0134$$

Therefore, the proportion of the rejects outside the revised limits is 0·0134, *i.e.*, 1·34%].

25. Derive the moment generating function of a random variable which has a normal distribution with mean μ and variance σ^2. Hence or otherwise prove that a linear combination of independent normal variates is also normally distributed.

An investor has the choice of two of four investments X_1, X_2, X_3, X_4. The profits from these may be assumed to be indepedetnly distributed, and

the profit from X_1 is $N(2, 1)$,
the profit from X_2 is $N(3, 3)$,
the profit from X_3 is $N(1, \frac{1}{4})$.
the profit from X_4 is $N(2\frac{1}{2}, 4)$.

(Profits are given in £ 1000 per annum).

Which pair should he choose to maximise his probability of making a total annual profit of at least £ 2000? **(London Univ. B.Sc. 1977)**

26. (*a*) State the important properties of the normal distribution and obtain from the tables the inter-quartile range in terms of its mean μ and standard deviation σ.

Find the mean and standard deviation as well as the inter-quartile range of the following data. Compare the inter-quartile range with that obtained from mean and standard deviation on the assumption of normality.

X (central values) ...	0	1	2	3	4	5	6
f (frequency) ...	5	9	15	32	21	10	8

(*b*) The following table gives Baseball throws for a distance by 303 first year high school girls:

Distance in feet	Number of girls	Distance in feet	Number of girls
15 and under 25	1	85 and under 95	44
25 and under 35	2	95 and under 105	31
35 and under 45	7	105 and under 115	27
45 and under 55	25	115 and under 125	11
55 and under 65	33	125 and under 135	4
65 and under 75	53	135 and under 145	1
75 and under 85	64		

(*i*) Fit a normal distribution and find the theoretical frequencies for the classes of the above frequency distribution.

(*ii*) Find the expected number of girls throwing baseballs at a distance exceeding 105 feet on the basis that the data fit a normal distribution.

27. (*a*) The table given below shows the distribution of heights among freshmen in a college :

Height in inches	61	62	63	64	65	66	67	68
Frequency	4	20	23	75	114	186	212	252
Height in inches	69	70	71	72	73	74		
Frequency	218	175	149	46	18	8		

By comparing the proportion of cases lying between $\mu \pm (2/3)\,\sigma$, $\mu \pm \sigma$, $\mu \pm 2\,\sigma$ and $\mu \pm 3\,\sigma$, for this distribution and for a normal curve, state whether the distribution may be considered normal.

(*b*) Fit a normal distribution to the following data of heights in cms of 200 Indian adult males :

Height in (cms)	*Frequency*
144 — 150	3
150 — 156	12
156 — 162	23
162 — 168	52
168 — 174	61
174 — 180	39
180 — 186	10

(*c*) Two hundred and fifty-five metal rods were cut roughly six inches over size. Finally the lengths of the oversize amount were measured exactly and grouped with 1-inch intervals, there being in all 12 groups. The frequency distribution for the 255 lengths was

Central value : x	1	2	3	4	5	6
Frequency : f	2	10	19	25	40	44
x	7	8	9	10	11	12
f	41	28	25	15	5	1

Fit a normal distribution to the data by the method of ordinates and calculate the expected frequencies.

28. (*a*) Let $X \sim N(\mu, \sigma^2)$. Let

$$\Phi(x) = P[X \leq x],$$

calculate the probabilities of the following events in terms of Φ :

(*i*) $\alpha X + \beta \leq t$, where α, β are finite constants.

(*ii*) $-X \geq t$

(*iii*) $|X| > t$ **[Poona Univ. B.E., 1991]**

(*b*) Determine C such that the following function becomes a distribution function:

$$F(x) = C \int_{-\infty}^{x} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] dy$$

29. (*a*) Determine the constant C so that $C.e^{-2x^2+x}, -\infty < x < \infty$, is a density function. If the random variable X has the resulting density function, then find (*i*) the mean of X, (*ii*) the variance of X and (*iii*) $P(X \geq 1/4)$.

Ans. (*i*) 0·25 (*ii*) 0·25 (*iii*) 0·5

(*b*) If $f(x) = k \cdot \exp\left\{-(9x^2 - 12x + 13)\right\}$, is the p.d.f. of a normal distribution (k, being a constant) find the mean and s.d. of the distribution.

(*c*) If X is a normal variate with p.d.f. $f(x) = 0{\cdot}03989 \exp(-0{\cdot}005x^2 + 0{\cdot}5x - 12{\cdot}5)$, express $f(x)$ in standard form and hence find the mean and variance of X. **[M.S. Baroda Univ. B.Sc., 1991]**

(*d*) Let the probability function of the normal distribution be

$$P(x) = ke^{-1/8x^2 + 2x}, \quad -\infty < x < \infty$$

Find k, μ and σ^2. **[Delhi Univ. B.Sc. (Stat. Hons.), 1985]**

(*e*) X_1, X_2, X_3, X_4 is a random sample from a normal distribution with mean 100 and variance 25 and $\bar{X} = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$.

State the distribution, expected value and variance of each of the following:

(*i*) $4\bar{X}$, (*ii*) $X_1 - 2X_2 + X_3 - 3X_4$,

(*iii*) $\frac{1}{25}\sum_{i=1}^{} \{X_i - 100\}^2$ **[Bangalore Univ. B.Sc., 1989]**

Ans. (*b*) **Mean** $= 2/3$, $\sigma = \frac{1}{3\sqrt{2}}$

30. If X is a normal variate with mean 50 and s.d. 10, find $P(Y \le 3137)$, where $Y = X^2 + 1$,

$$\left[\frac{1}{\sqrt{2\pi}}\int_0^{0.6} e^{-x^2/2}\,dx = 0{\cdot}2258\right]$$ **[Delhi Univ. B.Sc. (Hons.), 1990]**

Hint. Required Probability $= P(X^2 + 1 \le 3137) = P(-56 \le X \le 56)$.

Ans. 0·7258

31. Let X be normally distributed with mean μ and variance σ^2. Suppose σ^2 is some function of μ, say $\sigma^2 = h(\mu)$. Pick $h(.)$ so that $P(X \le 0)$ does not depend on μ for $\mu > 0$.

Ans. $P(X \le 0) = P(Z \le -\mu/\sqrt{h(\mu)}) = P(Z \le -1)$; independent of μ if we take $h(\mu) = \mu^2$.

32. (*a*) **If X is a standard normal variate, find $E|X|$ [Ans. $\sqrt{2/\pi} = 4/4$]**

(*b*) X is a random variable normally distributed with mean zero and variance σ^2. Find $E|X|$ **[Delhi Univ. B.Sc. (Stat. Hons.) 1990]**

Hint. $E|X|$ = Mean Deviation about origin

= M.D. about mean ($\because$ Mean = 0)

Ans. $\sqrt{(2/\pi)}$. $\sigma \approx \frac{4}{5}\sigma$

32. (*a*) X is a normal variate with mean 1 and variance 4, Y is another normal variate independent of X with mean 2 and variance 3. What is the distribution of $X + 2Y$? **[Punjab Univ. B.Sc. (Hons.) 1993]**

Ans. $X + 2Y \sim N(5, 16)$

(*b*) If X is a normal variate with mean 1 and S.D. 0 6, obtain $P[X>0]$, $P[|X-1|\geq 0{\cdot}6]$ and $P[-1{\cdot}8<X<2{\cdot}0]$. What is the distribution of $4X+5$?

34. (*a*) Let X and Y be two independent random variables each with a distribution which is $N(0, 1)$. Find the probability density function of $U=a_1X+a_2Y$, where a_1 and a_2 are constants.

(*b*) Show that if X_1, X_2 are mutually independent normal variates having means μ_1, μ_2 and standard deviations σ_1, σ_2 respectively, then $U=a_1X_1+a_2X_2$ is also normally distributed.

34. (*c*) If X_i, $(i=1, 2, ..., n)$ are independent $N(\mu_i, \sigma_i^2)$ variates, obtain the distribution of $\sum_{i=1}^{n} a_i X_i$

where a_i, $i=1, 2, ..., n$ are constants. Hence deduce the distributions of :

(*i*) X_1+X_2

(*ii*) X_1-X_2

(*iii*) $\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$; if X_i'*s* are i.i.d. $N(\mu, \sigma^2)$.

How do the results in (*i*) and (*ii*) compare with those in Poisson distribution and result in (*iii*) compare with Cauchy distribution ?

[Delhi Univ. B.Sc. (Stat. Hons.), 1991]

Hint. For Cauchy distribution, see Remark 4, § 8·9·1 .

35. (*a*) If X is normal with mean 2 and standard deviation 3, describe the distribution of $Y=\frac{1}{2}X-1$. Explain how you would find $P(Y\geq\frac{3}{2})$ from the tables.

Hint. (*a*) We are given that $X\sim N(\mu, \sigma^2)$ where $\mu=2, \sigma=3$. The distribution of the new variable $Y=aX+b$ is also normal with

$$\left.\begin{aligned} E(Y)=E(aX+b)=aE(X)+b=a\mu+b \\ \text{and } \operatorname{Var}(Y)=\operatorname{Var}(aX+b)=a^2\operatorname{Var}(X)=a^2\sigma^2 \end{aligned}\right\} \quad ...(*)$$

Hence $Y=\frac{1}{2}X-1\sim N(\mu_1, \sigma_1^2)$, where μ_1 and σ_1^2 are given by (*) with $a=\frac{1}{2}$ and $b=-1$, *i.e.*,

$$\mu_1=\tfrac{1}{2}\cdot 2-1=0\,;\ \sigma_1^2=(\tfrac{1}{2})^2\cdot 9=\tfrac{9}{4}.$$

Thus $Y\sim N(\mu_1, \sigma_1^2)$, where $\mu_1=0$, $\sigma_1=\frac{3}{2}$.

$$P(Y\geq\tfrac{3}{2})=P(Z\geq 1)=0{\cdot}5-P(0<Z<1)=0{\cdot}5-0{\cdot}3413=0{\cdot}1587\ .$$

(*b*) If X and Y are independent standard normal variables and if $Z=aX+bY+c$ where a, b and c are contants, what will be the distribution of Z ? What is the mean, median and standard deviation of the distribution of Z ?

Find $P(Z\leq 0{\cdot}1)$ if $a=1, b=-1$ and $c=0$. **(I.I.T. B. Tech. 1992)**

Hint. $Z\sim N(c, a^2+b^2)$

If $a=1, b=-1, c=0$ then $Z=X-Y\sim N(0, 2)$

$$\therefore\ P(Z \le 0{\cdot}1) = P\left(U \le \frac{1}{14{\cdot}142} \right); \qquad U = \frac{Z-0}{\sqrt{2}} \sim N(0, 1)$$

36. Let X be a random variable following normal distribution with mean μ and variance σ^2 and let r be a non-negative integer.

If $\mu_r' = E(X^r)$ and if $\mu_{2r} = [\, E(X-\mu)^{2r}\,]$, prove that

(*i*) $\mu'_{r+2} = 2\mu\mu'_{r+1} + (\sigma^2 - \mu^2)\,\mu_r' + \sigma^3 \dfrac{d\,\mu_r'}{d\,\sigma}$

(*ii*) $\mu_{2r+2} = \sigma^2\,\mu_{2r} + \sigma^3 \dfrac{d\,\mu_{2r}}{d\,\sigma}$ **[Madras Univ. B.Sc. (Main), Oct. 1989]**

Hint. (*i*)

$$\frac{d\,\mu_r'}{d\,\sigma} = -\int_{-\infty}^{\infty} \frac{x^r}{\sqrt{2\pi}\,\sigma^2} \exp\left\{-(x-\mu)^2/2\,\sigma^2\right\} dx$$

$$+ \int_{-\infty}^{\infty} \frac{x^r (x-\mu)^2}{\sqrt{2\pi}\,\sigma^4} \exp\left\{-(x-\mu)^2/2\,\sigma^2\right\} dx$$

$$= -\frac{\mu_r'}{\sigma} + \frac{\mu'_{r+2}}{\sigma^3} - \frac{2\mu\mu'_{r+1}}{\sigma^3} + \frac{\mu^2\mu_r'}{\sigma^3}$$

(*ii*) $$\frac{d\,\mu_{2r}}{d\,\sigma} = -\int_{-\infty}^{\infty} \frac{(x-\mu)^{2r}}{\sqrt{2\pi}\,\sigma^2} \exp\left\{-(x-\mu)^2/2\,\sigma^2\right\} dx$$

$$+ \int_{-\infty}^{\infty} \frac{(x-\mu)^{2r+2}}{\sqrt{2\pi}\,\sigma^4} \exp\left\{-(x-\mu)^2/2\,\sigma^2\right\} dx = -\frac{\mu_{2r}}{\sigma} + \frac{\mu_{2r+2}}{\sigma^3}$$

37. Prove that if the independent random variables X and Y have the probability densities,

$$\frac{h}{\sqrt{\pi}}\, e^{-h^2x^2} \text{ and } \frac{k}{\sqrt{\pi}}\, e^{-k^2y^2}, \quad -\infty < (x, y) < \infty$$

then the random variable $U = X + Y$ has the probability density,

$$\frac{l}{\sqrt{\pi}} \cdot e^{-l^2u^2}, \quad -\infty < u < \infty$$

where $$\frac{1}{l^2} = \frac{1}{h^2} + \frac{1}{k^2}$$

38. If $\left[\sum_{i=1}^{n} c_i\mu_i\right]^2 = 9\sum_{i=1}^{n} c_i^2\sigma_i^2$, find $P\left(0 \le Y \le 2\sum_{i=1}^{n} c_i\mu_i\right)$,

where $Y = \sum_{i=1}^{n} c_i X_i$, X_i being a normal variate with mean μ_i and variance σ_i^2.

(Allahabad Univ. B.Sc., 1988)

Hint.

We know $Y = \sum_{i=1}^{n} c_i X_i \sim N(\mu, \sigma^2)$, where $\mu = \sum_{i=1}^{n} c_i \mu_i$ and $\sigma^2 = \sum_{i=1}^{n} c_i^2 \sigma_i^2$

Since $\left(\sum_i c_i \mu_i \right)^2 = 9 \left(\sum_i c_i^2 \sigma_i^2 \right)$, we have $\mu^2 = 9\sigma^2$ or $\frac{\mu}{\sigma} = 3$

If we write $Z = \frac{Y - \mu}{\sigma}$, then $Z \sim N(0, 1)$.

$$P\left(0 \le Y \le 2 \sum_{i=1}^{n} c_i \mu_i\right) = P(0 \le Y \le 2\mu) = P(-3 \le Z \le 3) = 0{\cdot}9973$$

39. (*a*) Find the mean deviation about mean for the normal distribution $N(\mu, \sigma^2)$.

(*b*) If $X \sim N(\mu, \sigma^2)$, find the mean and variance of

$$Y = \frac{1}{2}\left[(x - \mu)/\sigma\right]^2$$ **[Punjabi Univ. M.A. (Eco.), 1991]**

Ans. $E(Y) = 1/2$, $\text{Var}(Y) = 1/2$

Remark. Also see Example 8·30, on Gamma distribution.

(*c*) Derive normal distribution as a limiting case of binomial distribution, clearly stating the conditions involved. **[Delhi Univ. B.A. (Stat. Hons.), 1981]**

40. If $f(x)$ is the density function for the normal distribution with mean zero and standard deviation σ, then show that

$$\int_{-\infty}^{+\infty} [f(x)]^2 \, dx = \frac{1}{2\sigma\sqrt{\pi}}$$

Hence show that if the normal distribution is grouped in intervals with total frequency N_1, and N_2 is the sum of the squares of the frequencies, an estimate of σ is $\frac{N_1^2}{2N_2\sqrt{\pi}}$ **(Gujarat Univ. B.Sc., 1992)**

Hint. $\int_{-\infty}^{\infty} [f(x)]^2 \, dx = \int_{-\infty}^{\infty} \left\{ \frac{1}{\sqrt{2\pi}\,\sigma} \exp(-x^2/2\sigma^2) \right\}^2 dx$

$$= \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} e^{-x^2/\sigma^2} \, dx = \frac{1}{2\pi\sigma^2} \cdot \frac{\sqrt{\pi}}{(1/\sigma)}$$

$$= \frac{1}{2\sigma\sqrt{\pi}} \qquad \left(\because \int_{-\infty}^{\infty} e^{-a^2x^2} \, dx = \sqrt{\pi}/a \right)$$

$$N_2 = \int_{-\infty}^{\infty} \{N_1 f(x)\}^2 \, dx = \frac{N_1^2}{2\sqrt{\pi}\,\sigma}$$

41. Obtain the normal distribution as a limiting case of Poisson distribution when the parameter $\lambda \to \infty$.

42. (*a*) If X is $N(0, 1)$, prove that the p.d.f. of $|X|$ is

$$h(x) = \begin{cases} \sqrt{2/\pi}\exp(-x^2/2), & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

(*b*) Let $X \sim N(0, 1)$ and $Y \sim N(0, 1)$ be independent random variables. Show that $X + Y$ is independent of $X - Y$.

43. If $X \sim N(\mu, 9^2)$ and $Y \sim N(\mu, 12^2)$ are independent, and if $P(X + 2Y \leq 3) = P(2X - Y \geq 4)$, determine μ.

[Calcutta Univ. B.Sc. (Maths Hons.), 1989]

44. If $X \sim N(0, 1)$ and $Y \sim N(0, 1)$, prove that

(*i*) $\text{Var}(\sin X) > \text{Var}(\cos X)$

(*ii*) $E|X - Y| \leq \sqrt{8/\pi}$

[Delhi Univ. B.A. Hons. (Spl. Course-Statistics), 1988]

Hint. (*i*) $X \sim N(0, 1) \Rightarrow \varphi_X(t) = E[\cos tX + i \sin tX) = e^{-t^2/2}$.

$\Rightarrow E(\cos tX) = e^{-t^2/2}$ and $E(\sin tX) = 0$.

Taking $t = 1$ and 2, we get:

$$E(\cos X) = e^{-1/2};\ E(\cos 2X) = e^{-2};\ E(\sin X) = E(\sin 2X) = 0.$$

$$\text{Var}(\cos X) = E(\cos^2 X) - (E\cos X)^2 = E\left[\frac{1 + \cos 2X}{2}\right] - [E\cos X]^2$$

$$= \frac{1}{2}(1 - e^{-1})^2 \simeq 0{\cdot}99$$

$$\text{Similarly Var}(\sin X) = E\left[\frac{1 - \cos 2X}{2}\right] - [E\sin X]^2 = \frac{1}{2}(1 - e^{-2}) \simeq 0{\cdot}43$$

(*ii*) Use $|X - Y| \leq |X| + |Y|$ and $E|X| = E|Y| = \sqrt{2/\pi}$

or $X - Y \sim N(0, \sigma^2 = 2)$; $E|X - Y| = \sqrt{2/\pi}\,\sigma = \sqrt{4/\pi} < \sqrt{(8/\pi)}$

45. Let X and Y be independent $N(0, 1)$ variates. Let $X = R\cos\theta$, $Y = R\sin\theta$. Show that R and θ are independent variates.

[Delhi Univ. B.A. Hons. (Spl. Course Statistics), 1985]

46. If $X \sim N(0, 1)$, find p.d.f. of $|X|$. Hence or otherwise evaluate $E|X|$. **[Delhi Univ. B.Sc. (Maths. Hons.), 1980]**

Hint. Distribution function $G_Y(y)$ of $Y = |X|$ is given by:

$$G_Y(y) = P(Y \leq y) = P(|X| \leq y) = P(-y \leq X \leq y)$$

$$= P(X \leq y) - P(X \leq -y)$$

$$G_Y(y) = F_X(y) - F_X(-y).$$

where $F(.)$ is the distribution function of X. Differentiating, the p.d.f. of $Y = |X|$ is given by

$$g_Y(y) = f_X(y) + f_X(-y) = 2f_X(y)$$

$$\Rightarrow \quad g_Y(y) = \sqrt{2/\pi}\,.\,e^{-y^2/2};\ y \geq 0 \qquad \text{[By symmetry, since } X \sim N(0, 1)\text{]}$$

8·2·15. The log-normal Distribution. The positive r.v. X is said to have a log-normal distribution if $\log_e X$ is normally distributed.

Let $Y = \log_e X \sim N(\mu, \sigma^2)$.

For $x > 0$,

$$F_X(x) = P(X \le x) = P(\log X \le \log x) = P(Y \le \log x)$$

(since $\log X$ is monotonic increasing function)

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\log x} \exp\left\{-(y-\mu)^2/2\sigma^2\right\} dy \qquad [\text{since } Y \sim N(\mu, \sigma^2)]$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{0}^{x} \exp\left\{-(\log u - \mu)^2/2\sigma^2\right\} \frac{du}{u} \qquad (y = \log u)$$

For $x \le 0$,

$$F_X(x) = P(X \le x) = 0$$

Let us define

$$f_X(u) = \begin{cases} \dfrac{1}{u\sigma\sqrt{2\pi}} \cdot \exp\left\{-(\log u - \mu)^2/2\sigma^2\right\}, u > 0 \\ 0, u \le 0 \end{cases} \qquad \ldots(8\cdot18)$$

Then $F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$ for every x and hence $f_X(x)$ defined in (8·18) is a p.d.f. of X.

Remark. If $X \sim N(\mu, \sigma^2)$, then $Y = e^X$ is called a log-normal random variable, since its logarithm $\log Y = X$, **is a normal r.v.**

Moments. The rth moment about origin is given by

$$\mu_r' = E(X^r) = E(e^{rY}) \qquad [\because Y = \log X \Rightarrow X = e^Y]$$

$$= M_Y(r) \qquad (m.g.f. \text{ of } Y, r \text{ being the parameter})$$

$$= \exp\left\{\mu r + \tfrac{1}{2} r^2 \sigma^2\right\} \qquad [\because Y \sim N(\mu, \sigma^2)]$$

$\ldots(8\cdot19)$

Remarks. 1. In particular if we take $\mu = \log \alpha$, $\alpha > 0$ *i.e.*, $\log X \sim N(\log \alpha, \sigma^2)$, then

$$\mu_r' = E(X^r) = \exp\left\{r \cdot \log \alpha + \tfrac{1}{2} r^2 \sigma^2\right\} = \alpha^r \cdot \exp\left\{r^2 \sigma^2/2\right\} \qquad \ldots(8\cdot19\,a)$$

$\therefore$ mean $= \mu_1' = \alpha\, e^{\sigma^2/2}$ and $\mu_2' = \alpha^2 e^{2\sigma^2}$

$$\mu_2 = \mu_2' - \mu_1'^2 = \alpha^2 e^{\sigma^2}(e^{\sigma^2} - 1)$$

2. It arises in problems of economics, biology, geology, and relibility theory. In particular it arises in the study of dimensions of particles under pulverisation.

3. If $X_1, X_2, \ldots, X_n$ is a set of independently identically distributed random variables such that mean of each $\log X_i$ is μ and its variance is σ^2, then the product $X_1 X_2 \ldots X_n$ is asymptotically distributed according to logarithmic normal distribution and with mean μ and variance $n\sigma^2$

EXERCISE 8(c)

1. (*a*) Let X be a non-negative random variable such that $\log X = Y$, (say), is normally distributed with mean μ and variance σ^2.

(*i*) Write down the probability density function of X. Find $E(X)$ and Var (X).

(*iii*) Find the median and the mode of the distribution of X.

(*b*) If X is a normally distributed with zero mean and variance σ^2 find the density function of $U = e^X$ Locate the mode of the distribution.

2. A random variable X has the probability density function:

$$f(x) = \frac{1}{\beta x \sqrt{2\pi}}, \exp\left[-\frac{1}{2\beta^2}(\log x - \alpha)^2\right], \quad x > 0.$$

Find $E(X)$ and Var (X). **[Punjab Univ. M.A. (Eco.), 1991]**

3. A random variate X has the p.d.f.,

$$f(x) = \begin{cases} \frac{1}{x\sqrt{2\pi}} e^{-(\log x)^2/2}, & x > 0 \\ 0, \text{ elsewhere} \end{cases}$$

Calculate the mean, mode, standard deviation and coefficient of skewness.

Ans. $\sqrt{e}$, $1/e$, $\sqrt{e(e-1)}$, and $(1 - e^{-3/2})/\sqrt{e-1}$

4. The random variable X has mean m and standard deviation s. If $Y = \log X$ is normally distributed with mean M and standard deviation S, prove that

$$(i) \quad m = \exp\left[M + \tfrac{1}{2}S^2\right], \qquad (ii) \quad 1 + \frac{s^2}{m^2} = e^{S^2}$$

5. Given that X_i are independent logarithmic normal variates with parameters μ_i and σ_i; $i = 2, ..., n$, find the sth raw moment of the variable

$$Y = \Pi (a_i X_i); \; i = 1, 2, ..., n$$

6. Show that the log-normal distribution is posilively skewed *i.e.*, mean > median > mode.

Ans. Let $Y = \log X \sim N(\mu, \sigma^2)$

$E(X) = e^{\mu + \sigma^2/2}$; Median $= e^{\mu}$; Mode $= e^{\mu - \sigma^2}$

7. If X and Y are two independent log-normal variates, then XY and X/Y are also log-normal variates.

Hint. Let $\log X \sim N(\mu_1, \sigma_1^2)$; $\log Y \sim N(\mu_2, \sigma_2^2)$; $U = XY$ and $V = (X/Y)$.

$$\left.\begin{aligned} \log U &= \log X + \log Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \\ \log V &= \log X - \log Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2) \end{aligned}\right\} \; (\because X \text{ and } Y \text{ are independent})$$

8. If $X \sim N(0, \sigma^2)$, obtain the distribution of e^X Find out the mean of the distribution. **[Delhi Univ. B.Sc. (Stat. Hons.), 1985]**

9. If $X \sim N(\mu, \sigma^2)$, find the p.d.f. of $Y = e^X$, using the result that $E(e^{tX}) = e^{\mu t + t^2\sigma^2/2}$.

Find the coefficient of variation of Y. **[Delhi Univ. M.A. (Eco.), 1991]**

8·3. Gamma Distribution. *The continuous random variable X which is distributed according to the probability law :*

$$f(x) = \begin{cases} \dfrac{e^{-x}\, x^{\lambda-1}}{\Gamma(\lambda)} ; \lambda > 0, 0 < x < \infty \\ 0, \text{ otherwise} \end{cases} \quad \ldots(8\cdot20)$$

is *known as a Gamma variate with parameter* λ and referred to as a $\gamma(\lambda)$ variate and its distribution is called the Gamma-distribution.

Remarks. 1. The function $f(x)$ defined above represents a probability function, since

$$\int_0^\infty f(x)\, dx = \frac{1}{\Gamma(\lambda)} \int_0^\infty e^{-x} x^{\lambda-1}\, dx = \frac{1}{\Gamma(\lambda)} \cdot \Gamma(\lambda) = 1$$

2. A continuous random variable X having the following p.d.f. is said to have a gamma distribution with two parameters a and λ.

$$\left.\begin{aligned} f(x) &= \frac{a^\lambda}{\Gamma\lambda} e^{-ax} x^{\lambda-1} ; a > 0, \lambda > 0 ;\ 0 < x < \infty \\ &= 0, \text{ otherwise} \end{aligned}\right\} \quad \ldots(8\cdot20\,a)$$

We write $X \sim \gamma(a, \lambda)$

Taking $a = 1$ in (8·20 a) we get (8·20). Hence we may write

$X \sim \gamma(\lambda) = (1, \lambda)$.

3. The *cumulative distribution function,* called incomplete gamma function is

$$F_X(x) = \begin{cases} \displaystyle\int_0^x f(u)\, du = \frac{1}{\Gamma\lambda} \int_0^x e^{-u} u^{\lambda-1}\, du, x > 0 \\ 0, \text{ otherwise} \end{cases} \quad \ldots(8\cdot20\,b)$$

8·3·1. M.G.F. of Gamma Distribution. M.G.F. about origin is given by

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} f(x)\, dx = \frac{1}{\Gamma(\lambda)} \int_0^\infty e^{tx} e^{-x} x^{\lambda-1}\, dx$$

$$= \frac{1}{\Gamma(\lambda)} \int_0^\infty e^{-(1-t)x} x^{\lambda-1}\, dx = \frac{1}{\Gamma(\lambda)} \cdot \frac{\Gamma(\lambda)}{(1-t)^\lambda}, \ |t| < 1$$

$$\therefore \quad M_X(t) = (1-t)^{-\lambda}, \ |t| < 1 \quad \ldots(8\cdot21)$$

8·3·2. Cumulant Generating Fucntion of Gamma Distribution. The cumulant generating fucntion $K_X(t)$ is given by

$$K_X(t) = \log M_X(t) = \log(1-t)^{-\lambda} = -\lambda \log(1-t) ; |t| < 1$$

$$= \lambda \left[t + \frac{t^2}{2} + \frac{t^3}{3} + \frac{t^4}{4} + \ldots \right]$$

$\therefore$ Mean $= \kappa_1 =$ Coefficient of t in $K_X(t) = \lambda$

$$\mu_2 = \kappa_2 = \text{Coefficient of } \frac{t^2}{2!} \text{ in } K_X(t) = \lambda$$

$$\kappa_3 = \text{Coefficient of } \frac{t^3}{3!} \text{ in } K_X(t) = 2\lambda$$

$$\kappa_4 = \text{Coefficient of } \frac{t^4}{4!} \text{ in } K_X(t) = 6\lambda$$

$\therefore$ $\mu_4 = \kappa_4 + 3\kappa_2^2 = 6\lambda + 3\lambda^2$

Hence $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{4\lambda^2}{\lambda^3} = \dfrac{4}{\lambda}$ and $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = 3 + \dfrac{6}{\lambda}$

Remarks 1. Like Poisson distribution, the mean and variance of the Gamma distribution are also equal. However, Poisson distribution is discrete while Gamma distribution is continuous.

2. Limiting form of Gamma distribution as $\lambda \to \infty$. We know that if $X \sim \gamma(\lambda)$, then $E(X) = \lambda = \mu$, (say), and $\text{Var}(X) = \lambda = \sigma^2$, (say). Then standard gamma variate is given by

$$Z = \frac{X - \mu}{\sigma} = \frac{X - \lambda}{\sqrt{\lambda}}.$$

$$M_Z(t) = e^{-\mu t/\sigma} M_X(t/\sigma) = e^{-\mu t/\sigma} (1 - t/\sigma)^{-\lambda} \qquad \text{[From (8·21)]}.$$

$$= e^{-t\lambda/\sqrt{\lambda}} \left(1 - \frac{t}{\sqrt{\lambda}}\right)^{-\lambda}$$

$$\Rightarrow \quad K_Z(t) = -\sqrt{\lambda}\, t - \lambda \log\left(1 - \frac{t}{\sqrt{\lambda}}\right)$$

$$= -\sqrt{\lambda}\, t + \lambda \left(\frac{t}{\sqrt{\lambda}} + \frac{t^2}{2\lambda} + \frac{t^3}{3\lambda^{3/2}} + \ldots \right)$$

$$= -\sqrt{\lambda}\, t + \sqrt{\lambda}\, t + \frac{t^2}{2} + o(\lambda^{-1/2})$$

where $O(\lambda^{-1/2})$ are terms containing $\frac{1}{2}$ and higher powers of λ in the denominator.

$$\therefore \quad \lim_{\lambda \to \infty} K_Z(t) = \frac{t^2}{2} \quad \Rightarrow \quad \lim_{\lambda \to \infty} M_Z(t) = e^{t^2/2},$$

which is the m.g.f. of a Standard Normal Variate. Hence by uniqueness theorem of m.g.f., Standard Gamma variate tends to Standard Normal Variate as $\lambda \to \infty$. In other words, Gamma distribution tends to Normal distribution for large values of parameter λ.

3. For the two parameter gamma distribution (8·20 *a*), we have

Exact Sampling Distributions (Chi-square Distribution)

13·1. Chi-Square Variate *(Pronounced as Ki - Sky without S).* The square of a standard normal variate is known as a chi-square variate with 1 degree of freedom (d.f.)

Thus if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$

and $Z^2 = \left(\frac{X - \mu}{\sigma}\right)^2$, is a chi-square variate with 1 d.f.

In general, if X_i, $(i = 1, 2, \ldots, n)$ are n independent normal variates with mean μ_i and variance σ_i^2, $(i = 1, 2, \ldots, n)$, then

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2, \text{ is a chi-square variate with } n \text{ d.f.} \quad \ldots(13\cdot1)$$

13·2. Derivation of the Chi-square Distribution.

First Method—Method of Moment Generating Function.

If X_i, $(i = 1, 2, \ldots, n)$ are independent $N(\mu_i, \sigma_i^2)$, we want the distribution of

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 = \sum_{i=1}^{n} U_i^2, \text{ where } U_i = \frac{X_i - \mu_i}{\sigma_i}$$

Since X_i's are independent, U_i's are also independent.

$$M_{\chi^2}(t) = M_{\sum U_i^2}(t) = \prod_{i=1}^{n} M_{U_i^2}(t) = [M_{U_i^2}(t)]^n,$$

since U_i's $\sim N(0, 1)$ are identically distributed.

Now

$$M_{U_i^2}(t) = E[\exp\{tU_i^2\}] = \int_{-\infty}^{\infty} \exp(tu_i^2) f(x_i) dx_i$$

$$= \int_{-\infty}^{\infty} \exp(tu_i^2) \frac{1}{\sigma\sqrt{2\pi}} \exp\{-(x_i - \mu)^2/2\sigma^2\} dx_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(tu_i^2) \exp(-u_i^2/2)\, du_i \qquad \left[u_i = \frac{x_i - \mu}{\sigma}\right]$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\left(\frac{1-2t}{2}\right)u_i^2\right\} du_i$$

$$= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\pi}\ ^*}{\left(\frac{1-2t}{2}\right)^{1/2}} = (1-2t)^{-1/2}$$

$$\therefore \quad M_{\chi^2}(t) = (1-2t)^{-n/2} \qquad \ldots(13\cdot1a)$$

which is the m.g.f. of a Gamma variate with parameters $\frac{1}{2}$ and $\frac{1}{2}n$.

Hence by uniqueness theorem of m.g.f.'s,

$$\chi^2 = \sum_i^n \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$$

is a Gamma variate with parameters $\frac{1}{2}$ and $\frac{1}{2}n$.

$$\therefore \quad dP(\chi^2) = \frac{(1/2)^{n/2}}{\Gamma(n/2)} \cdot [\exp(-\tfrac{1}{2}\chi^2)]\,(\chi^2)^{(n/2)-1} d\chi^2$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} [\exp(-\chi^2/2)]\,(\chi^2)^{(n/2)-1} d\chi^2,\ 0 \le \chi^2 < \infty \qquad \ldots(13\cdot2)$$

which is the required probability density function of chi-square distribution with n degrees of freedom.

Remarks 1. If a random variable X has a chi-square distribution with n d.f., we write $X \sim \chi^2_{(n)}$ and its *p.d.f.* is given by :

$$f(x) = \frac{1}{2^{n/2}\,\Gamma(n/2)} e^{-x/2}\, x^{(n/2)-1};\ 0 \le x < \infty \qquad \ldots(13\cdot2a)$$

2. If $X \sim \chi^2_{(n)}$, then $(X/2) \sim \gamma(n/2)$.

Proof. The p.d.f. of $Y = \frac{1}{2}X$, is given by :

$$g(y) = f(x) \cdot \left|\frac{dx}{dy}\right|$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} e^{-y} \cdot (2y)^{(n/2)-1} \cdot 2$$

$$= \frac{1}{\Gamma(n/2)} e^{-y}\, y^{(n/2)-1};\ 0 \le y < \infty$$

$$\Rightarrow \quad Y = (X/2) \sim \gamma(n/2)$$

Second Method—Method of Induction

If X_i is a $N(0, 1)$, then $X_i^2/2$ is a $\gamma(1/2)$ so that X_i^2 is a χ^2-variate with d.f. 1.

$^*\int_{-\infty}^{\infty} \exp(-a^2x^2)\,dx = \frac{\sqrt{\pi}}{a}$

If X_1 and X_2 are independent standard normal variates then $X_1^2 + X_2^2$ is a chi-square variate with 2 d.f. which may be proved as follows :

The joint probability differential of X_1 and X_2 is given by :

$$dP(x_1, x_2) = f(x_1, x_2)\, dx_1\, dx_2 = f_1(x_1) f_2(x_2) dx_1\, dx_2$$

$$= \frac{1}{2\pi} \exp\left\{-(x_1^2 + x_2^2)/2\right\} dx_1\, dx_2, -\infty < (x_1, x_2) \leq \infty$$

Let us now transform to polar co-ordinates by the substitution $x_1 = r \cos\theta$, $x_2 = r \sin\theta$. Jacobian of transformation J is given by

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial r} & \frac{\partial x_2}{\partial r} \\ \frac{\partial x_1}{\partial \theta} & \frac{\partial x_2}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{vmatrix} = r$$

Also we have $r^2 = x_1^2 + x_2^2$ and $\tan\theta = x_2/x_1$. As x_1 and x_2 range from $-\infty$ to $+\infty$, r varies from 0 to ∞ and θ from 0 to 2π. The joint probability differential of r and θ now becomes

$$dG(r, \theta) = \frac{1}{2\pi} \exp(-r^2/2)\, r\, dr d\theta\, ; 0 \leq r \leq \infty, 0 \leq \theta \leq 2\pi$$

Integrating over θ, the marginal distribution of r is given by

$$dG_1(r) = \int_0^{2\pi} dG(r, \theta) = r \exp(-r^2/2)\, dr \left[\frac{\theta}{2\pi}\right]_0^{2\pi}$$

$$= \exp(-r^2/2)\, r\, dr$$

$$\Rightarrow \quad dG_1(r^2) = \frac{1}{2} \exp(-r^2/2)\, dr^2$$

$$= \frac{1}{\Gamma(1)} \exp(-r^2/2)\, (r^2/2)^{1-1}\, d(r^2/2)$$

Thus $\frac{r^2}{2} = \frac{X_1^2 + X_2^2}{2}$ is a $\gamma(1)$ variate and hence $r^2 = X_1^2 + X_2^2$ is a χ^2-variate with 2 d.f.

For n variables X_i, $(i = 1, 2, \ldots, n)$ we transform $(X_1, X_2, \ldots, X_n)$ to $(\chi, \theta_1, \theta_2, \ldots, \theta_{n-1})$; (1 – 1 transformation) by

$$\left.\begin{aligned} x_1 &= \chi \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{n-1} \\ x_2 &= \chi \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{n-2} \sin\theta_{n-1} \\ x_3 &= \chi \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{n-3} \sin\theta_{n-2} \\ &\vdots \\ x_j &= \chi \cos\theta_1 \cos\theta_2 \ldots \cos\theta_{n-j} \sin\theta_{n-j+1} \\ &\vdots \\ x_n &= \chi \sin\theta_1 \end{aligned}\right\} \quad \ldots(13\cdot3$$

where $\chi > 0, -\pi < \theta_1 < \pi$ and $-\pi/2 < \theta_i < \pi/2$ for $i = 2, 3, \ldots (n-1)$.

Then $$x_1^2 + x_2^2 + \ldots + x_n^2 = \chi^2$$

and $$|J| = \chi^{n-1} \cos^{n-2}\theta_1 \cos^{n-3}\theta_2 \ldots \cos\theta_{n-2}$$

(*c.f.* Advanced Theory of Statistics Vol 1, by Kendall and Stuart.)

The joint distribution of $X_1, X_2, \ldots, X_n$ *viz.*,

$$dF(x_1, x_2, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp(-\Sigma x_i^2/2) \prod_{i=1}^{n} dx_i$$

transforms to

$$dG(\chi, \theta_1, \theta_2, \ldots, \theta_{n-1}) = \exp(-\chi^2/2)\, \chi^{n-1} \cos^{n-2}\theta_1 \cos^{n-3}\theta_2 \ldots \cos\theta_{n-2}\, d\chi d\theta_1 d\theta_2 \ldots d\theta_{n-1}$$

Integrating over $\theta_1, \theta_2, \ldots \theta_{n-1}$, we get the distribution of χ^2 as

$$dP(\chi^2) = k \exp(-\chi^2/2)\, (\chi^2)^{(n/2)-1}\, d\chi^2,\ 0 \le \chi^2 < \infty$$

The constant k is determined from the fact that the total probability is unity, *i.e.*,

$$\int_0^\infty dP(\chi^2) = 1 \Rightarrow k \int_0^\infty \exp(-\chi^2/2)\, (\chi^2)^{\frac{n}{2}-1}\, d\chi^2 = 1$$

$$\Rightarrow \qquad k = \frac{1}{2^{n/2}\,\Gamma(n/2)}$$

$$\therefore \qquad dP(\chi^2) = \frac{1}{2^{n/2}\,\Gamma(n/2)} \exp(-\chi^2/2)\, (\chi^2)^{\frac{n}{2}-1},\ 0 \le \chi^2 < \infty$$

Hence $$\frac{\chi^2}{2} = \frac{1}{2}\sum_{i=1}^{n} X_i^2$$ is a $\gamma(n/2)$ variate.

$$\Rightarrow \qquad \chi^2 = \sum_{i=1}^{n} X_i^2$$ is a chi-square variate with n degrees of freedom (d.f.) and (13·2) gives p.d.f. of chi-square distribution with n d.f.

Remarks 1. If X_i ; $i = 1, 2, \ldots, n$ are n independent normal variates with mean μ_i and S.D. σ_i, then $\sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$ is a χ^2-variate with n d.f.

2. In random sampling from a normal population with mean μ and S.D. σ, $\bar{x}$ is distributed normally about the mean μ with S.D. $\sigma/\sqrt{n}$.

$$\therefore \qquad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$\Rightarrow \qquad \left[\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right]^2$$ is a χ^2-variate with 1 d.f.

3. Normal distribution is a particular case of χ^2-distribution when $n = 1$, since for $n = 1$,

$$p(\chi^2) = \frac{1}{\sqrt{2}\,\Gamma(1/2)} \exp(-\chi^2/2)\,(\chi^2)^{\frac{1}{2}-1}\,d\chi^2,\ 0 \le \chi^2 < \infty$$

$$= \frac{1}{\sqrt{2\pi}} \exp(-\chi^2/2)\,d\chi,\ -\infty \le \chi < \infty$$

Thus χ is a standard normal variate.

13·3. M.G.F. of χ^2-distribution. Let $X \sim \chi^2_{(n)}$. then

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} f(x)dx$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} \int_0^\infty e^{tx} \cdot e^{-x/2}\, x^{(n/2)-1}\,dx$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} \int_0^\infty \exp\left[-\left(\frac{1-2t}{2}\right)x\right] \cdot x^{(n/2)-1}\,dx$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} \frac{\Gamma(n/2)}{[(1-2t)/2]^{n/2}} \quad \text{[Using Gamma Integral]}$$

$$= (1-2t)^{-n/2},\ |2t| < 1 \qquad \ldots(13\cdot4)$$

which is the required m.g.f. of a χ^2-variate with n d.f.

Remarks 1. Using Binomial expansion for negative index, we get from (13·4) if $|t| < \frac{1}{2}$.

$$M(t) = 1 + \frac{n}{2}(2t) + \frac{\frac{n}{2}\left(\frac{n}{2}+1\right)}{2!}(2t)^2 + \ldots + \frac{\frac{n}{2}\left(\frac{n}{2}+1\right)\left(\frac{n}{2}+2\right)\ldots\left(\frac{n}{2}+r-1\right)}{r!}(2t)^r + \ldots$$

$$\mu_r' = \text{Coefficient of } \frac{t^r}{r!} \text{ in the expansion of } M(t)$$

$$= 2^r \frac{n}{2}\left(\frac{n}{2}+1\right)\left(\frac{n}{2}+2\right)\ldots\left(\frac{n}{2}+r-1\right)$$

$$= n(n+2)(n+4)\ldots(n+2r-2) \qquad \ldots(13\cdot4a)$$

Remark. If n is even so that $n/2$ is a positive integer, then

$$\mu_r' = 2^r\,\Gamma[(n/2)+r]/\Gamma(n/2) \qquad \ldots(13\cdot4b)$$

13·3·1. Cumulant Generating Function of χ^2-distribution. If $X \sim \chi^2_{(n)}$, then

$$K_{\chi^2}(t) = \log M_X(t) = -\frac{n}{2}\log(1-2t)$$

$$= \frac{n}{2}\left[2t + \frac{(2t)^2}{2} + \frac{(2t)^3}{3} + \frac{(2t)^4}{4} + \ldots\right]$$

$\therefore$ κ_1 = Coefficient of t in $K(t) = n$

κ_2 = Coefficient of $\frac{t^2}{2!}$ in $K(t) = 2n$

κ_3 = Coefficient of $\frac{t^3}{3!}$ in $K(t) = 8n$

κ_4 = Coefficient of $\frac{t^4}{4!}$ in $K(t) = 48n$

In general,

$$\kappa_r = \text{Coefficient of } \frac{t^r}{r!} \text{ in } K(t) = n\,2^{r-1}(r-1)! \qquad \ldots(13\cdot4c)$$

Hence

$$\left.\begin{aligned} &\text{Mean} = \kappa_1 = n, \ \text{Variance} = \mu_2 = \kappa_2 = 2n \\ &\mu_3 = \kappa_3 = 8n, \ \mu_4 = \kappa_4 + 3\kappa_2^2 = 48n + 12n^2 \\ &\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{8}{n} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{12}{n} + 3 \end{aligned}\right\} \qquad \ldots(13\cdot4d)$$

13·3·2. Limiting Form of χ^2 Distribution for Large Degrees of Freedom. If $X \sim \chi^2_{(n)}$, then $M_X(t) = (1-2t)^{-n/2}, |t| < \frac{1}{2}$

The m.g.f. of standard χ^2-variate Z is given by

$$M_{\frac{X-\mu}{\sigma}}(t) = e^{-\mu t/\sigma} M_X(t/\sigma) \qquad [\mu = n, \sigma^2 = 2n]$$

or $$M_Z(t) = e^{-\mu t/\sigma}(1 - 2t/\sigma)^{-n/2}$$

$$= e^{-nt/\sqrt{2n}}\left(1 - \frac{2t}{\sqrt{2n}}\right)^{-n/2}$$

$\therefore$ $$K_Z(t) = \log M_Z(t) = -t\sqrt{\frac{n}{2}} - \frac{n}{2}\log\left(1 - t\sqrt{\frac{2}{n}}\right)$$

$$= -t\sqrt{\frac{n}{2}} + \frac{n}{2}\left[t\cdot\sqrt{\frac{2}{n}} + \frac{t^2}{2}\cdot\frac{2}{n} + \frac{t^3}{3}\left(\frac{2}{n}\right)^{3/2} + \ldots\right]$$

$$= -t\sqrt{\frac{n}{2}} + t\cdot\sqrt{\frac{n}{2}} + \frac{t^2}{2} + 0(n^{-1/2})$$

$$= \frac{t^2}{2} + 0(n^{-1/2}),$$

where $0(n^{-1/2})$ are terms containing $n^{1/2}$ and higher powers of n in the denominator.

$$\therefore \lim_{n\to\infty} K_Z(t) = \frac{t^2}{2} \Rightarrow M_Z(t) = e^{t^2/2}, \text{ as } n \to \infty,$$

which is the m.g.f. of a standard normal variate. Hence by uniqueness theorem of m.g.f. Z is asymptotically normal. In other words, standard χ^2 variate tends to standard normal variate as $n \to \infty$. Thus, χ^2-distribution tends to normal distribution for large d.f.

In practice for $n \geq 30$, the χ^2-approximation to normal distribution is fairly good. So whenever $n \geq 30$, we use the normal probability tables for testing the significance of the value of χ^2. That is why in the tables given in the Appendix, the significant values of χ^2 have been tabulated till $n = 30$ only.

Remark. For the distribution of χ^2-variate for large values of n, see Example 13·7 and also Remark 2 to § 13·7·1.

13·3·3. Characteristic Function of χ^2-distribution.

If $X \sim \chi^2_{(n)}$, then

$$\phi_X(t) = E\{\exp(itX)\} = \int_0^\infty \exp(itx) f(x)\, dx$$

$$= \frac{1}{2^{n/2}\,\Gamma(n/2)} \int_0^\infty \exp\left\{-\left(\frac{1-2it}{2}\right)x\right\} (x)^{\frac{n}{2}-1}\, dx$$

$$= (1-2it)^{-n/2} \qquad \ldots(13\cdot4e)$$

13·3·4. Mode and skewness of χ^2-distribution.

Let $X \sim \chi^2_{(n)}$, so that

$$f(x) = \frac{1}{2^{n/2}\,\Gamma(n/2)}\, e^{-x/2}\, x^{(n/2)-1},\ 0 \leq x < \infty \qquad \ldots(*)$$

Mode of the distribution is the solution of

$$f'(x) = 0 \quad \text{and} \quad f''(x) < 0$$

Logarithmic differentiation w.r.t. x in (*) gives :

$$\frac{f'(x)}{f(x)} = 0 - \frac{1}{2} + \left(\frac{n}{2} - 1\right) \cdot \frac{1}{x} = \frac{n-2-x}{2x} \qquad \ldots(13\cdot5)$$

Since $f(x) \neq 0$, $f'(x) = 0 \Rightarrow x = n-2$

It can be easily seen that at the point, $x = (n-2)$, $f''(x) < 0$.

Hence mode of the chi-square distribution with n d.f. is $(n-2)$.

Also Karl Pearson's coefficient of skewness is given by

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = \frac{n-(n-2)}{\sqrt{2n}} = \sqrt{\frac{2}{n}} \qquad \ldots(13\cdot6)$$

Since Pearson's coefficient of skewness is greater than zero for $n \geq 1$, the χ^2-distribution is positively skewed. Further since skewness is inversely proportional to the square root of d.f., it rapidly tends to symmetry as the d.f. increases and consequently as $n \to \infty$, the chi-square distribution tends to normal distribution.

13·3·5. Additive Property of χ^2-variates. *The sum of independent chi-square variates is also a χ^2-variate More precisely, if X_i, $(i = 1, 2, \ldots, k)$ are*

independent χ^2-variates with n_i d.f. respectively, then the sum $\sum_{i=1}^{k} X_i$ is also a chi-square variate with $\sum_{i=1}^{k} n_i$ d.f.

Proof. We have

$$M_{X_i}(t) = (1-2t)^{-n_i/2}; i = 1, 2, \ldots, k.$$

The *m.g.f.* of the sum $\sum_{i=1}^{k} X_i$ is given by

$$\begin{aligned} M_{\Sigma X_i}(t) &= M_{X_1}(t)\, M_{X_2}(t) \ldots M_{X_k}(t) \qquad [\because X_i\text{'s are independent}] \\ &= (1-2t)^{-n_1/2} (1-2t)^{-n_2/2} \ldots (1-2t)^{-n_k/2} \\ &= (1-2t)^{-(n_1+n_2+\ldots+n_k)/2} \end{aligned}$$

which is the *m.g.f.* of a χ^2-variate with $(n_1 + n_2 + \ldots + n_k)$ *d.f.* Hence by uniqueness theorem of *m.g.f.*'s, $\sum_{i=1}^{k} X_i$ is a χ^2-variate with $\sum_{i=1}^{k} n_i$ *d.f.*

Remarks 1. Converse is also true, *i.e.*, if X_i ; $i = 1, 2, \ldots, k$ are χ^2-variates with n_i ; $i = 1, 2, \ldots, k$ *d.f.* respectively and if $\sum_{i=1}^{k} X_i$ is a χ^2-variate with $\sum_{i=1}^{k} n_i$ *d.f.*, then X_i's are independent.

2. Another useful version of the converse is as follows :

If X and Y are independent non-negative variates such that $X + Y$ follows chi-square distribution with $n_1 + n_2$ *d.f.* and if one of them, say, X is a χ^2-variate with n_1 *d.f.* then the other, *viz.*, Y, is a χ^2-variate with n_2 *d.f.*

Proof. Since X and Y are independent variates, we have

$$\begin{aligned} M_{X+Y}(t) &= M_X(t)\, M_Y(t) \\ \Rightarrow \quad (1-2t)^{-(n_1+n_2)/2} &= (1-2t)^{-n_1/2} \cdot M_Y(t) \\ &\qquad \left[\because X + Y \sim \chi^2_{(n_1+n_2)} \text{ and } X \sim \chi^2_{(n_1)}\right] \\ \Rightarrow \quad M_Y(t) &= (1-2t)^{-n_2/2} \end{aligned}$$

which is the m.g.f. of χ^2-variate with n_2 *d.f.* Hence by uniqueness theorem of m.g.f.'s, $Y \sim \chi^2_{(n_2)}$

3. Still another form of the above theorem is "*Cochran theorem*" which is as follows :

Let $X_1, X_2, \ldots, X_n$ be independently distributed as standard normal variates, *i.e.*, $N(0, 1)$. Let

$$\sum_{i=1}^{n} X_i^2 = Q_1 + Q_2 + \ldots + Q_k$$

where each Q_i is a sum of squares of linear combinations of $X_1, X_2, \ldots, X_n$ with n_i degrees of freedom. Then if $n_1 + n_2 + \ldots + n_k = n$, the quantities $Q_1, Q_2 \ldots, Q_k$ are independent χ^2-variates with $n_1, n_2, \ldots, n_k$ d.f. respectively.

13·4. Chi-square Probability Curve. We get from (13·5)

$$f'(x) = \left[\frac{n - 2 - x}{2x}\right] f(x). \qquad \ldots(13·7)$$

Since $x > 0$ and $f(x)$ being p.d.f. is always non-negative, we get from (13·7) :

$$f'(x) < 0 \quad \text{if} \quad (n - 2) \leq 0,$$

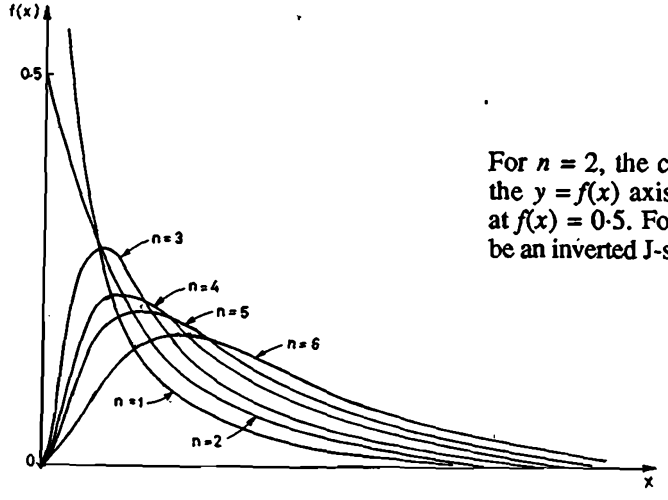
for all values of x. Thus the χ^2-probability curve for 1 and 2 degrees of freedom is monotonically decreasing.

When $n > 2$,

$$f'(x) = \begin{cases} > 0, \text{ if } x < (n - 2) \\ = 0, \text{ if } x = n - 2 \\ < 0, \text{ if } x > (n - 2) \end{cases}$$

This implies that for $n > 2$, $f(x)$ is monotonically increasing for $0 < x < (n - 2)$ and monotonically decreasing for $(n - 2) < x < \infty$, while at $x = n - 2$, it attains the maximum value.

For $n \geq 1$, as x increases, $f(x)$ decreases rapidly and finally tends to zero as $x \to \infty$. Thus for $n > 1$, the χ^2-probability curve is positively skewed [*c.f.* (13·6)] towards higher values of x. Moreover, x-axis is an asymptote to the curve. The shape of the curve for $n = 1, 2, 3, \ldots, 6$ is given below.



For $n = 2$, the curve will meet the $y = f(x)$ axis at $x = 0$, *i.e.*, at $f(x) = 0·5$. For $n = 1$, it will be an inverted J-shaped curve.

PROBABILITY CURVE OF CHI-SQUARE DISTRIBUTION

Theorem 13·1. *If χ_1^2 and χ_2^2 are two independent χ^2-variates with n_1 and n_2 d.f. respectively, then*

$$\frac{\chi_1^2}{\chi_2^2} \text{ is a } \beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right) \text{variate.}$$

(Gauhati Univ. M.Sc., 1992)

Proof. Since χ_1^2 and χ_2^2 are independent χ^2-variates with n_1 and n_2 d.f. respectively, their joint probability differential is given by the compound probability theorem as

$$dP(\chi_1^2, \chi_2^2) = dP_1(\chi_1^2)\, dP_2(\chi_2^2)$$

$$= \left[\frac{1}{2^{n_1/2}\Gamma(n_1/2)} \exp(-\chi_1^2/2)\,(\chi_1^2)^{(n_1/2)-1} d\chi_1^2\right]$$

$$\times \left[\frac{1}{2^{n_2/2}\,\Gamma(n_2/2)} \exp(-\chi_2^2/2)\,(\chi_2^2)^{(n_2/2)-1}\, d\chi_2^2\right]$$

$$= \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \exp\{-(\chi_1^2 + \chi_2^2)/2\}$$

$$\times (\chi_1^2)^{\frac{n_1}{2}-1} (\chi_2^2)^{\frac{n_2}{2}-1}\, d\chi_1^2 d\chi_2^2,\ 0 \le (\chi_1^2, \chi_2^2) < \infty$$

Let us make the transformation :

$u = \chi_1^2/\chi_2^2$ and $v = \chi_2^2$

so that $\chi_1^2 = uv$ and $\chi_2^2 = v$

Jacobian of transformation J is given by

$$J = \frac{\partial(\chi_1^2, \chi_2^2)}{\partial(u, v)} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v$$

Thus the joint distribution of random variables U and V becomes

$$dG(u, v) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \exp\{-(1+u)v/2\}$$

$$\times (uv)^{\frac{n_1}{2}-1} v^{\frac{n_2}{2}-1} |J|\, du\, dv,$$

$$= \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \exp\{-(1+u)v/2\}$$

$$\times u^{\frac{n_1}{2}-1} v^{\frac{n_1+n_2}{2}-1} du\, dv,\ 0 \le (u, v) < \infty$$

Integrating w.r.t. v over the range 0 to ∞, we get the marginal distribution of U as : $dG_1(u) = \int_0^\infty dG(u, v)$

$$= \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)}\, u^{(n_1/2)-1} du$$

$$\times \int_0^\infty \exp\left\{-\left(\frac{1+u}{2}\right)v\right\} v^{\{(n_1+n_2)/2\}-1} dv$$

$$= \frac{u^{(n_1/2)-1}}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \cdot \frac{\Gamma\{(n_1+n_2)/2\}}{[(1+u)/2]^{(n_1+n_2)/2}}\,du$$

$$= \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{u^{(n_1/2)-1}}{[+u]^{(n_1+n_2)/2}}\,du,\ 0 \le u < \infty$$

Hence $U = \dfrac{\chi_1^2}{\chi_2^2}$ is a $\beta_2\left(\dfrac{n_1}{2}, \dfrac{n_2}{2}\right)$ variate.

Theorem 13·2. *If χ_1^2 and χ_2^2 are independent χ^2-variates with n_1 and n_2 d.f. respectively, then*

$$U = \frac{\chi_1^2}{\chi_1^2 + \chi_2^2} \quad \text{and} \quad V = \chi_1^2 + \chi_2^2$$

are independently distributed, U as a $\beta_1\left(\dfrac{n_1}{2}, \dfrac{n_2}{2}\right)$ variate and V as a χ^2 variate with $(n_1 + n_2)$ d.f.

Proof. As the Theorem 13·1, we have

$$dP(\chi_1^2, \chi_2^2) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \exp\{-(\chi_1^2 + \chi_2^2)/2\} \times (\chi_1^2)^{(n_1/2)-1} (\chi_2^2)^{(n_2/2)-1}\,d\chi_1^2\,d\chi_2^2,\ 0 \le (\chi_1^2, \chi_2^2) < \infty$$

Let us transform to u and v defined as follows :

$$u = \frac{\chi_1^2}{\chi_1^2 + \chi_2^2} \quad \text{and} \quad v = \chi_1^2 + \chi_2^2$$

so that $\chi_1^2 = uv$ and $\chi_2^2 = v - \chi_1^2 = (1-u)v$

As χ_1^2 and χ_2^2 both range from 0 to ∞, u ranges from 0 to 1 and v from 0 to ∞.

Jacobian of transformation J is

$$J = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v$$

$$\therefore\ dG(u, v) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} \exp(-v/2)\,(uv)^{(n_1/2)-1} \times [(1-u)v]^{(n_2/2)-1} |J|\,du\,dv$$

$$= \frac{1}{2^{(n_1+n_2)/2}\,\Gamma(n_1/2)\,\Gamma(n_2/2)} u^{(n_1/2)-1}\,(1-u)^{(n_2/2)-1} \times \exp(-v/2).\ v^{\{(n_1+n_2)/2\}-1}\,du\,dv$$

$$= \left[\frac{\Gamma\{(n_1+n_2)/2\}}{\Gamma(n_1/2)\Gamma(n_2/2)} u^{(n_1/2)-1}\,(1-u)^{(n_2/2)-1}\,du\right] \times \left[\frac{1}{2^{(n_1+n_2)/2}\,\Gamma\{(n_1+n_2)/2\}} \exp(-v/2)\,v^{\{(n_1+n_2)/2\}-1}\,dv\right]$$

Since the joint probability differential of U and V is the product of their respective probability differentials, U and V are independently distributed, with

$$dG_1(u) = \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} u^{(n_1/2)-1} (1-u)^{(n_2/2)-1} du, \quad 0 \le u \le 1$$

and

$$dG_2(v) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma\{(n_1+n_2)/2\}} \exp(-v/2)\, v^{\{(n_1+n_2)/2\}-1} dv, \quad 0 \le v < \infty$$

i.e., U as a $\beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ variate and V as a χ^2-variate with $(n_1 + n_2)$ d.f

Remark. The results in Theorems 13·1 and 13·2 can be summarised as follows :

If $X \sim \chi^2_{(n_1)}$ and $Y \sim \chi^2_{(n_2)}$ are independent chi-square variates then :

(*i*) $X + Y \sim \chi^2_{(n_1+n_2)}$ *i.e.*, the sum of two independent chi-square variates is also a chi-square variate.

(*ii*) $\frac{X}{Y} \sim \beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ *i.e.*, the ratio of two independent chi-square variates is a β_2-variate.

(*iii*) $\frac{X}{X+Y} \sim \beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$

Theorem 13·3. *In a random and large sample,*

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(n_i - np_i)^2}{np_i}\right], \qquad \ldots(13\cdot8)$$

follows chi-square distribution approximately with $(k-1)$ degrees of freedom, where n_i is the observed frequency and np_i is the corresponding expected frequency of the ith class, $(i = 1, 2, \ldots, k)$, $\sum_{i=1}^{k} n_i = n$.

Proof. Let us consider a random sample of size n, whose members are distributed at random in k classes or cells. Let p_i be the probability that sample observation will fall in the i th cell, $(i = 1, 2, \ldots, k)$. Then the probability P of there being n_i members in the ith cell, $(i = 1, 2, \ldots, k)$ respectively is given by the multinomial probability law, by the expression

$$P = \frac{n!}{n_1!\ n_2!\ \ldots\ n_k!}\, p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k},$$

where $\sum_{i=1}^{k} n_i = n$ and $\sum_{i=1}^{k} p_i = 1$.

If n is sufficiently large so that n_i, $(i = 1, 2, \ldots, k)$ are not small then using Stirling's approximation to factorials for large n, *viz.*,

$\lim_{n \to \infty} (n!) \approx \sqrt{2\pi}\, e^{-n}\, n^{n + \frac{1}{2}}$, we get

$$P \approx \frac{\sqrt{2\pi}\, e^{-n}\, n^{n+\frac{1}{2}}}{(\sqrt{2\pi})^k e^{-(n_1 + n_2 + \ldots + n_k)}} \times \frac{p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}}{n_1^{n_1 + \frac{1}{2}}\, n_2^{n_2 + \frac{1}{2}} \ldots\, n_k^{n_k + \frac{1}{2}}}$$

$$\approx \frac{e^{-n}\, n^{n+\frac{1}{2}} \left(\frac{np_1}{n_1}\right)^{n_1 + \frac{1}{2}} \left(\frac{np_2}{n_2}\right)^{n_2 + \frac{1}{2}} \cdots \left(\frac{np_k}{n_k}\right)^{n_k + \frac{1}{2}}}{(\sqrt{2\pi})^{k-1}\, e^{-n}\, n^{n_1 + n_2 + \ldots + n_k + (k/2)}\, (p_1 p_2 \ldots p_k)^{1/2}}$$

$$\approx C \prod_{i=1}^{k} \left(\frac{np_i}{n_i}\right)^{n_i + \frac{1}{2}}$$

where $C = \dfrac{1}{(2\pi)^{(k-1)/2}\, n^{(k-1)/2}\, (p_1 p_2 \ldots p_k)^{1/2}}$,

is a constant independent of n_i's.

$$\therefore \qquad \log P \approx \log C + \sum_{i=1}^{k} (n_i + \tfrac{1}{2}) \log \left(\frac{np_i}{n_i}\right)$$

$$\Rightarrow \qquad \log (P/C) \approx \sum_{i=1}^{k} (n_i + \tfrac{1}{2}) \log \left(\frac{\lambda_i}{n_i}\right), \qquad \ldots(*)$$

where $\lambda_i = np_i$ is the expected frequency for the ith cell, *i.e.*,

$$E(n_i) = np_i = \lambda_i, \ (i = 1, 2, \ldots, k).$$

Let us define

$$\xi_i = \frac{n_i - \lambda_i}{\sqrt{\lambda_i}},$$

so that $n_i - \lambda_i = \xi_i \sqrt{\lambda_i} \quad \Rightarrow \quad n_i = \lambda_i + \xi_i \sqrt{\lambda_i} \qquad \ldots(**)$

Substituting in (*), we get

$$\log (P/C) \approx \sum_{i=1}^{k} (\lambda_i + \xi_i \sqrt{\lambda_i} + \tfrac{1}{2}) \log \left[\frac{\lambda_i}{\lambda_i + \xi_i \sqrt{\lambda_i}}\right]$$

$$= \sum_{i=1}^{k} (\lambda_i + \xi_i \sqrt{\lambda_i} + \tfrac{1}{2}) \log [1/\{1 + \xi_i/\sqrt{\lambda_i}\}]$$

$$= -\sum_{i=1}^{k} (\lambda_i + \xi_i \sqrt{\lambda_i} + \tfrac{1}{2}) \log \{1 + (\xi_i/\sqrt{\lambda_i})\}$$

If we assume that ξ_i is small compared with λ_i, the expansion of $\log 1 + (\xi_i/\sqrt{\lambda_i})$ in ascending powers of $\xi_i/\sqrt{\lambda_i}$ is valid.

$$\therefore \quad \log P/C \approx -\sum_{i=1}^{k} (\lambda_i + \xi_i\sqrt{\lambda_i} + \tfrac{1}{2}\lambda_i)\left[\frac{\xi_i}{\sqrt{\lambda_i}} - \frac{1}{2}\frac{\xi_i^2}{\lambda_i} + O\,(1/\lambda_i^{3/2})\right]$$

$$\approx -\sum_{i=1}^{k} \left[\xi_i\sqrt{\lambda_i} - \tfrac{1}{2}\xi_i^2 + \xi_i^2 + O(\lambda_i^{-1/2})\right],$$

neglecting higher powers of $\xi_i/\sqrt{\lambda_i}$ if ξ_i is small compred with λ_i.
Since n is large, so is $\lambda_i = np_i$. Hence $O(\lambda_i^{-1/2}) \to 0$ for large n.

Also $$\sum_{i=1}^{k} \xi_i\sqrt{\lambda_i} = \sum_{i=1}^{k} (n_i - \lambda_i) = \sum_{i=1}^{k} n_i - \sum_{i=1}^{k} \lambda_i$$

$$= \sum_{i=1}^{k} n_i - n\sum_{i=1}^{k} p_i = n - n = 0 \qquad (\because \Sigma n_i = n,\ \Sigma p_i = 1)$$

$$\therefore \quad \log (P/C) \approx -\left[\sum_{i=1}^{k} \xi_i\sqrt{\lambda_i} + \frac{1}{2}\sum_{i=1}^{k} \xi_i^2 + O(\lambda_i^{-1/2})\right] \approx -\frac{1}{2}\sum_{i=1}^{k} \xi_i^2$$

$$\Rightarrow \qquad P \approx C \exp\left(-\frac{1}{2}\sum_{i=1}^{k} \xi_i^2\right),$$

which shows that ξ_i, $(i = 1, 2, \ldots, k)$ are distributed as independent standard normal variates.

Hence $$\sum_{i=1}^{k} \xi_i^2 = \sum_{i=1}^{k} \left[\frac{(n_i - \lambda_i)^2}{\lambda_i}\right],$$

being the sum of the squares of k independent standard normal variates is a χ^2-variate with $(k-1)$ d.f., one d.f. being lost because of the linear constraint

$$\sum_{i=1}^{k} \xi_i\sqrt{\lambda_i} = \Sigma(n_i - \lambda_i) = 0 \Rightarrow \sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \lambda_i \qquad \ldots(***)$$

Remarks 1. If O_i and E_i $(i = 1, 2, \ldots, k)$, be a set of observed and expected frequencies, then

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i}\right], \quad \left(\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i\right) \qquad \ldots(13\cdot8a)$$

follows chi-square distribution with $(k-1)$ d.f

Another convenient form of this formula is as follows :

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{O_i^2 + E_i^2 - 2O_iE_i}{E_i}\right) = \sum_{i=1}^{k} \left(\frac{O_i^2}{E_i} + E_i - 2O_i\right)$$

$$= \sum_{i=1}^{k} \left(\frac{O_i^2}{E_i}\right) + \sum_{i=1}^{k} E_i - 2\sum_{i=1}^{k} O_i$$

$$= \sum_{i=1}^{k} \left(\frac{O_i^2}{E_i} \right) - N, \qquad \ldots(13\cdot8b)$$

where $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = N$ (say), is the total frequency.

2. Conditions for the Validity of χ^2-test. χ^2-test is an approximate test for large values of n. For the validity of chi-square test of '*goodness of fit*' between theory and experiment, the following conditions must be satisfied :

(*i*) The sample observations should be independent.

(*ii*) Constraints on the cell frequencies, if any, should be linear, *e.g.*, $\Sigma n_i = \Sigma \lambda_i$ or $\Sigma O_i = \Sigma E_i$.

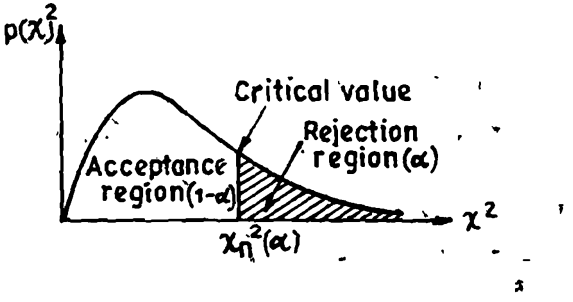
(*iii*) N, the total frequency should be reasonably large, say, greater than 50.

(*iv*) No theoretical cell frequency should be less than 5. (The chi square distribution is essentially a continuous distribution but it cannot maintain its character of continuity if cell frequency is less than 5). If any theoretical cell frequency is less than 5, then for the application of χ^2-test, it is pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and finally adjust for the d.f. lost in pooling.

3. It may be noted that the χ^2-test depends only on the set of observed and expected frequencies and on degrees of freedom (*d.f.*). It does not make any assumptions regarding the parent population from which the observations are taken. Since χ^2 defined in (13·8) does not involve any population parameters, it is termed as a statistic and the test is known as *Non-Parametric Test* or *Distribution-Free Test.*

4. Critical Values. Let $\chi_n^2(\alpha)$ denote the value of chi-square for *n.d.f.* such that the area to the right of this point is α, *i.e.*,

$$P[\chi^2 > \chi_n^2(\alpha)] = \alpha \qquad \ldots(13\cdot8c)$$



The value $\chi_n^2(\alpha)$ defined in (13·8c) is known as the *upper (right-tailed) α-point or Critical Value or Significant Value of chi-square for n d.f.* and has been tabulated for different values of n and α in Table VI in the Appendix at the end of the book. From these tables we observe that the critical values of χ^2 increase as n (d.f.) increases and level of significance (α) decreases.

The critical values for left-tailed test or two tailed tests can be obtained from the above table, as discussed in Remark 1 to § 16·7·4.

13·6. Linear Transformation. Let us suppose that the given set of variables $X' = (x_1, x_2, \ldots x_n)$ is transformed to a new set of variables $Y' = (y_1, y_2, \ldots, y_n)$ by means of the linear transformation :

$$\left.\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ &\vdots \\ y_n &= a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n \end{aligned}\right\} \quad \ldots(13\cdot9)$$

i.e., $$y_i = a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n \; ; i = 1, 2, \ldots, n$$

In matrix notation, this system of linear equations can be expressed symbolically as

$$\mathbf{Y} = \mathbf{AX} \qquad \ldots(13\cdot10)$$

where $$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \; \mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \; \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \ldots a_{1n} \\ a_{21} & a_{22} \ldots a_{2n} \\ \vdots & \vdots \quad \vdots \\ a_{n1} & a_{n2} \ldots a_{nn} \end{pmatrix}$$

From matrix theory, we know that the system (13·10) has a unique solution iff $|\mathbf{A}| \neq 0$. In other words, we can express X uniquely in terms Y if A is non-singular and the solution is given by

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{Y} \qquad \ldots(13\cdot10a)$$

where $\mathbf{A}^{-1}$ is the inverse of the square matrix A.

The linear transformation defined in (13·9) or (13·10) is said to be *orthogonal* if

$$\mathbf{X'X} = \mathbf{Y'Y} \qquad \ldots(13\cdot11)$$

$$\Rightarrow \quad \mathbf{X'X} = (\mathbf{AX})'\,\mathbf{AX} = \mathbf{X'}(\mathbf{A'A})\mathbf{X}$$

$$\Rightarrow \quad \mathbf{A'A} = \mathbf{I}_n \qquad \ldots(13\cdot11a)$$

$\Rightarrow$ A is an orthogonal matrix.

More elaborately

$$\mathbf{X'X} = \mathbf{Y'Y}$$

$$\Rightarrow \quad \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{in}x_n)^2, \qquad \ldots(*)$$

for every set of variables, $(x_1, x_2, \ldots, x_n)$.

If we write $\delta_{ij} = \sum_{k=1}^{n} a_{ik}\,a_{kj}, \; (i, j = 1, 2, \ldots, n)$,

then (*) implies that δ_{ij} is a Kronecker delta so that

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \qquad \ldots(13\cdot11b)$$

whence it follows that A is an orthogonal matrix.

Exact Sampling Distributions

(CONTINUED)

(t, F AND Z DISTRIBUTIONS)

14·1. Introduction. The entire large sample theory was based on the application of *"Normal Test"* (*c.f.* § 12·9). However, if the sample size n is small, the distribution of the various statistics, *e.g.*, $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ or $Z = (X - nP)/\sqrt{nPQ}$ etc., are far from normality and as such *'normal test'* cannot be applied if n is small. In such cases exact sample tests, pioneered by W.S. Gosset (1908) who wrote under the pen name of Student, and later on developed and extended by Prof. R.A. Fisher (1926), are used. In the following sections we shall discuss

(i) t-test, (ii) F-test, and (iii) Fisher's z-transformation.

The exact sample tests can, however, be applied to large samples also though the converse is not true. In all the exact sample tests, the basic assumption is that *"The population(s) from which sample(s) are drawn is (are) normal, i.e., the parent population(s) is (are) normally distributed."*

14·2. Student's 't'. *Definition.* Let x_i, $(i = 1, 2, \ldots, n)$ be a random sample of size n from a normal population with mean μ and variance σ^2. Then Student's t is defined by the statistic

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \qquad \ldots(14\cdot1)$$

where $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, is the sample mean and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad \ldots(14\cdot1a)$$

is an unbiased estimate of the population variance σ^2, and it follows Student's t-distribution with $\nu = (n-1)$ *d.f.* with probability density function,

$$f(t) = \frac{1}{\sqrt{\nu}\, B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{\nu}\right]^{(\nu+1)/2}}; \quad -\infty < t < \infty \qquad \ldots(14\cdot2)$$

Remarks 1. A statistic t following Student's t-distribution with n d.f. will be abbreviated as $t \sim t_n$.

2. If we take $\nu = 1$ in (14·2), we get

$$f(t) = \frac{1}{B\left(\frac{1}{2}, \frac{1}{2}\right)} \cdot \frac{1}{(1 + t^2)},$$

$$= \frac{1}{\pi} \cdot \frac{1}{(1 + t^2)}; -\infty < t < \infty \qquad [\because \Gamma(\tfrac{1}{2}) = \sqrt{\pi}]$$

which is the p.d.f. of standard Cauchy distribution. Hence, when $\nu = 1$ Student's t distribution reduces to Cauchy distribution.

14·2·1. Derivation of Student's t-distribution. The expression (14·1) can be re-written as

$$t^2 = \frac{n(\bar{x} - \mu)^2}{S^2} = \frac{n(\bar{x} - \mu)^2}{ns^2/(n-1)} \qquad [\because ns^2 = (n-1)\,S^2]$$

$$\Rightarrow \quad \frac{t^2}{(n-1)} = \frac{(\bar{x} - \mu)^2}{\sigma^2/n} \quad \frac{1}{ns^2/\sigma^2} = \frac{(\bar{x} - \mu)^2/(\sigma^2/n)}{ns^2/\sigma^2}$$

Since x_i, $(i = 1, 2, \ldots, n)$ is a random sample from the normal population with mean μ and variance σ^2,

$$\bar{x} \sim N(\mu, \sigma^2/n) \Rightarrow \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Hence $\frac{(\bar{x} - \mu)^2}{\sigma^2/n}$, being the square of a standard normal variate is a chi-square variate with 1 d.f.

Also $\frac{ns^2}{\sigma^2}$ is a χ^2-variate with $(n-1)$ d.f. (*c.f. Theorem 13·5*).

Further since $\bar{x}$ and s^2 are independently distributed (*c.f. Theorem* 13·5), $\frac{t^2}{n-1}$, being the ratio of two independent χ^2-variates with 1 and $(n-1)$ d.f. respectively, is a $\beta_2\left(\frac{1}{2}, \frac{n-1}{2}\right)$ variate and its distribution is given by :

$$dF(t) = \frac{1}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{(t^2/\nu)^{\frac{1}{2} - 1}}{\left[1 + \frac{t^2}{\nu}\right]^{(\nu+1)/2}} d(t^2/\nu),\ 0 \le t^2 < \infty$$

[where $\nu = (n-1)$]

$$= \frac{1}{\sqrt{\nu}\, B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{\nu}\right]^{(\nu+1)/2}} dt;\ -\infty < t < \infty$$

the factor 2 disappearing since the integral from $-\infty$ to ∞ must be unity. This is the required probability function as given in (14·2) of Student's t-distribution with $\nu = (n-1)$ d.f.

Remarks on Student's 't'. 1. *Importance of Student's t-distribution in Statistics.* W.S. Gosset, who wrote under pseudonym (pen-name) of Student

defined his t in a slightly different way, *viz.*, $t = (\bar{x} - \mu)/s$ and investigated its sampling distribution, somewhat empirically, in a paper entitled '*The probable error of the mean*', published in 1908. Prof. R.A. Fisher, later on defined his own 't' and gave a rigorous proof for its sampling distribution in 1926. The salient feature of 't' is that both the statistic and its sampling distribution are functionally independent of σ, the population standard deviation.

The discovery of 't' is regarded as a landmark in the history of statistical inference because of the following reason. Before Student gave his 't' it was customary to replace σ^2 in $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$, by its unbiased estimate S^2 to give $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$ and then normal test was applied even for small samples. It has been found that although the distribution of t is asymptotically normal for large n (*c.f.* § 14·2·5), it is far from normality for small samples. The Student's t ushered in an era of exact sample distributions (and tests) and since its discovery many important contributions have been made towards the development and extension of small (exact) sample theory.

2. *Confidence or Fiducial Limits for* μ. If $t_{0.05}$ is the tabulated value of t for $\nu = (n - 1)$ d.f. at 5% level of significance, *i.e.*,

$$P(|t| > t_{0.05}) = 0.05 \Rightarrow P(|t| \leq t_{0.05}) = 0.95,$$

the 95% confidence limits for μ are given by :

$$|t| \leq t_{0.05}, \text{ i.e., } \left|\frac{\bar{x} - \mu}{S/\sqrt{n}}\right| \leq t_{0.05}$$

$$\Rightarrow \qquad \bar{x} - t_{0.05} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{0.05} \frac{S}{\sqrt{n}}$$

Thus, 95% confidence limits for μ are :

$$\bar{x} \pm t_{0.05} \cdot \frac{S}{\sqrt{n}} \qquad \ldots[14\cdot2(a)]$$

Similarly, 99% confidence limits for μ are :

$$\bar{x} \pm t_{0.01} \frac{S}{\sqrt{n}} \qquad \ldots[14\cdot2(b)]$$

where $t_{0.01}$ is the tabulated value of t for $\nu = (n - 1)$ d.f. at 1% level of significance.

14·2·2. Fisher's 't' (*Definition*). It is the ratio of a standard normal variate to the square root of an independent chi-square variate divided by its degrees of freedom. If ξ is a $N(0, 1)$ and χ^2 is an independent chi-square variate with n *d.f.*, then Fisher's t is given by

$$t = \xi \Big/ \sqrt{\frac{\chi^2}{n}} \qquad \ldots(14\cdot3)$$

and it follows student's 't' distribution with n degrees of freedom.

14·2·3. Distribution of Fisher's 't'. Since ξ and χ^2 are independent, their joint probability differential is given by

$$dF(\xi, \chi^2) = \frac{1}{\sqrt{2\pi}} \cdot \exp(-\xi^2/2) \frac{\exp(-\chi^2/2)(\chi^2)^{\frac{n}{2}-1}}{2^{n/2}\,\Gamma(n/2)} d\xi\, d\chi^2$$

Let us transform to new variates t and u by the substitution

$$t = \frac{\xi}{\sqrt{\chi^2/n}} \text{ and } u = \chi^2 \quad \Rightarrow \quad \xi = t\sqrt{u/n} \text{ and } \chi^2 = u$$

Jacobian of transformation J is given by

$$J = \frac{\partial(\xi, \chi^2)}{\partial(t, u)} = \begin{vmatrix} \sqrt{u/n} & t/(2\sqrt{un}) \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{u}{n}}$$

The joint distribution of t and u becomes

$$dG(t,u) = \frac{1}{\sqrt{2\pi}\, 2^{n/2}\, \Gamma(n/2)\, \sqrt{n}} \exp\left\{-\frac{u}{2}\left(1 + \frac{t^2}{n}\right)\right\} u^{\frac{n}{2}-\frac{1}{2}} du\, dt;$$

Integrating w.r.t. 'u' over the range 0 to ∞, the marginal distribution of t becomes

$$dG_1(t) = \frac{1}{\sqrt{2\pi}\, 2^{n/2}\, \Gamma(n/2)\, \sqrt{n}} \left[\int_0^\infty \exp\left\{-\frac{u}{2}\left(1 + \frac{t^2}{n}\right)\right\} u^{(n-1)/2}\, du\right] dt$$

$$= \frac{1}{\sqrt{2\pi}\, 2^{n/2}\, \Gamma(n/2)\, \sqrt{n}} \frac{\Gamma[(n+1)/2]}{\left[\frac{1}{2}\left(1 + \frac{t^2}{n}\right)\right]^{(n+1)/2}} dt$$

$$\therefore\ dG_1(t) = \frac{\Gamma(n+1)/2]}{\sqrt{n}\,\Gamma(n/2)\Gamma(\frac{1}{2})} \cdot \frac{1}{\left[1 + \frac{t^2}{n}\right]^{(n+1)/2}} dt, \quad -\infty < t < \infty$$

$$= \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)\left[1 + \frac{t^2}{n}\right]^{(n+1)/2}} dt, \quad -\infty < t < \infty$$

which is same as the probability function of Student's t-distribution with n d.f.

Remarks 1. In Fisher's 't' the d.f. is the same as the d.f. of chi-square variate.

2. Student's 't' may be regarded as a particular case of Fisher's 't' as explained below.

$$\text{Since } \bar{x} \sim N(\mu, \sigma^2/n), \qquad \xi = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \qquad \ldots(*)$$

and

$$\chi^2 = \frac{ns^2}{\sigma^2} = \sum_{i=1}^{n} (x_i - \bar{x})^2/\sigma^2 \qquad \ldots(**)$$

is independently distributed as chi-square variate with $(n-1)$ d.f. Hence Fisher's t is given by

$$t = \frac{\xi}{\sqrt{\chi^2/(n-1)}} = \frac{\sqrt{n}\,(\bar{x}-\mu)}{\sigma} \cdot \frac{\sigma}{\sqrt{\Sigma(x_i-\bar{x})^2/(n-1)}}$$

$$= \frac{\sqrt{n}\,(\bar{x}-\mu)}{S} = \frac{\bar{x}-\mu}{S/\sqrt{n}} \qquad \ldots(***)$$

and it follows Student's t-distribution with $(n-1)$ d.f. (*c.f.* Remark 1 above.)

Now, (***) is same as Student's 't' defined in (14·1). Hence Student's 't' is a particular case of Fisher's 't'.

14·2·4. Constants of t-distribution. Since $f(t)$ is symmetrical about the line $t = 0$, all the moments of odd order about origin vanish, *i.e.*,

$$\mu'_{2r+1} \text{ (about origin)} = 0 \; ; \; r = 0, 1, 2, \ldots$$

In particular,

$$\mu_1' \text{ (about origin)} = 0 = \text{Mean}$$

Hence central moments coincide with moments about origin.

$$\therefore \qquad \mu_{2r+1} = 0, \quad (r = 1, 2, \ldots) \qquad \ldots(14\cdot4)$$

The moments of even order are given by

$$\mu_{2r} = \mu'_{2r} \text{ (about origin)}$$

$$= \int_{-\infty}^{\infty} t^{2r} f(t)\, dt = 2\int_0^{\infty} t^{2r} f(t)\, dt$$

$$= 2.\frac{1}{B\left(\frac{1}{2}, \frac{n}{2}\right)\sqrt{n}} \int_0^{\infty} \frac{t^{2r}}{\left[1+\frac{t^2}{n}\right]^{(n+1)/2}} dt$$

This integral is absolutely convergent if $2r < n$.

Put $1 + \frac{t^2}{n} = \frac{1}{y} \;\Rightarrow\; t^2 = n(1-y)/y$ *i.e.*, $2t\,dt = -\frac{n}{y^2}dy$

When $t = 0$, $y = 1$ and when $t = \infty$, $y = 0$. Therefore,

$$\mu_{2r} = \frac{2}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_1^0 \frac{t^{2r}}{(1/y)^{(n+1)/2}} \cdot \frac{-n}{2ty^2} dy$$

$$= \frac{n}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 (t^2)^{(2r-1)/2}\, y^{\{(n+1)/2\}-2}\, dy$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 \left[n\left(\frac{1-y}{y}\right)\right]^{r-\frac{1}{2}} y^{\{(n+1)/2\}-2}\, dy$$

$$= \frac{n^r}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^1 y^{\frac{n}{2}-r-1} (1-y)^{r-\frac{1}{2}} dy$$

$$= \frac{n^r}{B\left(\frac{1}{2}, \frac{n}{2}\right)} . B\left(\frac{n}{2} - r, r + \frac{1}{2}\right), n > 2r. \quad \ldots[14\cdot4(a)]$$

$$= n^r \frac{\Gamma[(n/2) - r]\, \Gamma(r + \frac{1}{2})}{\Gamma(\frac{1}{2})\, \Gamma(n/2)}$$

$$= n^r \frac{(r - \frac{1}{2})(r - \frac{3}{2}) \ldots \frac{3}{2} \cdot \frac{1}{2} \Gamma(\frac{1}{2})\, \Gamma[(n/2) - r]}{\Gamma(\frac{1}{2})\, [(n/2) - 1][(n/2) - 2] \ldots [(n/2) - r] \Gamma[(n/2) - r]}$$

$$= n^r \frac{(2r-1)(2r-3)\ldots 3 \cdot 1}{(n-2)(n-4)\ldots(n-2r)}, \frac{n}{2} > r \quad \ldots[14\cdot4\,(b)]$$

In particular

$$\mu_2 = n \frac{1}{(n-2)} = \frac{n}{n-2}, [n > 2] \quad \ldots[14\cdot4(c)]$$

and $$\mu_4 = n^2 \frac{3 \cdot 1}{(n-2)(n-4)} = \frac{3n^2}{(n-2)(n-4)}, [n > 4] \quad \ldots[14\cdot4(d)]$$

Hence $$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0 \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = 3\left(\frac{n-2}{n-4}\right)$$

Remarks 1. As $n \to \infty$, $\beta_1 = 0$ and

$$\beta_2 = \lim_{n \to \infty} 3\left(\frac{n-2}{n-4}\right) = 3 \lim_{n \to \infty} \left[\frac{1 - (2/n)}{1 - (4/n)}\right] = 3 \quad \ldots[14\cdot4(e)]$$

2. Changing r to $(r-1)$ in [14·4(*b*)], dividing and simplifying, we shall get the recurrence relation for the moments as

$$\frac{\mu_{2r}}{\mu_{2r-2}} = \frac{n(2r-1)}{(n-2r)}, \frac{n}{2} > r \quad \ldots[14\cdot4(c)]$$

3. Moment Generating Function of t-distribution. From [14·4(*b*)] we observe that if $t \sim t_n$, then all the moments of order $2r < n$ exist but the moments of order $2r \geq n$ do not exist. Hence the m.g.f. of t-distribution does not exist.

Example 14·1. *Express the constants y_0, a and m of the distribution :*

$$dF(x) = y_0 \left[1 - \frac{x^2}{a^2}\right]^m dx, \; -a \leq x \leq a \quad \ldots(*)$$

in terms of its μ_2 and β_2.

Show that if x is related to a variable t by the equation

$$x = \frac{at}{\{2(m+1) + t^2\}^{1/2}}, \quad \ldots(**)$$

then t has Student's distribution with 2(m + 1) degrees of freedom. Use the transformation to calculate the probability that t ≥ 2 when the degrees of freedom are 2 and also when 4. **(*Madras Univ. M.Sc., 1991*)**

Solution. First of all we shall determine the constant from the consideration that total probability is unity.

$$\therefore \qquad y_0 \int_{-a}^{a} \left(1 - \frac{x^2}{a^2}\right)^m dx = 1$$

$$\Rightarrow \qquad 2y_0 \int_{0}^{a} \left(1 - \frac{x^2}{a^2}\right)^n dx = 1$$

(∵ Integrand is an even function of x)

$$\Rightarrow \qquad 2y_0 \int_0^{\pi/2} \cos^{2m}\theta \,.\, a\cos\theta\, d\theta = 1 \qquad (x = a\sin\theta)$$

$$\Rightarrow \qquad 2ay_0 \int_0^{\pi/2} \cos^{2m+1}\theta\, d\theta = 1$$

But we have the Beta integral,

$$2\int_0^{\pi/2} \sin^p\theta \cos^q\theta\, d\theta = B\left(\frac{p+1}{2}, \frac{q+1}{2}\right) \qquad ...(1)$$

$$\therefore \qquad ay_0 . 2\int_0^{\pi/2} \cos^{2m+1}\theta \sin^0\theta\, d\theta = 1$$

$$\Rightarrow \qquad ay_0 B(m+1, \tfrac{1}{2}) = 1 \qquad \text{[Using (1)]}$$

$$\Rightarrow \qquad y_0 = \frac{1}{a\, B(m+1, \frac{1}{2})} \qquad ...(2)$$

Since the given probability function is symmetrical about the line $x = 0$, we have as in § 14·2·4.

$$\mu_{2r+1} = \mu'_{2r+1} = 0;\ r = 0, 1, 2, \ldots \qquad [\because \text{Mean} = \text{Origin}]$$

The moments of even order are given by

$$\mu_{2r} = \mu'_{2r} \text{ (about origin)}$$

$$= \int_{-a}^{a} x^{2r} f(x)\, dx = y_0 \int_{-a}^{a} x^{2r} \left(1 - \frac{x^2}{a^2}\right)^m dx$$

$$= 2y_0 \int_0^a x^{2r} \left(1 - \frac{x^2}{a^2}\right)^m dx$$

$$= 2y_0 \int_0^{\pi/2} (a \sin\theta)^{2r} \cos^{2m}\theta .\ a\cos\theta\, d\theta \qquad [x = a\sin\theta]$$

$$= y_0\, a^{2r+1} . 2 \int_0^{\pi/2} \sin^{2r}\theta . \cos^{2m+1}\theta\, d\theta$$

$$= y_0\, a^{2r+1} B(r + \tfrac{1}{2}, m + 1) \qquad [\text{Using (1)}]]$$

$$= a^{2r} \frac{B(r+\frac{1}{2},\ m+1)}{B(m+1,\ \frac{1}{2})} = a^{2r} . \frac{\Gamma\left(r+\frac{1}{2}\right)\Gamma\left(m+\frac{3}{2}\right)}{\Gamma\left(m+r+\frac{3}{2}\right)\Gamma\left(\frac{1}{2}\right)} \quad ...(***)$$

In particular, $\mu_2 = a^2 . \dfrac{\Gamma\{m + (3/2)\} . \frac{1}{2}\Gamma(1/2)}{\{m + (3/2)\}\ \Gamma\{m + (3/2)\}\ \Gamma(1/2)} = \dfrac{a^2}{2m + 3}$

$\therefore \quad a^2 = (2m + 3)\mu_2 \qquad ...(3)$

Also $\quad \mu_4 = a^4 \dfrac{\Gamma(5/2)}{\Gamma\{m + (7/2)\}} \times \dfrac{\Gamma\{m + (3/2)\}}{\Gamma(1/2)}$

$$= \frac{3a^4}{(2m + 5)\ (2m + 3)} \qquad (\textit{On simplification})$$

$\therefore \quad \beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{3(2m + 3)}{(2m + 5)}$

$\Rightarrow \quad m = \dfrac{9 - 5\beta_2}{2(\beta_2 - 3)} \qquad (\textit{On simplification}) \ ... \ (4)$

Equations (2), (3) and (4) express the constants y_0, a and m in terms of μ_2 and β_2.

$$x = \frac{at}{[2(m + 1) + t^2]^{1/2}} \quad \Rightarrow \quad \frac{x^2}{a^2} = \frac{t^2}{2(m + 1) + t^2}$$

i.e., $\quad 1 - \dfrac{x^2}{a^2} = \dfrac{2(m + 1)}{2(m + 1) + t^2} = \left(1 + \dfrac{t^2}{n}\right)^{-1}, (n = 2m + 2)$

Also $\quad dx = a\left[\dfrac{dt}{(n + t^2)^{1/2}} - t \cdot \dfrac{1}{2} \dfrac{2t\, dt}{(n + t^2)^{3/2}}\right]$

$$= a \frac{1}{(n + t^2)^{1/2}} \left[1 - \frac{t^2}{n + t^2}\right] dt$$

$$= \frac{an}{(n + t^2)^{3/2}} dt = \frac{a}{\sqrt{n}} \cdot \frac{1}{[1 + (t^2/n)]^{3/2}} dt$$

Hence the p.d.f. of X transforms to

$$dF(t) = y_0 \frac{1}{\left[1 + \dfrac{t^2}{n}\right]^m} \cdot \frac{a}{\sqrt{n}} \frac{dt}{\left[1 + \dfrac{t^2}{n}\right]^{3/2}}$$

$$= \frac{1}{a\,B\left(m+1, \frac{1}{2}\right)} \cdot \frac{a}{\sqrt{n}} \cdot \frac{dt}{\left[1+\frac{t^2}{n}\right]^{m+(3/2)}}$$

$$= \frac{1}{\sqrt{n}\,B\left(\frac{n}{2}, \frac{1}{2}\right)} \cdot \frac{dt}{\left[1+\frac{t^2}{n}\right]^{(n+1)/2}}, \quad -\infty < t < \infty \quad \ldots(5)$$

which is the probability differential of Student's t-distribution with $n = 2(m+1)$ d.f. Hence the result.

For 2 d.f. *i.e.*, $n = 2$, we get $2(m+1) = 2 \Rightarrow m = 0$. Hence from (**), we get (for $m = 0$),

$$x = \frac{at}{(2+t^2)^{1/2}} \quad \Rightarrow \quad x = \frac{\sqrt{2}}{\sqrt{3}}a, \text{ when } t = 2.$$

$$\therefore \qquad P(t \geq 2) = P\left(X \geq \sqrt{(2/3)}\,a\right) = \int_{a\sqrt{(2/3)}}^{a} dF(x)$$

$$= \int_{a\sqrt{(2/3)}}^{a} \frac{1}{a\,B(1, \frac{1}{2})}\,dx \qquad \text{[From (*), since } m = 0\text{]}$$

$$= \frac{1}{2a}\left(a - \frac{\sqrt{2}}{\sqrt{3}}a\right) = \frac{\sqrt{3}-\sqrt{2}}{2\sqrt{3}}$$

$$\left[\because B(1, \tfrac{1}{2}) = \frac{\Gamma 1\,\Gamma(1/2)}{\Gamma(3/2)} = \frac{\Gamma(1/2)}{(1/2)\,\Gamma(1/2)} = 2\right]$$

For 4 d.f., *i.e.*, $n = 4$, we get $m = 1$. Proceeding exactly similarly we shall obtain

$$P(t \geq 2) = \frac{1}{2} - \frac{5\sqrt{2}}{16}$$

EXERCISE 14(a)

1. (*a*) Given that

(*i*) u is normally distributed with zero mean and unit variance,

(*ii*) v^2 has a chi-square distribution with n degrees of freedom, and

(*iii*) u and v are independently distributed,

find the distribution of the variable

$$t = \frac{u\sqrt{n}}{v}$$

(*b*) Find the variance of the t distribution with n degrees of freedom, $(n > 2)$.

(*c*) If the variable t has Student's t distribution with 2 degrees of freedom, prove that

$$P(t \geq 2) = \frac{3 - \sqrt{6}}{6}$$

[*Shivaji Univ. B.Sc., 1990*]

2. (*a*) State, (without proof), the sampling distribution of Student's t. Who discovered it?

(*b*) 'Discovery of Student's t is regarded as a landmark in the history of statistical inference'. Elucidate.

(*c*) Let t be distributed as Student's t-distribution with 2 d.f. Find the probability $P(-\sqrt{2} \leq t \leq \sqrt{2})$.

3. (*a*) Show that

$$E(T^r) = \begin{cases} \dfrac{k^{r/2}\,\Gamma\left(\dfrac{1+r}{2}\right) \cdot \Gamma\left(\dfrac{k-r}{2}\right)}{\Gamma(1/2) \cdot \Gamma(k/2)}, & \text{if } r \text{ is even for } -1 < r < k \\ 0, & \text{if } r \text{ is odd} \end{cases}$$

where T has Student's t-distribution with k degrees of freedom.

(*b*) For the t-distribution with n d.f., establish the recurrence relation

$$\mu_{2r} = \frac{n(2r-1)}{(n-2r)} \cdot \mu_{2r-2}, \; n > 2r$$

[*Poona Univ. B.Sc., 1990; Delhi Univ. B.Sc. (Stat. Hons.), 1992*]

(*c*) For how many d.f. does (*i*) χ^2-distribution reduce to negative exponential distribution and (*ii*) t-distribution reduce to Cauchy distribution?

4. Suppose $X_1, X_2, \ldots, X_n$ $(n > 1)$ are independent variates each distributed as $N(0, \sigma^2)$. Find the p.d.f. of

$$W = X_1 \Big/ \left\{\frac{1}{n}\sum_{i=1}^{n} X_i^2\right\}^{1/2}$$

Why does not W follow the t-distribution?

[*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

5. Let $x_1, x_2, \ldots, x_n$ be independent observations from a normal universe with mean μ and variance σ^2 and let $\bar{x}$ and s^2 be the sample mean and sum of the squares of the deviations from the mean respectively. Let x' be one more observation independent of previous ones. Show that

$$\frac{x' - \bar{x}}{s}\left[\frac{n(n-1)}{n+1}\right]^{1/2}$$

has a Student t-distribution with $(n-1)$ degrees of freedom.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1989*]

6. (*a*) Let X_1 and X_2 be two independent normal variates with the same normal distribution $N(\mu, \sigma^2)$. Obtain the distribution of

$$Y = \frac{X_1 + X_2 - 2\mu}{\sqrt{|X_1 - X_2|^2}}$$

Ans. Standard Cauchy distribution.

(*b*) If X is t-distributed with k degrees of freedom, show that

$$\frac{1}{1 + (X^2/k)},$$

has a beta distribution.

[***Delhi Univ. B.Sc. (Maths. Hons.), 1988***]

7. Define Student's t-statistic and state its probability density function.

If x_i $(i = 1, 2, \ldots, n)$, is a random sample of n independent observations from a normal population with mean μ and variance σ^2, show that

$$U = \frac{(\bar{x} - \mu)\sqrt{n(n-1)}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \text{ where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

conforms to Student's t-variate. If x is an additional observation drawn independently from the same normal population, show that

$$W = \frac{(x - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \times \sqrt{\frac{n(n-1)}{n+1}}$$

also conforms to Student's t-variate.

8. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, and $\bar{X}$ and S^2, respectively, be the sample mean and sample variance. Let $X_{n+1} \sim N(\mu, \sigma^2)$, and assume that $X_1, X_2, \ldots, X_n, X_{n+1}$ are independent. Obtain the sampling distribution of

$$U = \frac{(X_{n+1} - \bar{X})}{S} \cdot \sqrt{\frac{n}{n+1}}; \quad \left[S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

9. If the random variables X_1 and X_2 are independent and follow chi-square distribution with n d.f., show that $\dfrac{\sqrt{n}\,(X_1 - X_2)}{2\sqrt{X_1 X_2}}$ is distributed as Student's t with n d.f., independently of $X_1 + X_2$.

[***Calcutta Univ. B.Sc. (Hons.), 1992***]

Hint. $p(x_1, x_2) = \dfrac{1}{2^n [\Gamma(n/2)]^2} \cdot e^{-(x_1 + x_2)/2}\, x_1^{(n/2)-1}\, x_2^{(n/2)-1}$;

$$0 \le x_1 < \infty,\ 0 \le x_2 < \infty$$

Put $\quad u = \dfrac{\sqrt{n}\,(x_1 - x_2)}{2\sqrt{x_1 x_2}}$ and $v = x_1 + x_2$

$$\Rightarrow \quad x_1 = \frac{v}{2}\left[1 + \frac{1}{\sqrt{\left(1 + \frac{n}{u^2}\right)}}\right], \quad x_2 = \frac{v}{2}\left[1 - \frac{1}{\sqrt{\left(1 + \frac{n}{u^2}\right)}}\right]$$

Jacobian of transformation is $J = \dfrac{\partial(x_1, x_2)}{\partial(u, v)} = \dfrac{v}{2\sqrt{n}\,[1 + u^2/n]^{3/2}}$

The joint p.d.f. of U and V becomes

$$g(u, v) = p(x_1, x_2) \mid J \mid = \frac{1}{2^{2n-1}\,\Gamma(n/2)\,\Gamma(n/2)\,\sqrt{n}} \cdot \frac{e^{-v/2}\, v^{n-1}}{(1 + u^2/n)^{(n+1)/2}};$$

$$-\infty < u < \infty,\ 0 \le v < \infty$$

Using Legender's duplication formula, *viz.*,

$$\Gamma n = 2^{n-1}\,\Gamma(n/2)\,\Gamma\left(\frac{n+1}{2}\right)\sqrt{\pi} \quad \Rightarrow \quad \Gamma(n/2) = \frac{\Gamma n\,\sqrt{\pi}}{2^{n-1}\,\Gamma\left(\frac{n+1}{2}\right)}, \text{ we get}$$

$$2^{2n-1}\,\Gamma(n/2)\,\Gamma(n/2)\,\sqrt{n} = \frac{2^{2n-1}\cdot\sqrt{n}\,\sqrt{\pi}}{2^{n-1}\,\Gamma\left(\frac{n+1}{2}\right)}\,\Gamma\left(\frac{n}{2}\right)\sqrt{n}$$

$$= 2^n\,\sqrt{n}\,\sqrt{n}\,B\left(\tfrac{1}{2}, n/2\right) \qquad \left[\because \sqrt{\pi} = \Gamma\left(\tfrac{1}{2}\right)\right]$$

$$g(u, v) = \left(\frac{1}{2^n\,\Gamma n}\, e^{-v/2}\, v^{n-1}\right)\left[\frac{1}{\sqrt{n}\,B\left(\frac{1}{2}, n/2\right)} \cdot \frac{1}{\left(1 + \frac{u^2}{n}\right)^{(n+1)/2}}\right];$$

$$0 < v < \infty,\ -\infty < u < \infty.$$

10. Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent random samples from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively. If $\bar{X}$ and $\bar{Y}$ denote the corresponding sample means and if

$$(m-1)S_1^2 = \sum_{i=1}^{m} (X_i - \bar{X})^2,\ (n-1)\,S_2^2 = \sum_{j=1}^{n} (Y_j - \bar{Y})^2,$$

obtain the sampling distribution of

$$\frac{a(\bar{X} - \mu_1) + b(\bar{Y} - \mu_2)}{\left[\left\{\dfrac{(m-1)\,S_1^2 + (n-1)\,S_2^2}{(m+n-2)}\right\}\left\{\dfrac{a^2}{m} + \dfrac{b^2}{n}\right\}\right]^{1/2}}$$

where a and b are two fixed real numbers.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1989*]

11. If $I_x(p, q)$ represents the incomplete Beta function defined by

$$I_x(p, q) = \frac{1}{B(p, q)} \int_0^x t^{p-1}\,(1-t)^{q-1}\,dt;\ p > 0, q > 0,$$

show that the distribution function $F(.)$ of Student's t-distribution is given by

$$F(t) = 1 - \frac{1}{2} I_x\left(\frac{n}{2}, \frac{1}{2}\right), \text{ where } x = \left(1 + \frac{t^2}{n}\right)^{-1}.$$

[*Delhi Univ. M.Sc. (Stat.), 1990; Nagpur Univ. M.Sc. (Stat.), 1991*]

Hint. If $f(.)$ is p.d.f. of t-distribution with n d.f, then

$$F(t) = \int_{-\infty}^{t} f(u)\,du = 1 - \int_{t}^{\infty} f(u)\,du$$

$$= 1 - \frac{1}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{t}^{\infty} \left(1 + \frac{u^2}{n}\right)^{-(n+1)/2} du$$

$$= 1 + \frac{1}{2\,B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{\left(1+\frac{t^2}{n}\right)^{-1}}^{0} z^{(n/2)-1}(1-z)^{-1/2}\,dz,$$

where $\left[\frac{1}{z} = 1 + \frac{u^2}{n}\right]$

$$= 1 - \frac{1}{2\,B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{0}^{x} z^{(n/2)-1}\,(1-z)^{-1/2}\,dz, \quad \left[x = \left(1 + \frac{t^2}{n}\right)^{-1}\right]$$

$$= 1 - \frac{1}{2} I_x\left(\frac{n}{2}, \frac{1}{2}\right)$$

12. Show that for t-distribution with n d.f., mean deviation about mean is given by

$$\sqrt{n}\ \Gamma\left(\frac{n-1}{2}\right) \Big/ \sqrt{\pi}\ \Gamma(n/2)$$

(Shivaji Univ. B.Sc. Oct., 1992)

Hint. $E(t) = 0$.

M.D. about mean $= \int_{-\infty}^{\infty} |t|\,f(t)\,dt$

$$= \frac{1}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{-\infty}^{\infty} \frac{|t|\,dt}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}}$$

$$= \frac{2}{\sqrt{n}\,B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{0}^{\infty} \frac{t\,dt}{\left(1 + \frac{t^2}{n}\right)^{(n+1)/2}}$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_{0}^{\infty} \frac{dy}{(1+y)^{(n+1)/2}}, \qquad \left[\left(\frac{t^2}{n} = y\right)\right]$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \int_0^\infty \frac{y^{1-1}}{(1+y)^{\frac{n-1}{2}+1}} dy$$

$$= \frac{\sqrt{n}}{B\left(\frac{1}{2}, \frac{n}{2}\right)} B\left(\frac{n-1}{2}, 1\right)$$

13. If $X \sim t_{(n)}$, show that

$$(n - \tfrac{1}{2}) \log\left[1 + \frac{x^2}{n}\right] \sim \chi^2_{(1)},$$

for large n.

You may assume that for large n,

$$\frac{\Gamma\left(\frac{n}{2} + \frac{1}{2}\right)}{\Gamma\left(\frac{1}{2} + n\right)\sqrt{\frac{1}{2}n}} \simeq \left(1 - \frac{1}{4n}\right)$$

14. If $\bar{X}$ and $\hat{\sigma}^2 = S^2$ be the usual sample mean and sample variance based on a random sample of n observations from $N(\mu, \sigma^2)$, and if $T = (\bar{X} - \mu)\sqrt{n}/S$, prove that

(*i*) $\text{Var}(T) = (n-1)/(n-3)$

(*ii*) $\text{Cov}(\bar{X}, T) = \sigma \dfrac{\sqrt{n-1}\ \Gamma[(n-2)/2]}{\sqrt{2n}\ \Gamma[(n-1)/2]}$

(*iii*) $r(\bar{X}, T) = \left[\tfrac{1}{2}(n-3)\right]^{1/2} \Gamma\left[\tfrac{1}{2}(n-2)\right] / \Gamma\left[\tfrac{1}{2}(n-1)\right]$

14·2·5. Limiting Form of t-distribution. *As $n \to \infty$, the p.d.f. of t-distribution with n d.f. viz.,*

$$f(t) = \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \to \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \quad -\infty < t < \infty$$

Proof. $\lim_{n \to \infty} \dfrac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} = \lim_{n \to \infty} \dfrac{1}{\sqrt{n}} \dfrac{\Gamma[(n+1)/2]}{\Gamma(\frac{1}{2})\,\Gamma(n/2)}$

$$= \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\pi}} \left(\frac{n}{2}\right)^{\frac{1}{2}} = \frac{1}{\sqrt{2\pi}}$$

$\left[\because \Gamma(\tfrac{1}{2}) = \sqrt{\pi} \text{ and } \lim_{n \to \infty} \dfrac{\Gamma(n+k)}{\Gamma(n)} = n^k, \text{ (c.f. Remark to § 14·5·7)}\right]$

$$\therefore \lim_{n \to \infty} f(t) = \lim_{n \to \infty} \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \lim_{n \to \infty} \left[\left(1 + \frac{t^2}{n}\right)^n\right]^{-\frac{1}{2}}$$

$$\times \lim_{n \to \infty} \left(1 + \frac{t^2}{n}\right)^{-\frac{1}{2}}$$

$$= \frac{1}{\sqrt{2\pi}} \exp(-t^2/2), \quad -\infty < t < \infty$$

Hence for large d.f. t-distribution tends to standard normal distribution.

14·2·6. Graph of t-distribution. The p.d.f. of t-distribution with n d.f. is

$$f(t) = C.\left[1 + \frac{t^2}{n}\right]^{-(n+1)/2}, \quad -\infty < t < \infty$$

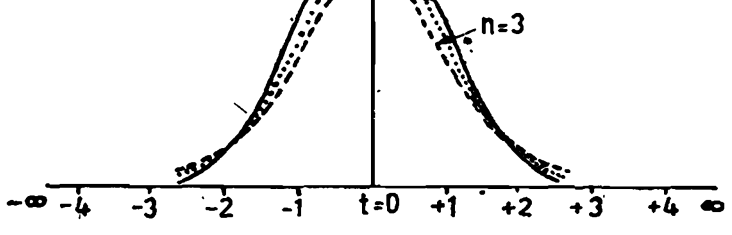
Since $f(-t) = f(t)$, the probability curve is symmetrical about the line $t = 0$. As t increases, $f(t)$ decreases rapidly and tends to zero as $t \to \infty$, so that t-axis is an asymptote to the curve. We have shown that

$$\mu_2 = \frac{n}{n-2}, n > 2; \quad \beta_2 = \frac{3(n-2)}{(n-4)}, n > 4$$

Hence for $n > 2$, $\mu_2 > 1$ *i.e.*, the variance of t-distribution is greater than that of standard normal distribution and for $n > 4$, $\beta_2 > 3$ and thus t-distribution is more flat on the top than the normal curve. In fact, for small n, we have

$$P[|t| \geq t_0] \geq P[|Z| \geq t_0], \quad Z \sim N(0, 1)$$

i.e., the tails of the t-distribution have a greater probability (area) than the tails of standard normal distribution. Moreover we have also seen [§ 14·2·5] that for large n (*d.f.*), t-distribution tends to standard normal distribution.

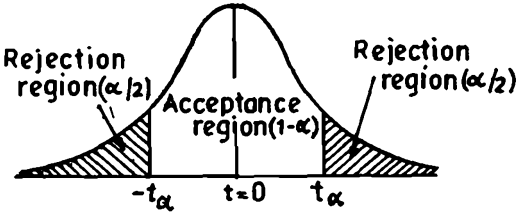


14·2·7. Critical Values of t. The critical (or significant) values of t at level of significance α and *d.f.* υ for *two-tailed test* are given by the equation

$$P[|t| > t_\upsilon(\alpha)] = \alpha \qquad \ldots(14\cdot5)$$

$$\Rightarrow \qquad P[|t| \leq t_\upsilon(\alpha)] = 1 - \alpha. \qquad \ldots(14\cdot5a)$$

CRITICAL VALUES OF t-DISTRIBUTION



The values $t_\nu(\alpha)$ have been tabulated in Fisher and Yates' Tables, for different values of α and ν and are given in the Appendix at the end of the book.

Since t-distribution is symmetric about $t = 0$, we get from (14·5)

$$P(t > t_\nu(\alpha)] + P[t < -t_\nu(\alpha)] = \alpha$$

$$\Rightarrow \quad 2P[t > t_\nu(\alpha)] = \alpha$$

$$\Rightarrow \quad P[t > t_\nu(\alpha)] = \alpha/2$$

$$\Rightarrow \quad P[t > t_\nu(2\alpha)] = \alpha \qquad \ldots(14\cdot5b)$$

$t_\nu(2\alpha)$ (from the Tables in the Appendix) gives the significant value of t for a single-tail test, [Right-tail or Left-tail-since the distribution is symmetrical], at level of significance α and ν *d.f.*

Hence the significant values of t at level of significance 'α' for a single tailed test can be obtained from those of two-tailed test by looking the values at level of significance '2α'.

For example,

t_8 (0·05) for single-tail test = t_8 (0·10) for two-tail test = 1·86

t_{15}(0·01) for single-tail test = t_{15} (0·02) for two-tail test = 2·60.

14·2·8. Applications of t-distribution. The t-distribution has a wide number of applications in Statistics, some of which are enumerated below.

(*i*) To test if the sample mean ($\bar{x}$) differs significantly from the hypothetical value μ of the population mean.

(*ii*) To test the significance of the difference between two sample means.

(*iii*) To test the significance of an observed sample correlation co-efficient and sample regression coefficient.

(*iv*) To test the significance of observed partial and multiple correlation coefficients.

In the following sections we will discuss these applications in detail, one by one.

14·2·9. t-Test for Single Mean. Suppose we want to test :

(*i*) if a random sample x_i ($i = 1, 2, \ldots, n$) of size n has been drawn from a normal population with a specified mean, say μ_0, or

(*ii*) if the sample mean differs significantly from the hypothetical value μ_0 of the population mean.

Under the null hypothesis H_0 :

(i) The sample has been drawn from the population with mean μ or (ii) there is no significant difference between the sample mean $\bar{x}$ and the population mean μ,

the statistic
$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \qquad \ldots(14\cdot6)$$

where
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \ldots[14\cdot6(a)]$$

follows Student's *t*-distribution with $(n-1)$ d.f.

We now compare the calculated value of t with the tabulated value at certain level of significance. If calculated $|t|$ > tabulated t, null hypothesis is rejected and if calculated $|t|$ < tabulated t, H_0 may be accepted at the level of significance adopted.

Remarks 1. *On computation of S^2 for numerical problems.* If $\bar{x}$ comes out in integers, the formula (14·6*a*) can be conveniently used for computing S^2. However, if $\bar{x}$ comes in fractions then the formula (14·6*a*) for computing S^2 is very cumbersome and is not recommended. In that case, step deviation method, given below, is quite useful.

If we take, $d_i = x_i - A$, where A is any arbitrary number then

$$S^2 = \frac{1}{n-1}\left[\sum(x_i - \bar{x})^2\right] = \frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right] \qquad \ldots[14\cdot6(b)]$$

$$= \frac{1}{n-1}\left[\sum d_i^2 - \frac{(\sum d_i)^2}{n}\right], \qquad \ldots[14\cdot6(c)]$$

since variance is independent of change of origin.

Also, in this case $\bar{x} = A + \frac{\sum d_i}{n}$. $\qquad \ldots[14\cdot6(d)]$

2. We know, the sample variance

$$s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

$$ns^2 = (n-1)S^2$$

$$\Rightarrow \quad \frac{S^2}{n} = \frac{s^2}{n-1} \qquad \ldots[14\cdot6(e)]$$

Hence for numerical problems, the test statistic (14·6) on using [14·6(*e*)] becomes

$$t = \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}} = \frac{\bar{x} - \mu_0}{\sqrt{s^2/(n-1)}} \sim t_{n-1} \qquad \ldots[14\cdot6(f)]$$

3. Assumptions for Student's t-test. The following assumptions are made in the Student's *t*-test :

(*i*) The parent population from which the sample is drawn is normal.

(*ii*) The sample observations are independent, *i.e.*, the sample is random.

(*iii*) The population standard deviation σ is unknown.

Example 14·2. *A machinist is making engine parts with axle diameters of 0·700 inch. A random sample of 10 parts shows a mean diameter of 0·742 inch with a standard deviation of 0·040 inch. Compute the statistic you would use to test whether the work is meeting the specifications. Also state how you would proceed further.*

Solution. Here we are given :

$\mu = 0{\cdot}700$ inches, $\bar{x} = 0{\cdot}742$ inches, $s = 0{\cdot}040$ inches and $n = 10$

Null Hypothesis, $H_0 : \mu = 0{\cdot}700$, *i.e., the product is conforming to specifications.*

Alternative Hypothesis, $H_1 : \mu \neq 0{\cdot}700$

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{\bar{x} - \mu}{\sqrt{S^2/n}} = \frac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{(n-1)}$$

Now
$$t = \frac{\sqrt{9}(0{\cdot}742 - 0{\cdot}700)}{0{\cdot}040} = 3{\cdot}15$$

How to proceed further. Here the test statistic 't' follows Student's t-distribution with $10 - 1 = 9$ d.f. We will now compare this calculated value with the tabulated value of t for 9 d.f. and at certain level of significance, say 5%. Let this tabulated value be denoted by t_0.

(*i*) If calculated 't' *viz.*, $3{\cdot}15 > t_0$, we say that the value of t is significant. This implies that $\bar{x}$ differs significantly from μ and H_0 is rejected at this level of significance and we conclude that the product is not meeting the specifications.

(*ii*) If calculated $t < t_0$, we say that the value of t is not significant, *i.e.*, there is no significant difference between $\bar{x}$ and μ. In other words, the deviation $(\bar{x} - \mu)$ is just due to fluctuations of sampling and null hypothesis H_0 may be retained at 5% level of significance, *i.e.*, we may take the product conforming to specifications.

Example 14·3. *The mean weekly sales of soap bars in departmental stores was 146·3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153·7 and showed a standard deviation of 17·2. Was the advertising campaign successful ?*

Solution. We are given : $n = 22$, $\bar{x} = 153{\cdot}7$, $s = 17{\cdot}2$.

Null Hypothesis. The advertising campaign is not successful, *i.e.*,

$H_0 : \mu = 146{\cdot}3$

Alternative Hypothesis. $H_1 : \mu > 146{\cdot}3$ (Right-tail).

Test Statistic. Under the null hypothesis, the test statistic is :

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{22-1} = t_{21}$$

Now
$$t = \frac{153{\cdot}7 - 146{\cdot}3}{\sqrt{(17{\cdot}2)^2/21}} = \frac{7{\cdot}4 \times \sqrt{21}}{17{\cdot}2} = 9{\cdot}03$$

Conclusion. Tabulated value of t for 21 *d.f.* at 5% level of significance for *single-tailed test* is 1·72. Since calculated value is much greater than the

tabulated value, it is highly significant. Hence we reject the null hypothesis and conclude that the advertising campaign was definitely successful in promoting sales.

Example 14·4. *A random sample of 10 boys had the following I.Q.'s : 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q. of 100 ? Find a reasonable range in which most of the mean I.Q. values of samples of 10 boys lie.*

[*Madras Univ. B.E., April 1990*]

Solution. *Null hypothesis, H_0 : The data are consistent with the assumption of a mean I.Q. of 100 in the population, i.e., $\mu = 100$.*

Alternative hypothesis, $H_1 : \mu \neq 100$.

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{(\bar{x} - \mu)}{\sqrt{S^2/n}} \sim t_{(n-1)},$$

where $\bar{x}$ and S^2 are to be computed from the sample values of I.Q.'s.

CALCULATIONS FOR SAMPLE MEAN AND S.D.

X	$(X - \bar{x})$	$(X - \bar{x})^2$
70	−27·2	739·84
120	22·8	519·84
110	12·8	163·84
101	3·8	14·44
88	−9·2	84·64
83	−14·2	201·64
95	−2·2	4·84
98	0·8	0·64
107	9·8	96·04
100	2·8	7·84
Total 972		1833·60

Hence $n = 10, \bar{x} = \frac{972}{10} = 97{\cdot}2$ and $S^2 = \frac{1833{\cdot}60}{9} = 203{\cdot}73$

$$\therefore \quad |t| = \frac{|97{\cdot}2 - 100|}{\sqrt{203{\cdot}73/10}} = \frac{2{\cdot}8}{\sqrt{20{\cdot}37}} = \frac{2{\cdot}8}{4{\cdot}514} = 0{\cdot}62$$

Tabulated $t_{0{\cdot}05}$ for $(10 - 1)$ *i.e.*, 9 d.f. for two-tailed test is 2·262.

Conclusion. Since calculated t is less than tabulated $t_{0{\cdot}05}$ for 9 d.f., H_0 may be accepted at 5% level of significance and we may conclude that the data are consistent with the assumption of mean I.Q. of 100 in the population.

The 95% confidence limits within which the mean I.Q. values of samples of 10 boys will lie are given by

$$\bar{x} \pm t_{0{\cdot}05}\, S/\sqrt{n} = 97{\cdot}2 \pm 2{\cdot}262 \times 4{\cdot}514$$

$= 97{\cdot}2 \pm 10{\cdot}21 = 107{\cdot}41$ and $86{\cdot}99$

Hence the required 95% confidence interval is [86·99, 107·41].

Remark. Aliter for computing $\bar{x}$ and S^2. Here we see that $\bar{x}$ comes in fractions and as such the computation of $(x - \bar{x})^2$ is quite laborious and time consuming. In this case we use the method of step deviations to compute $\bar{x}$ and S^2, as given below.

X	$d = X - 90$	d^2
70	−20	400
120	30	900
110	20	400
101	11	121
88	−2	4
83	−7	49
95	5	25
98	8	64
107	17	289
100	10	100
Total	$\Sigma d = 72$	$\Sigma d^2 = 2352$

Here $d = X - A$, where $A = 90$

$$\therefore \quad \bar{x} = A + \frac{1}{n}\Sigma d = 90 + \frac{72}{10} = 97{\cdot}2$$

and $$S^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right] = \frac{1}{9}\left[2352 - \frac{(72)^2}{10}\right] = 203{\cdot}73$$

Example 14·5. *The heights of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches ? Test at 5% significance level, assuming that for 9 degrees of freedom $P(t > 1{\cdot}83) = 0{\cdot}05$.*

Solution. *Null Hypothesis,* $H_0 : \mu = 64$ inches.

Alternative Hypothesis, $H_1 : \mu > 64$ inches.

CALCULATIONS FOR SAMPLE MEAN AND S.D.

x	70	67	62	68	61	68	70	64	64	66	Total 660
$x - \bar{x}$	4	1	−4	2	−5	2	4	−2	−2	0	0
$(x - \bar{x})^2$	16	1	16	4	25	4	16	4	4	0	90

$$\bar{x} = \frac{\Sigma x}{n} = \frac{660}{10} = 66$$

$$S^2 = \frac{1}{n-1}\Sigma (x-\bar{x})^2 = \frac{90}{9} = 10$$

Test Statistic. Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{\sqrt{S^2/n}} = \frac{66 - 64}{\sqrt{10/10}} = 2,$$

which follows Student's t-distribution with $10 - 1 = 9$ *d.f.*

Tabulated value of t for 9 *d.f.* at 5% level of significance for single (right) tail-test is 1·833. (This is the value $t_{0\cdot10}$ for 9 *d.f.* in the two-tailed Table given in the Appendix.)

Conclusion. Since calculated value of t is greater than the tabulated value, it is significant. Hence H_0 is rejected at 5% level of significance and we conclude that the average height is greater than 60 inches.

Example 14·6. *A random sample of 16 values from a normal population showed a mean of 41·5 inches and the sum of squares of deviations from this mean equal to 135 square inches. Show that the assumption of a mean of 43·5 inches for the population is not reasonable. Obtain 95 per cent and 99 per cent fiducia. limits for the same.*

You may use the following information from statistical tables :

$$v = 15, \begin{cases} P = 0{\cdot}05, \; t = 2{\cdot}131 \\ P = 0{\cdot}01, \; t = 2{\cdot}947 \end{cases}$$

Solution. We are given $n = 16$, $\bar{x} = 41{\cdot}5$ inches and

$$\Sigma(x - \bar{x})^2 = 135 \text{ sq. inches.}$$

$$\therefore \qquad S^2 = \frac{1}{n-1}\Sigma(x-\bar{x})^2 = \frac{135}{15} = 9 \quad \Rightarrow \quad S = 3.$$

Null Hypothesis, H_0 : $\mu = 43{\cdot}5$ inches, *i.e.*, the data are consistent with the assumption that the mean height in the population is 43·5 inches.

Alternative Hypothesis, H_1 : $\mu \neq 43{\cdot}5$ inches.

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

Now $$|t| = \frac{|41{\cdot}5 - 43{\cdot}5|}{3/4} = \frac{8}{3} = 2{\cdot}667$$

Here number of degrees of freedom is $(16 - 1) = 15$.

We are given :

$t_{0\cdot05}$ for 15 d.f. = 2·131 and $t_{0\cdot01}$ for 15 d.f. = 2·947

Conclusion. Since calculated $|t|$ is greater than 2·131, null hypothesis is rejected at 5% level of significance and we conclude that the assumption of mean of 43·5 inches for the population is not reasonable.

Remark. Since calculated $|t|$ is less than 2·947, null hypothesis ($\mu = 43{\cdot}5$) may be accepted at 1% level of significance.

95% fiducial limits for μ : (d.f. = 15)

$$\bar{x} \pm t_{0\cdot05} \times \frac{S}{\sqrt{n}} = 41\cdot5 \pm 2\cdot131 \times \frac{3}{4} = 41\cdot5 \pm 1\cdot598$$

$$\therefore \quad 39\cdot902 < \mu < 43\cdot098$$

99% fiducial limits for μ : (*d.f.* = 15)

$$\bar{x} \pm t_{0\cdot01} \times \frac{S}{\sqrt{n}} = 41\cdot5 \pm 2\cdot947 \times \frac{3}{4} = 43\cdot71 \text{ and } 39\cdot29$$

$$\therefore \quad 39\cdot29 < \mu < 43\cdot71$$

EXERCISE 14(b)

1. (*a*) Write a short note on Student's *t*-distribution and point out its uses.

(*b*) Show how the *t*-distribution has been found useful in testing whether the mean of small sample is significantly different from a hypothetical value.

(*c*) It is desired to test the hypothesis that the mean of a normal population is $\mu = \mu_0$ against the alternative that $\mu \neq \mu_0$. Explaining the assumptions involved, develop the statistic suitable for testing this hypothesis if the size of the sample is small. What modification do you suggest when the sample size is large ?

2. What is a test of significance ?

To test the hypothesis that the mean of a normal distribution is zero, two independent observations x_1 and x_2 are taken from the distribution. Show that the hypothesis is rejected at 10% level of significance, using *t* test with equal tail ends, if

$$|x_1 + x_2| > |x_1 - x_2| \tan 81^0$$

3. It is required to test that the mean of a normal population is zero. A random sample drawn from the population gives the values $x_1, x_2, \ldots, x_n$. Show that the *t*-test for acceptance of the hypothesis reduces to

$$\left(\sum_{i=1}^{n} x_i\right) \leq \frac{n \cdot t_\alpha^2}{t_\alpha^2 + (n-1)} \left(\sum_{i=1}^{n} x_i^2\right)$$

where t_α is the value of Student's *t* at the desired level of significance α for $(n-1)$ *d.f.*

4. (*a*) Find the Student's *t* for following variate values in a sample of eight : –4, –2, –2, 0, 2, 2, 3, 3, taking the mean of the universe to be zero. How would you proceed further ?

(*b*) Ten individuals are chosen at random from a normal population and their heights are found to be 63, 63, 66, 67, 68, 69 70, 70, 71, 71 inches. Test if the sample belongs to the population whose mean heights is 66″

[Given $t_{0\cdot05} = 2\cdot62$ for 9 *d.f.*]

(*c*) A random sample of 9 experimental animals under a certain diet gave the following increase in weight : $\Sigma x_i = 45$ lbs, $\Sigma x_i^2 = 279$ lbs., where x_i denotes the increase in weight of the *i*th animal. Assuming that the increase in weight is normally distributed as $N(\mu, \sigma^2)$ variate, test $H_0 : \mu = 1$ against $H_1 : \mu \neq 1$ at 5% level. Given $P(|t| > 2\cdot306) = 0\cdot05$ for 8 degrees of freedom.

[Calcutta Univ. B.Sc. (Maths.Hons.), 1991]

5. A manufacturer of gunpowder has developed a new powder which is designed to produce a muzzle velocity equal to 3000 ft/sec. Seven shells are loaded with the charge and the muzzle velocities measured.

The resulting velocities are as follows : 3,005; 2,935; 2,965; 2,995; 3,905, 2,935; and 2,905. Do these data present sufficient evidence to indicate that the average velocity differs from 3,000 ft./sec.

6. The average length of time for students to register for summer classes at a certain college has been 50 minutes with a standard deviation of 10 minutes. A new registration procedure using modern computing machines is being tried. If a random sample of 12 students had an average registration time of 42 minutes with s.d. of 11·9 minutes under the new system, test the hypothesis that the population mean has not changed, using ·05 as level of significance.

7. The nine items of a sample had the following values : 45, 47, 50, 52, 48, 47, 49, 53 and 51.

Does the mean of the nine items differ significantly from the assumed population mean of 47·5 ? Given that

$$\nu = 8, \begin{cases} P = 0{\cdot}945 \text{ for } t = 1{\cdot}8 \\ P = 0{\cdot}953 \text{ for } t = 1{\cdot}9 \end{cases}$$

8. A time study engineer developed a new sequence of operation elements that he hopes will reduce the mean cycle time of a certain production process. The results of a time study of 20 cycles are given below :

cycle time in minutes

12·25	11·97	12·15	12·08	12·31	12·28	11·94	11·89	12·16	12·04
12·09	12·15	12·14	12·47	11·98	12·04	12·11	12·25	12·15	12·34

If the present mean cycle time is 12·5 minutes, should he adopt the new sequence ?

9. (*a*) The average breaking strength of steel rods is specified to be 18·5 thousand pounds. To test this a sample of 14 rods was tested. The mean and standard deviations obtained were 17·85 and 1·955 thousand pounds respectively. Is the result of the experiment significant ? Also obtain the 95 per cent fiducial limits from the sample for the average breaking strength of steel rods.

(*b*) A sample of 9 shafts is inspected from a production line. The following measurements are the diameters (in mm.) of shafts : 45·010, 45·020, 45·021, 45·015, 45·019, 45·018, 45·020, 45·023 and 45·005. If the production line meets the specifications laid by the I.S.I., with S.D. 0·006 mm, estimate the 95% confidence interval within which the true diameter of the shaft lies.

[*Madras Univ. B.E., 1989*]

10. a random sample of 8 envelopes is taken from letter box of a post office and their weights in grams are found to be 12·1, 11·9, 12·4, 12·3, 11·9, 12·1, 12·4, 12·1.

(*a*) Find 99% confidence limits for the mean weight of the envelopes received at that post office.

(*b*) Using the result of part (*a*), does this sample indicate at 1% level that the average weight of envelopes received at that post office is 12·35 gms.

11. A random sample of nine from men of a large city gave a mean height 68 inches and the unbiased estimate of the population variance found from the

sample was 4·5 inches. Proceed as far as you can to test for a mean height of 68·5 inches for the men of the city. Also state how you would proceed further.

14·2·10. t-Test for Difference of Means. Suppose we want to test if two independent samples x_i ($i = 1, 2, \ldots, n_1$) and y_j, ($j = 1, 2, \ldots, n_2$) of sizes n_1 and n_2 have been drawn from two normal populations with means μ_X and μ_Y respectively.

Under the null hypothesis (H_0) *that the samples have been drawn from the normal populations with means* μ_X *and* μ_Y *and under the assumption that the population variance are equal, i.e.,* $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (say), the statistic

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \ldots(14\cdot7)$$

where $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i, \qquad \bar{y} = \frac{1}{n_2}\sum_{j=1}^{n_2} y_j$

and $S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right] \qquad \ldots[14\cdot7(a)]$

is an unbiased estimate of the common population variance σ^2, follows Student's t-distribution with $(n_1 + n_2 - 2)$ d.f.

Proof. *Distribution of t defined in (14·7).*

$$\xi = \frac{(\bar{x} - \bar{y}) - E(\bar{x} - \bar{y})}{\sqrt{V(\bar{x} - \bar{y})}} \sim N(0, 1)$$

But $E(\bar{x} - \bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_X - \mu_Y$

and $V(\bar{x} - \bar{y}) = V(\bar{x}) + V(\bar{y})$

[The covariance term vanishes since samples are independent.]

$$= \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \qquad \text{(By assumption)}$$

$$\therefore \quad \xi = \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1) \qquad \ldots(*)$$

Let $\chi^2 = \left[\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{j=1}^{n_2} (y_j - \bar{y})^2\right] \Big/ \sigma^2$

$$= \left[\sum_i (x_i - \bar{x})^2/\sigma^2\right] + \left[\sum_j (y_j - \bar{y})^2/\sigma^2\right] = \frac{n_1 s_X^2}{\sigma^2} + \frac{n_2 s_Y^2}{\sigma^2}$$

$\ldots(**)$

Since $n_1 s_X^2/\sigma^2$ and $n_2 s_Y^2/\sigma^2$ are independent χ^2-variates with $(n_1 - 1)$ and $(n_2 - 1)$ d.f. respectively, by the additive property of chi-square distribution, χ^2 defined in (**) is a χ^2-variate with $(n_1 - 1) + (n_2 - 1)$, *i.e.*, $n_1 + n_2 - 2$ *d.f.*

Further, since sample mean and sample variance are independently distributed, ξ and χ^2 are independent random variables.

Hence Fisher's t statistic is given by

$$t = \frac{\xi}{\sqrt{\frac{\chi^2}{n_1 + n_2 - 2}}}$$

$$= \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \times \frac{1}{\left[\frac{1}{n_1 + n_2 - 2}\left\{\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right\}/\sigma^2\right]^{1/2}}$$

$$= \frac{(\bar{x} - \bar{y}) - (\mu_X - \mu_Y)}{S\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where $$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right]$$

and it follows Student's t-distribution with $(n_1 + n_2 - 2)$ d.f. (*c.f.* Remark § 14·2·3, page 14·4).

Remarks 1. S^2, defined in 14·7(a) is an unbiased estimate of the common population variance σ^2, since

$$E(S^2) = \frac{1}{n_1 + n_2 - 2} E\left[\sum_i (x_i - \bar{x})^2 + \sum_j (y_j - \bar{y})^2\right]$$

$$= \frac{1}{n_1 + n_2 - 2} E\left[(n_1 - 1) S_X^2 + (n_2 - 1) S_Y^2\right]$$

$$= \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1) E(S_X^2) + (n_2 - 1) E(S_Y^2)\right]$$

$$= \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2\right] = \sigma^2$$

2. An important deduction which is of much practical utility is discussed below :

Suppose we want to test if : (a) two independent samples x_i ($i = 1, 2, \ldots, n_1$), and y_j ($j = 1, 2, \ldots, n_2$), have been drawn from the populations with same means or (b) the two sample means $\bar{x}$ and $\bar{y}$ differ significantly or not.

Under the null hypothesis H_0 that (a) *samples have been drawn from the populations with the same means,i.e.,* $\mu_X = \mu_Y$ *or (b) the sample means* $\bar{x}$ *and* $\bar{y}$ *do not differ significantly,* [From (14·7)] the statistic :

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad [\because \mu_X = \mu_Y, \text{ under } H_0] \qquad \ldots(14\cdot8)$$

where symbols are defined in (14·7*a*),follows Student's t-distribution with $(n_1 + n_2 - 2)$ d.f.

3. *On the assumption of t-test for difference of means.* Here we make the following three fundamental assumptions :

(*i*) Parent populations, from which the samples have been drawn are normally distributed.

(*ii*) The population variances are equal and unknown, *i.e.*, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, (say), where σ^2 is unknown.

(*iii*) The two samples are random and independent of each other.

Thus before applying t-test for testing the equality of means it is theoretically desirable to test the equality of population variances by applying F-test. If the variances do not come out to be equal then t-test becomes invalid and in that case Behren's 'd'-test based on fiducial intervals is used. For practical problems, however, the assumptions (*i*) and (*ii*) are taken for granted.

4. Paired t-test For Difference of Means. Let us now consider the case when (*i*) the sample sizes are equal, *i.e.*, $n_1 = n_2 = n$ (say), and (*ii*) the two samples are not independent but the sample observations are paired together, *i.e.*, the pair of observations (x_i, y_i), $(i = 1, 2, \ldots, n)$ corresponds to the same (ith) sample unit. The problem is to test if the sample means differ significantly or not.

For example,suppose we want to test the efficacy of a particular drug, say, for inducing sleep. Let x_i and y_i $(i = 1, 2, \ldots, n)$ be the readings, in hours of sleep, on the ith individual, before and after the drug is given respectively. Here instead of applying the difference of the means test discussed in § 14·2·10, we apply the paired t-test given below.

Here we consider the increments, $d_i = x_i - y_i$, $(i = 1, 2, \ldots, n)$.

Under the null hypothesis, H_0 that *increments are due to fluctuations of sampling, i.e., the drug is not responsible for these increments,* the statistic.

$$t = \frac{\bar{d}}{S/\sqrt{n}} \qquad \ldots(14\cdot9)$$

where $$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 \qquad \ldots[14\cdot9(a)]$$

follows Student's t-distribution with $(n - 1)$ d.f.

Example 14·7. *Below are given the gain in weights (in lbs.) of pigs fed on two diets A an B.*

Gain in weight

Diet A : 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

Diet B : 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22

Test, if the two diets differ significantly as regards their effect on increase in weight.

Solution. *Null hypothesis,* $H_0 : \mu_X = \mu_Y$, *i.e., there is no significant difference between the mean increase in weight due to diets A and B.*

Alternative hypothesis, $H_1 : \mu_X \neq \mu_Y$ (two-tailed).

Diet A

	X	$X - \bar{X}$	$(X - \bar{X})^2$
	25	−3	9
	32	4	16
	30	2	4
	34	6	36
	24	−4	16
	14	−14	196
	32	4	16
	24	−4	16
	30	2	4
	31	3	9
	35	7	49
	25	−3	9
Total	336	0	380

Diet B

	Y	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
	44	14	196
	34	4	16
	22	−8	64
	10	−20	400
	47	17	289
	31	1	1
	40	10	100
	30	0	0
	32	2	4
	35	5	25
	18	−12	144
	21	−9	81
	35	5	25
	29	−1	1
	22	−8	64
Total	450	0	1410

Under null hypothesis (H_0) :

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2}$$

Here $\left.\begin{matrix} n_1 = 12, \\ \Sigma x = 336 \\ \Sigma(x - \bar{x})^2 = 380 \end{matrix}\right\}$ and $\left.\begin{matrix} n_2 = 15 \\ \Sigma y = 450 \\ \Sigma(y - \bar{y})^2 = 1410 \end{matrix}\right\}$

$$\therefore \quad \bar{x} = \frac{336}{12} = 28, \quad \bar{y} = \frac{450}{15} = 30$$

and $$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2\right] = 71{\cdot}6$$

$$\therefore \quad t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{28 - 30}{\sqrt{71{\cdot}6\left(\frac{1}{12} + \frac{1}{15}\right)}}$$

$$= \frac{-2}{\sqrt{10{\cdot}74}} = -0{\cdot}609$$

Tabulate $t_{0{\cdot}05}$ for $(12 + 15 - 2) = 25$ d.f. is 2·06.

Conclusion. Since calculated $|t|$ is less than tabulated t, H_0 may be accepted at 5% level of significance and we may conclude that the two diets do not differ significantly as regards their effect on increase in weight.

Remark. Here $\bar{x}$ and $\bar{y}$ come out to be integral values and hence the direct method of computing $\Sigma(x-\bar{x})^2$ and $\Sigma(y-\bar{y})^2$ is used. In case $\bar{x}$ and (or) $\bar{y}$ comes out to be fractional, then the step deviation method is recommended for computation of $\Sigma(x-\bar{x})^2$ and $\Sigma(y-\bar{y})^2$.

Example 14·8. *Samples of two types of electric light bulbs were tested for length of life and following data were obtained :*

	Type I	*Type II*
Sample No.	$n_1 = 8$	$n_2 = 7$
Sample Means	$\bar{x}_1 = 1{,}234$ *hrs.*	$\bar{x}_2 = 1{,}036$ *hrs.*
Sample S.D.'s	$s_1 = 36$ *hrs.*	$s_2 = 40$ *hrs.*

Is the difference in the means sufficient to warrant that type 1 is superior to type II regarding length of life ?

Solution. *Null Hypothesis,* $H_0 : \mu_X = \mu_Y$, *i.e.,* the two types I and II of electric bulbs are indentical.

Alternative Hypothesis, $H_1 : \mu_X > \mu_Y$, *i.e.,* type 1 is superior to type II.

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t_{13},$$

where $$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2\right]$$

$$= \frac{1}{n_1 + n_2 - 2}\left[n_1 s_1^2 + n_2 s_2^2\right] = \frac{1}{13}\left[8 \times (36)^2 + 7 \times (40)^2\right] = 1659{\cdot}08$$

$$\therefore \quad t = \frac{1234 - 1036}{\sqrt{1659{\cdot}08\left(\frac{1}{8} + \frac{1}{7}\right)}} = \frac{198}{\sqrt{1659{\cdot}08 \times 0{\cdot}2679}} = 9{\cdot}39$$

Tabulated value of t for 13 *d.f.* at 5% level of significance for right (single) tailed test is 1·77. [This is the value of $t_{0{\cdot}10}$ for 13 *d.f.* from two-tail tables given in Appendix].

Conclusion. Since calculated 't' is much greater than tabulated 't', it is highly significant and H_0 is rejected. Hence the two types of electric bulbs differ significantly. Further since $\bar{x}_1$ is much greater than $\bar{x}_2$, we conclude that type I is definitely superior to type II.

Example 14·9. *The heights of six randomly chosen sailors are in inches : 63, 65, 68, 69, 71, and 72. Those of 10 randomly chosen soldiers are 61, 62, 65, 66, 69, 69, 70, 71, 72 and 73. Discuss, the light that these data throw on the suggestion that sailors are on the average taller than soldiers.*

Solution. If the heights of sailors and soldiers be represented by the variables X and Y respectively then the *Null Hypothesis* is, $H_0 : \mu_X = \mu_Y$, *i.e., the sailors are not on the average taller than the soldiers.*

Alternative Hypothesis, $H_1 : \mu_X > \mu_Y$ (Right-tailed).

Under H_0, the test statistic is :

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t_{14}$$

Sailors			*Soldiers*		
X	$d = X - A = X - 68$	d^2	Y	$D = Y - B = Y - 66$	D^2
63	−5	25	61	−5	25
65	−3	9	62	−4	16
68	0	0	65	−1	1
69	1	1	66	0	0
71	3	9	69	3	9
72	4	16	69	3	9
			70	4	16
			71	5	25
Total	0	60	72	6	36
			73	7	49
			Total	18	186

$$\therefore \quad \bar{x} = A + \frac{\Sigma d}{n_1}$$

$$= 68 + 0 = 68$$

$$\text{and } \Sigma(x - \bar{x})^2 = \Sigma d^2 - \frac{(\Sigma d)^2}{n_1}$$

$$= 60 - 0 = 60$$

$$\bar{y} = B + \frac{\Sigma D}{n_2}$$

$$= 66 + \frac{18}{10} = 67{\cdot}8$$

$$\text{and } \Sigma(y - \bar{y})^2 = \Sigma D^2 - \frac{(\Sigma D)^2}{n_2}$$

$$= 186 - \frac{324}{10} = 153{\cdot}6$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2\right] = \frac{1}{14}(60 + 153{\cdot}6) = 15{\cdot}2571$$

$$\therefore \quad t = \frac{68 - 67{\cdot}8}{\sqrt{15{\cdot}2571}\left(\frac{1}{6} + \frac{1}{10}\right)^{1/2}} = \frac{0{\cdot}2}{\sqrt{15{\cdot}2571 \times 0{\cdot}2667}} = 0{\cdot}099$$

Tabulated $t_{0{\cdot}05}$ for 14 *d.f.* for single-tail test is 1·76.

Conclusion. Since calculated t is much less than 1·76, it is not at all significant at 5% levels of significance. Hence null hypothesis may be retained at 5% level of significance and we conclude that the data are inconsistent with the suggestion that the sailors are on the average taller than soldiers.

Example 14·10. *A certain stimulus administered to each of the 12 patients resulted in the following increase of blood pressure :*

5, 2, 8, –1, 3, 0, –2, 1, 5, 0, 4 and 6

Can it be concluded that the stimulus will, in general, be accompanied by an increase in blood pressure ? **[*Delhi Univ. B.Sc. 1989*]**

Solution. Here we are given the increments in blood pressure *i.e.*,

$d_i\ (= x_i - y_i)$.

Null Hypothesis, $H_0 : \mu_X = \mu_Y$, *i.e.*, there is no significant difference in the blood pressure readings of the patients before and after the drug. In other words, the given increments are just by chance (fluctuations of sampling) and not due to the stimulus.

Alternative Hypothesis, $H_1 : \mu_X < \mu_Y$, *i.e.*, the stimulus results is an increase in blood pressure.

Test Statistic. Under H_0, the test statistic is :

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$$

d	5	2	8	–1	3	0	–2	1	5	0	4	6	31
d^2	25	4	64	1	9	0	4	1	25	0	16	36	185

$$S^2 = \frac{1}{n-1}\Sigma(d - \bar{d})^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right]$$

$$= \frac{1}{11}\left[185 - \frac{(31)^2}{12}\right] = \frac{1}{11}(185 - 80{\cdot}08) = 9{\cdot}5382$$

and $$\bar{d} = \frac{\Sigma d}{n} = \frac{31}{12} = 2{\cdot}58$$

$$\therefore \quad t = \frac{\bar{d}}{S/\sqrt{n}} = \frac{2{\cdot}58 \times \sqrt{12}}{\sqrt{9{\cdot}5382}} = \frac{2{\cdot}58 \times 3{\cdot}464}{3{\cdot}09} = 2{\cdot}89$$

Tabulated $t_{0{\cdot}05}$ for 11 d.f. for right-tail test is 1·80. [This is the value of $t_{0{\cdot}10}$ for 11 *d.f* in the Table for two-tailed test given in the Appendix].

Conclusion. Since calculated $t > t_{0.05}$, H_0 is rejected at 5% level of significance. Hence we conclude that the stimulus will, in general, be accompanied by an increase in blood pressure.

Example 14·11. *In a certain experiment to compare two types of pig foods A and B, the following results of increase in weights were observed in pigs :*

Pig number		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Total*
Increase in weight in lb	*Food A*	*49*	*53*	*51*	*52*	*47*	*50*	*52*	*53*	*407*
	Food B	*52*	*55*	*52*	*53*	*50*	*54*	*54*	*53*	*423*

(*i*) *Assuming that the two samples of pigs are independent, can we conclude that food B is better than food A ?*

(*ii*) *Also examine the case when the same set of eight pigs were used in both the foods.*

Solution. *Null Hypothesis,* H_0. If the increase in weights due to foods *A* and *B* are denoted by *X* and *Y* respectively then $H_0 : \mu_X = \mu_Y$, *i.e.*, there is no significant difference in increase in weights due to diets *A* and *B*.

Alternative Hypothesis, $H_1 : \mu_X < \mu_Y$ (Left-tailed).

(*i*)If the two samples of pigs be assumed to be independent, then we will apply *t*-test for difference of means to test H_0.

Test Statistic. Under $H_0 : \mu_X = \mu_Y$, the test criterion is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2}$$

Food A			*Food B*		
X	$d = X - 50$	d^2	*Y*	$D = Y - 52$	D^2
49	−1	1	52	0	0
53	3	9	55	3	9
51	1	1	52	0	0
52	2	4	53	1	1
47	−3	9	50	−2	4
50	0	0	54	2	4
52	2	4	54	2	4
53	3	9	53	1	1
	7	37		7	23

$$\therefore \bar{x} = 50 + \frac{7}{8} = 50{\cdot}875 \qquad \bar{y} = 52 + \frac{7}{8} = 52{\cdot}875$$

and $\Sigma(x - \bar{x})^2 = \Sigma d^2 - \frac{(\Sigma d)^2}{n_1} = 37 - \frac{49}{8} = 30{\cdot}875$ $\Sigma(y - \bar{y})^2 = \Sigma D^2 - \frac{(\Sigma D)^2}{n_2} = 23 - \frac{49}{8} = 16{\cdot}875$

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2\right]$$

$$= \frac{1}{14}(30{\cdot}875 + 16{\cdot}875) = 3{\cdot}41$$

$$\therefore \quad t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{50{\cdot}875 - 52{\cdot}875}{\sqrt{3{\cdot}41\left(\frac{1}{8} + \frac{1}{8}\right)}} = -2{\cdot}17$$

Tabulated $t_{0{\cdot}05}$ for $(8 + 8 - 2) = 14$ d.f. for one-tail test is 1·76.

Conclusion. The critical region for the left-tail test is $t < -1{\cdot}76$. Since calculated t is less than $-1{\cdot}76$, H_0 is rejected at 5% level of significance. Hence we conclude that the foods A and B differ significantly as regards their effect on increase in weight. Further, since $\bar{y} > \bar{x}$, food B is superior to food A.

(*ii*) If the same set of pigs is used in both the cases, then the readings X and Y are not independent but they are paired together and we apply the paired t-test for testing H_0.

Under $H_0 : \mu_X = \mu_Y$, the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{(n-1)}$$

X	49	53	51	52	47	50	52	53	Total
Y	52	55	52	53	50	54	54	53	
$d = X - Y$	−3	−2	−1	−1	−3	−4	−2	0	−16
d^2	9	4	1	1	9	16	4	0	44

$$\therefore \quad \bar{d} = \frac{\Sigma d}{n} = \frac{-16}{8} = -2$$

and $$S^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right] = \frac{1}{7}\left[44 - \frac{256}{8}\right] = 1{\cdot}714$$

$$\therefore \quad |t| = \frac{|\bar{d}|}{\sqrt{S^2/n}} = \frac{2}{\sqrt{1{\cdot}7143/8}} = \frac{2}{0{\cdot}4629} = 4{\cdot}32$$

Tabulated $t_{0{\cdot}05}$ for $(8 - 1) = 7$ d.f. for one-tail test is 1·90.

Conclusion. Here also the observed value of '*t*' is significant at 5% level of significance and we conclude that food *B* is superior to food *A*.

EXERCISE 14 (c)

1. Explain, stating clearly the assumptions involved, the *t*-test for testing the significance of the difference between the two sample means.

2. Two independent samples of 8 and 7 items respectively had the following values

Sample I...	9	11	13	11	15	9	12	14
Sample II...	10	12	10	14	9	8	10	

Is the difference between the means of samples significant ?

3. (*a*) Two horses *A* and *B* were tested according to the time (in seconds) to run a particular track with the following results :

Horse A :	28	30	32	33	33	29	34
Horse B :	29	30	30	24	27	29	

Test whether the two horses have the same running capacity. [5 per cent values of *t* for 11 and 12 degrees of freedom respectively are 2·20 and 2·18].

Ans. Calculated $t = 2{\cdot}5$ (approx.)

(*b*) The gain in weight of two random samples of rats fed on two different diets *A* and *B* are given below. Examine whether the difference in mean increases in weight is significant.

Diet A :	13	14	10	11	12	16	10	8	
Diet B :	7	10	12	8	10	11	9	10	11

4. (*a*) Show how you would use Student's *t*-test to decide whether the two sets of observations

[17, 27, 18, 25, 27, 29, 27, 23, 17] and [16, 16, 20, 16, 20, 17, 15, 21]

indicate samples drawn from the same universe.

(*b*) A reading test is given to an elementary school class that consists of 12 Anglo-American children and 10 Mexican-American children. The results of the test are :

Anglo-American	*Mexican-American*
$\bar{x}_1 = 74$	$\bar{x}_2 = 70$
$s_1 = 8$	$s_2 = 10$

Is the difference between the means of the two groups significant at the 0·05 level ? Given $t_{20} = 2{\cdot}086$, $t_{22} = 2{\cdot}074$ at 5% level.

[Delhi Univ. M.C.A., 1986]

5. (*a*) For a random sample of 10 pigs, fed on a diet *A*, the increases in weight in a certain period were 10, 6, 16, 17, 13, 12, 8, 14, 15, 9 lbs.

For another random sample of 12 pigs fed on diet *B*, the increases in the same period were 7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17 lbs.

Find if the two samples are significantly different regarding the effect of diet, given that for d.f. $\nu = 20, 21, 22$, the five per cent values of *t* are respectively 2·09, 2·07, 2·06.

Ans. $t = 1{\cdot}51$; Sample means do not differ significantly.

(*b*) Two independent samples of rats chosen among both the series had the following increase in weights when fed on a diet. Can you say that the mean increase in weight differs significantly with sex ?

Male : 96, 88, 97. 89, 92, 95 and 90

Female : 112, 80, 98, 100, 84, 82, 89, 95, 100 and 96.

6. (*a*) Ten soldiers visit a riffle range for two consecutive weeks. For the first week their scores are

67, 24, 57, 55, 63, 54, 56, 68, 33, 43

and during the second week they score in the same order—

70, 38, 58, 58, 56, 67, 68, 72, 42, 38

Examine if there is any significant difference in their performance.

(*b*) Two independent groups of 10 children were tested to find how many digits they could repeat from memory after hearing them. The results are as follows :

Group A	:	8	6	5	7	6	8	7	4	5	6
Group B	:	10	6	7	8	6	9	7	6	7	7

Is the difference between the mean scores of the two groups significant ?

(*c*) Measurements of the fat content of two kinds of ice cream, Brand *A* and Brand *B*, yielded the following sample data :

Brand A	:	13·5	14·0	13·6	12·9	13·0
Brand B	:	12·9	13·0	12·4	13·5	12·7

Test the null hypothesis $\mu_1 = \mu_2$, (where μ_1 and μ_2 are the respective true average fat contents of the two kinds of ice cream), against the alternative hypothesis $\mu_1 \neq \mu_2$ at the level of significance $\alpha = 0{\cdot}05$.

[Madras Univ. B.E., 1990]

7. (*a*) A random sample of 16 values from a normal population has a mean of 41·5 inches and sum of squares of deviations from the mean is equal to 135 inches. Another sample of 20 values form an unknown population has a mean of 43·0 inches and sum of squares of deviations from their mean is equal to 171 inches. Show that the two samples may be regarded as coming from the same normal population.

(*b*) A company is interested in knowing if there is a difference in the average salary received by foremen in two divisions. Accordingly samples of 12 foremen in the first division and 10 foremen in the second division are selected at random. Based upon experience, foremen's salaries are known to be approximately normally distributed, and the standard deviations are about the same.

	First Division	*Second division*
Sample size	12	10
Average monthly salary of foremen (Rs.)	1,050	980
Standard deviation of salaries (Rs.)	68	74

The table value of t for 20 d.f. at 5% level of significance is 2·086.

Ans. $t = 2{\cdot}2$. Reject $H_0 : \mu_X = \mu_Y$

(*c*) The average number of articles produced by two machines per day are 200 and 250 with standard deviations 20 and 25 respectively on the basis of records of 25 days production. Can you regard both the machines equally efficient at 1% level of significance ?

Ans. $t = -7{\cdot}65$. **Hint.** Here $n_1 = n_2 = 25$.

8. Eleven school boys were given a test in Statistics. They were given a month's tuition and a second test was held at the end of it. Do the marks give evidence that the students have benefited by the extra coaching ?

Boys	1	2	3	4	5	6	7	8	9	10	11
Marks in 1st test	23	20	19	21	18	20	18	17	23	16	19
Marks in 2nd test	24	19	22	18	20	22	20	20	23	20	18

Ans. H_0: $\mu_1 = \mu_2$; H_1 : $\mu_1 < \mu_2$. Paired $|t| = 1{\cdot}483$. Not significant. Hence, students have not benefited from extra coaching.

9. (*a*) The following table gives the additional hours of sleep gained by 10 patients in an experiment to test the effect of a drug. Do these data give evidence that the drug produces additional hours of sleep ?

Patients	1	2	3	4	5	6	7	8	9	10
Hours gained :	0·7	0·1	0·2	1·2	0·31	0·4	3·7	0·8	3·8	2·0

(*b*) A drug was administered to 10 patients, and the increments in their blood pressure were recorded to be 6, 3, –2, 4, –3, 4, 6,0, 3, 2. Is it reasonable to believe that the drug has no effect on change of blood pressure ? Use 5% significance level, and assume that for 9 degrees of freedom, $P(t > 2{\cdot}26) = 0{\cdot}025$. **[Calcutta Univ. B.Sc.(Maths. Hons.), 1986]**

(*c*) The scores of 10 candidates prior and after training are given below :

Prior :	84	48	36	37	54	69	83	96	90	65
After :	90	58	56	49	62	81	84	86	84	75

Is the training effective ? **[Calicut Univ. B.Sc., Oct. 1992]**

10. The following table gives measurements of blood pressure on subjects by two investigators :

Subject No. :	1	2	3	4	5	6	7	8	9	10
Investigator I :	70	68	56	75	80	90	68	75	56	58
Investigator II :	68	70	52	73	75	78	67	70	54	55

No other details of the experiment were given.

(*i*) If a valid inference has to be drawn about the difference between the investigators, mention the precautions that should have been taken in conducting the experiment with respect to the time of measurement, interval between the first and second measurements, the order in which the investigators measure, etc.

(*ii*) After the experiment was conducted it was discovered that all the subjects were unrelated except that No. 10 was the father of No. 9. Assuming that all the precautions you mention in (*a*) are satisfied, analyse the data to draw an inference on the difference between the investigators. 5 per cent values of the *t*-statistic corresponding to various degrees of freedom are as follows :

5 per cent values of *t*...	2·40	2·31	2·26	2·23	2·10	2·09
Degrees of freedom...	7	8	9	10	18	19

11. The following are the values of the cephalic index found in two samples of skulls, one consisting of 15 and the other of 13 individuals.

Sample I :	74·1	77·7	74·0	74·4	73·8	79·3	75·8	82·8
	72·2	75·2	78·2	77·1	78·4	76·3	76·8	
Sample II :	70·8	74·9	74·2	70·4	69·2	72·2	76·8	72·4
	77·4	78·1	72·8	74·3	74·7			

(*i*) Test the hypothesis that the means of population I and population II could be equal.

(*ii*) Is it possible that the sample II has come from a population of mean 72·0 ?

(*iii*) Obtain confidence limits for the mean of population I and for the mean of population II.

(Assume that the distribution of cephallic indices for a homogeneous population is normal.)

12. (*a*) The following table gives the gain in weight in decagrams in a feeding experiment with pigs on the relative value of limestone and bone meal for bone development.

Limestone	49·2	53·3	50·6	52·0	46·8	50·5	52·1	53·0
Bone meal	51·5	54·9	52·2	53·3	51·6	54·1	54·2	53·3

Test for the significance of difference between the means in two ways :

(*i*) by assuming that the values are paired.

(*ii*) by assuming that the values are not paired.

(*b*) The following table shows the mean number of bacterial colonies per plate obtainable by four slightly different methods from soil samples taken at 4 P.M. and 8 P.M. respectively.

Method	*A*	*B*	*C*	*D*
4 P.M.	29·75	27·50	30·25	27·80
8 P.M.	39·20	40·60	36·20	42·40

Are there significantly more bacteria at 8 P.M. than at 4 P.M. ?

[Given $t_{0.05}(3) = 3{\cdot}18$ and $t_{0.01}(3) = 5{\cdot}84$]

13. (*a*) It is believed that glucose treatment will extend the sleep time of mice. In an experiment to test this hypothesis ten mice selected at random are given glucose treatment and are found to have a mean hexabarbital sleep time of 47·2 min with a standard deviation of 9·3 min. A further sample of ten untreated mice are found to have a mean hexabarbital sleep time of 28·5 min. with a standard deviation of 7·2 min. Are these results significant evidence in favour of the hypothesis ?

Find 95% confidence limits for the population mean difference in sleep time. State any assumptions made concerning the data in carrying out the test and finding the limits. [***Bangalore Univ. B.E., Oct. 1992***]

(*b*) An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material I were tested, by exposing each piece to a machine measuring wear. Ten pieces of material II were similarly tested. In each case the depth of wear was observed. The sample of material I gave an average (coded) wear 8·5 units with a standard deviation of 0·4

while the sample of material II gave an average of 8·1 and a standard deviation of 0·5. Test the hypothesis that the two types of material exhibit the same mean abrasive wear at the 0·10 level of significance. Assume the populations to be approximately normal with equal variances.

If the level of significance is 0·01, what will be your conclusion ?

[*Delhi Univ. M.E., 1992*]

14·2·11. t-test For Testing Significance of an Observed sample Correlation Coefficient. If r is the observed correlation coefficient in a sample of n pairs of observations from a bivariate normal population, then Prof. Fisher proved that under the null hypothesis $H_0 : \rho = 0$, *i.e., population correlation coefficient is zero*, the statistic :

$$t = \frac{r}{\sqrt{(1 - r^2)}}\sqrt{(n - 2)} \qquad \ldots(14{\cdot}9)$$

follows Student's t-distribution with $(n - 2)$ d.f. (c.f. Remark to § 14·3 page 14·41).

If the value of t comes out to be significant, we reject H_0 at the level of significance adopted and conclude that $\rho \neq 0$, *i.e.*, 'r' is significant of correlation in the population.

If t comes out to be non-significant then H_0 may be accepted and we conclude that variables may be regarded as uncorrelated in the population.

Example 14·12. *A random sample of 27 pairs of observations from a normal population gave a correlation coefficient of 0·6. Is this significant of correlation in the population ?*

Solution. We set up the null hypothesis, $H_0 : \rho = 0$, *i.e.*, the *observed sample correlation coefficient is not significant of any correlation in the population.*

Under H_0 : $\quad t = \dfrac{r\sqrt{(n - 2)}}{\sqrt{(1 - r^2)}} \sim t_{(n-2)}$

Here $\quad t = \dfrac{0{\cdot}6\sqrt{27 - 2}}{\sqrt{(1 - 0{\cdot}36)}} = \dfrac{3}{\sqrt{0{\cdot}64}} = 3{\cdot}75$

Tabulated $t_{0{\cdot}05}$ for $(27 - 2) = 25$ d.f. is 2·06.

Conclusion. Since calculated t is much greater than the tabulated t, it is significant and hence H_0 is discredited at 5% level of significance. Thus we conclude that the variables are correlated in the population.

Example 14·13. *Find the least value of r in a sample of 18 pairs of observations from a bi-variate normal population, significant at 5% level of significance.*

Solution. Here $n = 18$. From the tables $t_{0{\cdot}05}$ for $(18 - 2) = 16$ d.f. is 2·12

Under $H_0 : \rho = 0$, $\quad t = \dfrac{r\sqrt{(n - 2)}}{\sqrt{(1 - r^2)}} \sim t_{(n-2)}$

In order that the calculated value of t is significant at 5% level of significance, we should have

$$\left|\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}\right| > t_{0.05} \Rightarrow \left|\frac{r\sqrt{16}}{\sqrt{(1-r^2)}}\right| > 2.12$$

$$\Rightarrow \quad 16r^2 > (2.12)^2(1-r^2) \Rightarrow 20.493r^2 > 4.493$$

$$\Rightarrow \quad r^2 > \frac{4.493}{20.493} = 0.2192$$

Hence $|r| > 0.4682$

Example 14.14. *A coefficient of correlation of 0.2 is derived from a random sample of 625 pairs of observations. (i) Is this value of r significant ? (ii) What are the 95% and 99% confidence limits to the correlation coefficient in the population ?*

Solution. *Under the null hypothesis* $H_0 : \rho = 0$, *i.e., the value of* $r = 0.2$ *is not significant; the test statistics is :*

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Now $$t = \frac{0.2 \times \sqrt{(625-2)}}{\sqrt{(1-0.04)}} = 5.09$$

Since $d.f. = 625 - 2 = 623$, the significant values of t are same as in the case of normal distribution, *viz.*, $t_{0.05} = 1.96$ and $t_{0.01} = 2.58$. Since calculated t is much greater than these values; it is highly significant. Hence $H_0 : \rho = 0$ is rejected and we conclude that the sample correlation is significant of correlation in the population.

95% Confidence Limits for ρ *(population correlation coefficient)* are

$$r \pm 1.96 \text{ S.E. } (r) = r \pm 1.96\,(1-r^2)/\sqrt{n} \quad \text{[Since } n \text{ large]}$$

$$= 0.2 \pm (1.96 \times 0.96/\sqrt{625})$$

$$= 0.2 \pm 0.075 = (0.125, 0.275)$$

99% Confidence Limits for ρ are :

$$0.2 \pm 2.58 \times 0.0384 = 0.2 \pm 0.099 = (0.101, 0.299)$$

EXERCISE 14 (d)

1. A restaurant owner ranked his 17 waiters in terms of their speed and efficiency on the job. He correlated these ranks with the total amount of tips each of these waiters received for a one-week period. The obtained value of correlation coefficient is 0.438. What do you conclude ?

Given : $t_{15}(0.05) = 2.131$, $t_{16}(0.05) = 2.120$ for two-tailed test.

[Delhi Univ. M.C.A., 1990]

2. Test the significance of the values of correlation coefficient 'r' obtained from samples of size n pairs from a bivariate normal population.

(*i*) $r = 0.6, n = 38$ (*ii*) $r = 0.5, n = 11$

Ans. (*i*) $t = 4.5$; Significant at 5% level; $H_0 : \rho = 0$ rejected.

(*ii*) $t = 1.73$; Not significant at 5% level.

(*i*) Consistent Statistic
(*ii*) Unbiased Statistic
(*iii*) Sufficient Statistic
(*iv*) Efficiency. [***Delhi Univ. B.Sc. (Stat. Hons.), 1987, 1982***]

2. What do you understand by Point Estimation ? When would you say that estimate of a parameter is good ? In particular, discuss the requirements of consistency and unbiasedness of an estimate. Give an example to show that a consistent estimate need not be unbiased.

[***Delhi Univ. B.Sc. (Stat. Hons.), 1992, 1986***]

3. Discuss the terms (*i*) estimate, (*ii*) consistent estimate, (*iii*) unbiased estimate, of a parameter and show that sample mean is both consistent and unbiased estimate of the population mean.

[***Calcutta Univ. B.Sc. (Maths. Hons.), 1986***]

4. (*a*) If $s_1^2, s_2^2, \ldots, s_r^2$ are r sample variances based on random samples of sizes $n_1, n_2, \ldots, n_r$ respectively, and if T is some statistic given by

$$T = \frac{n_1 s_1^2 + n_2 s_2^2 + \ldots + n_r s_r^2}{a},$$

for estimating σ^2 as an unbiased estimator, find the value, of a, supposing population is very large and for every sample

$$s^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

Ans. $a = (n_1 + n_2 + \ldots + n_r) - r.$

(*b*) If $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_r$ are the sample means based on samples of sizes $n_1, n_2, n_3, \ldots, n_r$ respectively, an unbiased estimator

$$t = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \ldots + n_r\bar{X}_r}{k}$$

has been defined to estimate μ. Find the value of k.

Ans. $k = n_1 + n_2 + \ldots + n_r.$

5. (*a*) For the geometric distribution,

$$f(x, \theta) = \theta\,(1-\theta)^{x-1},\ (x = 1, 2, \ldots),\ 0 < \theta < 1,$$

Obtain an unbiased estimator of $1/\theta$. [**Ans.** $E(\bar{X}) = 1/\theta$.]

(*b*) The random variable X takes the values 1 and 0 with respective probabilities θ and $1 - \theta$. Independent observations $X_1, X_2, \ldots, X_n$ on X are available. Write $\xi = X_1 + X_2 + \ldots + X_n$.

Show that $\xi\,(n - \xi)/n(n-1)$ is an unbiased estimate of $\theta(1-\theta)$.

6. Show that if T is an unbiased estimator of a parameter θ, then $\lambda_1 T + \lambda_2$ is an unbiased estimator of $\lambda_1\theta + \lambda_2$, where λ_1 and λ_2 are known constants, but T^2 is a biased estimator of θ^2.

7. For the following cases determine if the given estimator is unbiased for the parametric function. When it is biased, derive an unbiased estimator from it. $\bar{x}$ is the sample mean.

Proof. Let (x_i, y_i), $(i = 1, 2, \ldots, n)$ be a random sample of size n drawn from an uncorrelated bivariate normal population ($\rho = 0$) in which $E(X) = E(Y) = 0$ and $V(X) = \sigma_X^2$, $V(Y) = \sigma_Y^2$. Let the variable Y be transformed to the variable Z by means of a linear orthogonal transformation, *viz.*,

$$\mathbf{Z} = \mathbf{C}\mathbf{Y}$$

where $\mathbf{Z}_{n\times 1} = (z_1, z_2, \ldots, z_n)'$, $\mathbf{Y}_{n\times 1} = (y_1, y_2, \ldots, y_n)'$ and $C_{n\times n} = (c_{ij})$, $\mathbf{C}$ is an orthogonal matrix. Let us, in particular, take

$$c_{11} = c_{12} = \ldots = c_{1n} = 1/\sqrt{n},$$

so that

$$z_1 = \frac{1}{\sqrt{n}}(y_1 + y_2 + \ldots + y_n) = \sqrt{n}\,\bar{y}$$

Now proceeding as in (*Theorem 13·5*), we get

$$\sum_{i=2}^{n} z_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = n s_Y^2$$

Since in a bivariate normal distribution, the marginal distributions of X and Y are also normal, we have $Y \sim N(0, \sigma_Y^2)$. Hence by Fisher's Lemma (*Theorem 13·4*) z_i, $(i = 1, 2, \ldots, n)$ are independent $N(0, \sigma_Y^2)$.

Now

$$r = \frac{\text{Cov}(X, Y)}{s_X s_Y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n\, s_X\, s_Y}$$

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

$$\therefore \quad \sqrt{n}\, s_Y\, r = \frac{\sum (x_i - \bar{x})\, y_i}{\sqrt{n}\, s_X} = z_2, \text{ (say)}, \qquad \ldots(**)$$

[since the sum of the squares of coefficients of $y_1, y_2, \ldots, y_n$ in (**) is unity.]
From (*) and (**), we get

$$n s_Y^2 = \sum_{i=2}^{n} z_i^2 = \sum_{i=3}^{n} z_i^2 + z_2^2 = \sum_{i=3}^{n} z_i^2 + n r^2 s_Y^2$$

$$\Rightarrow \qquad (1 - r^2)\, n\, s_Y^2 = \sum_{i=3}^{n} z_i^2 \qquad \ldots(***)$$

Since z_i, $(i = 1, 2, \ldots, n)$ are independent $N(0, \sigma_Y^2)$;

$\therefore$ (z_i/σ_Y), $(i = 1, 2, \ldots, n)$ are independent $N(0, 1)$.

Hence from (**),

$$U = \frac{z_2^2}{\sigma_Y^2} = \frac{n r^2 s_Y^2}{\sigma_Y^2}$$

being the square of a standard normal variate is a χ^2-variate with 1 d.f. and from (***),

$$V = \sum_{i=3}^{n} z_i^2/\sigma_Y^2 = \sum_{i=3}^{n} (z_i/\sigma_Y)^2 = \frac{(1-r^2)\, n\, s_Y^2}{\sigma_Y^2},$$

being the sum of squares of $(n-2)$ independent standard normal variates is an independent χ^2-variate with $(n-2)$ d.f.

Further, since z_2 and $(z_3, z_4, \ldots, z_n)$ are independent $r.v.'s$, U and V are independent chi-square variates with 1 and $(n-2)$ d.f. respectively.

$$\therefore \quad \frac{U}{U+V} = \frac{nr^2 s_Y^2/\sigma_Y^2}{[nr^2 s_Y^2 + (1-r^2)\, n\, sy^2]/\sigma_Y^2} \sim \beta_1\left(\frac{1}{2}, \frac{n-2}{2}\right)$$

[*c.f.* Theorem 13·2]

$$\Rightarrow \qquad r^2 \sim \beta_1\left(\frac{1}{2}, \frac{n-2}{2}\right)$$

Hence the probability function of r^2 is given by

$$dF(r^2) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} (r^2)^{(1/2)-1} [1-r^2]^{\frac{n-2}{2}-1} d(r^2),\ 0 \le r^2 \le 1$$

$$\Rightarrow \quad dF(r) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} [1-r^2]^{(n-4)/2}\, dr,\ -1 \le r \le 1$$

the factor 2 disappearing from the fact that total probability in the range $-1 \le r \le 1$ must be unity.

Remark. *If* $\rho = 0$, *then* $t = \frac{r}{\sqrt{(1-r^2)}}\sqrt{(n-2)}$ *is distributed as Student's t with* $(n-2)$ *d. f.*

Proof.

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \qquad \ldots(*)$$

$$\Rightarrow \qquad \frac{t^2}{n-2} = \frac{r^2}{(1-r^2)} = \frac{1}{1-r^2} - 1$$

$$\Rightarrow \qquad (1-r^2) = \left[1 + \frac{t^2}{(n-2)}\right]^{-1} \qquad \ldots(**)$$

From (*),

$$dt = \sqrt{(n-2)}\, d[r/\sqrt{(1-r^2)}\,]$$

$$dt = \sqrt{(n-2)} \left[\frac{dr}{\sqrt{(1-r^2)}} + \left(-\frac{r}{2}\right)\frac{(-2r)dr}{(1-r^2)^{3/2}}\right]$$

$$\Rightarrow \quad dt = \sqrt{(n-2)}\, \frac{dr}{\sqrt{(1-r^2)}}\left[1 + \frac{r^2}{1-r^2}\right]$$

$$\Rightarrow \quad dt = \sqrt{(n-2)} \times \frac{dr}{(1-r^2)^{3/2}},\ \text{i.e.,}\ dr = \frac{1}{\sqrt{(n-2)}}(1-r^2)^{3/2}\, dt$$

As r ranges from -1 to 1, from (*), t ranges from $-\infty$ to ∞.

When $\rho = 0$, the p.d.f. of 'r' is given by (14·12) and it transforms to

$$dG(t) = \frac{1}{B\left(\frac{1}{2}, \frac{n-2}{2}\right)} [1-r^2]^{(n-4)/2} \frac{1}{\sqrt{(n-2)}} (1-r^2)^{3/2}\, dt$$

$$= \frac{1}{\sqrt{(n-2)}\, B\left(\frac{1}{2}, \frac{n-2}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{n-2}\right]^{(n-1)/2}}$$

[From (**)]

$$= \frac{1}{\sqrt{(n-2)}\, B\left(\frac{1}{2}, \frac{n-2}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{n-2}\right]^{(n-2+1)/2}},$$

$$-\infty < t < \infty$$

which is the p.d.f. of t-distribution with $(n-2)$d.f.

Hence $t = \frac{r}{\sqrt{(1-r^2)}} \cdot \sqrt{(n-2)} \sim t_{(n-2)}$

Example 14·15. *(a) If (x_i, y_i) is a random sample drawn from an uncorrelated bivariate normal population, derive the distribution of*

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2\, \Sigma(y_i - \bar{y})^2}}$$

(b) Further, when $n = 5$ and if $P(|r| \geq C) = \alpha$, show that C is a root of the equation,

$$C\sqrt{(1-C^2)} + \sin^{-1} C + \frac{\pi(\alpha - 1)}{2} = 0$$

Solution. *(a) c.f.* § 14·3.

(b) $P(|r| \geq C) = 1 - P(|r| \leq C) = 1 - P(-C \leq r \leq C)$

$$= 1 - 2P(0 \leq r \leq C) = 1 - 2\int_0^C f(r)\, dr$$

[$\because f(r)$ is symmetrical about $r = 0$]

When $n = 5$, $f(r) = \frac{1}{B\left(\frac{1}{2}, \frac{3}{2}\right)} \cdot (1-r^2)^{\frac{1}{2}}\, dr$ [*c.f.* Equation (14·12)]

$$\therefore \quad P(|r| \geq C) = 1 - 2\frac{\Gamma(2)}{\Gamma(1/2)\Gamma(3/2)} \int_0^C (1-r^2)^{1/2}\, dr$$

$$= 1 - 2 \times \frac{1}{\frac{1}{2}\pi} \left[\tfrac{1}{2} r\,(1-r^2)^{1/2} + \tfrac{1}{2}\sin^{-1} r\right]_0^C$$

$$= 1 - \frac{4}{\pi}\left[\tfrac{1}{2} C\,(1-C^2)^{1/2} + \tfrac{1}{2}\sin^{-1} C\right] = \alpha, \text{ (Given)}$$

$$\therefore \qquad 1 - \frac{2}{\pi}\left[C(1 - C^2)^{\frac{1}{2}} + \sin^{-1} C\right] = \alpha$$

$$\Rightarrow \qquad C(1 - C^2)^{1/2} + \sin^{-1} C + (\alpha - 1)\frac{\pi}{2} = 0$$

14·4. Non-central t-distribution. The non-central t-distribution is the distribution of the ratio of a normal variate with possibly non-zero mean and variance unity, to the square root of an independent χ^2-variate divided by its degrees of freedom. If $X \sim N(\mu, 1)$ and Y is an independent χ^2-variate with n *d.f.*, then

$$t' = \frac{X}{\sqrt{Y/n}} \qquad \ldots(14\cdot13)$$

is said to have a non-central t-distribution with n *d.f.* and non-centrality parameter μ. Non-central t-distribution is required for the power functions of certain tests concerning normal population.

p.d.f. of 't''. Since $X \sim N(\mu, 1)$, its *p.d.f.* $f(.)$ is

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}(x - \mu)^2\right]$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}(\mu^2 + x^2)\right] \sum_{i=0}^{\infty} \frac{(\mu x)^i}{i\,!}, \quad -\infty < x < \infty$$

Since $Y \sim \chi^2_{(n)}$, its *p.d.f.* $g(.)$ is

$$g(y) = \frac{1}{2^{n/2}\,\Gamma(n/2)} e^{-y/2}\, y^{(n/2)-1}, \quad 0 < y < \infty$$

Since X and Y are independent, their joint p.d.f. becomes

$$f(x, y) = \frac{1}{\sqrt{2\pi}\,2^{n/2}\,\Gamma(n/2)} \exp\left[-\tfrac{1}{2}(\mu^2 + x^2 + y)\right] y^{(n/2)-1} \sum_{i=0}^{\infty} \frac{(\mu x)^i}{i\,!}$$

Let us transform to new variables t' and z by the substitution :

$$t' = \frac{x}{\sqrt{y/n}} = \frac{\sqrt{n}\,x}{\sqrt{y}}, \quad z = +\sqrt{y}$$

$$\Rightarrow \qquad x = zt'/\sqrt{n}, \quad y = z^2$$

Jacobian of transformation J is

$$J = \begin{vmatrix} \frac{\partial x}{\partial t'} & \frac{\partial x}{\partial z} \\ \frac{\partial y}{\partial t'} & \frac{\partial y}{\partial z} \end{vmatrix} = \begin{vmatrix} \frac{z}{\sqrt{n}} & \frac{t'}{\sqrt{n}} \\ 0 & 2z \end{vmatrix} = \frac{2z^2}{\sqrt{n}}$$

The joint p.d.f. of t' and z becomes

$$h(t', z) = \frac{\exp(-\mu^2/2)}{\sqrt{2\pi}\,2^{n/2}\,\Gamma(n/2)} (z^2)^{\frac{n}{2}-1} \sum_{i=0}^{\infty} \frac{(\mu z t'/\sqrt{n})^i}{i\,!} \cdot \frac{2z^2}{\sqrt{n}};$$

$$\times \exp\left[-\frac{1}{2}\left(1+\frac{t'^2}{n}\right)z^2\right]; \quad -\infty < t' < \infty,\ 0 < z < \infty$$

$$= \frac{\exp(-\mu^2/2)}{\sqrt{\pi}\, 2^{(n-1)/2}\,\Gamma(n/2)} \sum_{i=0}^{\infty}\left[\frac{(\mu t')^i}{i!\, n^{(i+1)/2}} \cdot \exp\left\{-\frac{1}{2}\left(1+\frac{t'^2}{n}\right)z^2\right\} z^{n+i}\right]$$

Integrating w.r.t. z in the range 0 to ∞, we get the p.d.f. of t'

$$h_1(t') = \frac{\exp(-\mu^2/2)}{\sqrt{\pi}\, 2^{(n-1)/2}\,\Gamma(n/2)} \times \sum_{i=0}^{\infty}\left[\frac{(\mu t')^i}{i!\, n^{(i+1)/2}} \int_0^{\infty} \exp\left\{-\frac{1}{2}\left(1+\frac{t'^2}{n}\right)z^2\right\} z^{n+i}\, dz\right]$$

$$= \frac{\exp(-\mu^2/2)}{\sqrt{\pi} 2^{(n-1)/2}\,\Gamma(n/2)} \times \sum_{i=0}^{\infty}\left[\frac{(\mu t')^i}{i!\, n^{(i+1)/2}} \int_0^{\infty} \exp\left\{-\left(1+\frac{t'^2}{n}\right)v\right\} \cdot (2v)^{(n+i-1)/2}\, dv\right]$$

$$= \frac{\exp(-\mu^2/2)}{\sqrt{\pi}\,\Gamma(n/2)} \sum_{i=0}^{\infty}\left[\frac{\mu^i 2^{i/2}\,\Gamma\left(\frac{n+i+1}{2}\right)}{i!\, n^{(i+1)/2}} \cdot \frac{t'^i}{\left(1+\frac{t'^2}{n}\right)^{(n+i+1)/2}}\right]$$

...[14·13(*a*)]

which is the p.d.f. of non-central t-distribution with n d.f. and non-centrality element μ.

Remark. If $\mu = 0$, we get from [14·13 (*a*)]

$$h_1(t') = \frac{1}{\sqrt{\pi}\,\Gamma(n/2)} \cdot \frac{\Gamma[(n+1)/2]}{\sqrt{n}} \cdot \frac{1}{\left[1+\frac{t'^2}{n}\right]^{(n+1)/2}}$$

$$= \frac{1}{\sqrt{n}\, B\left(\frac{1}{2}, \frac{n}{2}\right)}\left[1+\frac{t'^2}{n}\right]^{-(n+1)/2}, \quad -\infty < t' < \infty$$

which is the p.d.f. of central t-distribution with n d.f.

14·5. F-statistic. Definition. *If X and Y are two independent chi-square variates with ν_1 and ν_2 d.f. respectively, then F-statistic is defined by*

$$F = \frac{X/\nu_1}{Y/\nu_2} \qquad ...(14·14)$$

In other words, F is defined as the ratio of two independent chi-square variates divided by their respective degrees of freedom and it follows Snedecor's F-distribution with (ν_1, ν_2) d.f. with probability function given by

$$f(F) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{F^{\frac{\nu_1}{2}-1}}{\left[1 + \frac{\nu_1}{\nu_2}F\right]^{(\nu_1+\nu_2)/2}}, \quad 0 \le F < \infty \qquad \ldots[14\cdot14(a)$$

Remarks 1. The sampling distribution of F-statistic does not involve any population parameters and depends only on the degrees of freedom ν_1 and ν_2.

2. A statistic F following Snedecor's F-distribution with (ν_1, ν_2) $d.f.$ will be denoted by $F \sim F(\nu_1, \nu_2)$.

14·5·1 Derivation of Snedecor's F-distribution. Since X and Y are independent chi-square variates with ν_1 and ν_2 d.f. respectively, their joint probability differential is given by

$$dF(x, y) = \left\{\frac{1}{2^{\nu_1/2}\,\Gamma(\nu_1/2)} \exp(-x/2)\, x^{(\nu_1/2)-1}\, dx\right\} \times \left\{\frac{1}{2^{\nu_2/2}\,\Gamma(\nu_2/2)} \exp(-y/2)\, y^{(\nu_2/2)-1}\, dy\right\}$$

$$= \frac{1}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \exp\{-(x+y)/2\} \times x^{(\nu_1/2)-1}\, y^{(\nu_2/2)-1}\, dx\, dy, \quad 0 \le (x, y) < \infty$$

Let us make the following transformation of variables :

$$F = \frac{x/\nu_1}{y/\nu_2} \text{ and } u = y, \text{ so that } 0 \le F < \infty, 0 < u < \infty$$

$$\therefore \quad x = \frac{\nu_1}{\nu_2} Fu = \frac{\nu_1}{\nu_2} Fu \quad \text{and} \quad y = u$$

Jacobian of transformation J is given by

$$J = \frac{\partial(x, y)}{\partial(F, u)} = \begin{vmatrix} \frac{\nu_1}{\nu_2} u & 0 \\ \frac{\nu_1}{\nu_2} F & 1 \end{vmatrix} = \frac{\nu_1 u}{\nu_2}$$

Thus the distribution of the transformed variable is

$$dG(F, u) = \frac{1}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \exp\left\{-\frac{u}{2}\left(1 + \frac{\nu_1}{\nu_2}F\right)\right\} \times \left(\frac{\nu_1}{\nu_2} Fu\right)^{(\nu_1/2)-1} u^{(\nu_2/2)-1} |J|\, du\, dF$$

$$= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \exp\left\{-\frac{u}{2}\left(1 + \frac{\nu_1}{\nu_2}F\right)\right\} \times u^{\{(\nu_1+\nu_2)/2\}-1}\, F^{(\nu_1/2)-1}\, du\, dF; \quad 0 < u < \infty, 0 \le F < \infty$$

Integrating out u over the range 0 to ∞, the distribution of F becomes

$$g_1(F)\,dF = \frac{(\nu_1/\nu_2)^{(\nu_1/2)}\,F^{(\nu_1/2)-1}\,dF}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \times \left[\int_0^\infty \exp\left\{-\frac{u}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right\} u^{\{(\nu_1+\nu_2)/2\}-1}\,du\right]$$

$$= \frac{(\nu_1/\nu_2)^{(\nu_1/2)}\,F^{(\nu_1/2)-1}}{2^{(\nu_1+\nu_2)/2}\,\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \times \frac{\Gamma[(\nu_1+\nu_2)/2]}{\left[\frac{1}{2}\left(1+\frac{\nu_1}{\nu_2}F\right)\right]^{(\nu_1+\nu_2)/2}}\,dF$$

$$\therefore \quad g_1(F) = \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)} \cdot \frac{F^{(\nu_1/2)-1}}{\left[1+\frac{\nu_1}{\nu_2}F\right]^{(\nu_1+\nu_2)/2}},\ 0 \le F < \infty$$

which is the required probability function of F-distribution with (ν_1, ν_2) d.f.

Aliter

$$F = \frac{x/\nu_1}{y/\nu_2}$$

$\therefore \frac{\nu_1}{\nu_2}F = \frac{x}{y}$, being the ratio of two independent chi-square variates with ν_1 and ν_2 d.f. respectively is a $\beta_2\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)$ variate. Hence the probability function of F is given by

$$dP(F) = \frac{1}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)} \cdot \frac{\left(\frac{\nu_1}{\nu_2}F\right)^{(\nu_1/2)-1}}{\left[1+\frac{\nu_1}{\nu_2}F\right]^{(\nu_1+\nu_2)/2}}\,d\left(\frac{\nu_1}{\nu_2}F\right)$$

$$= \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)} \cdot \frac{F^{(\nu_1/2)-1}}{\left[1+\frac{\nu_1}{\nu_2}F\right]^{(\nu_1+\nu_2)/2}}\,dF,\ 0 \le F < \infty$$

14·5·2. Constants of F-distribution.

$$\mu'_r \text{ (about origin)} = E(F^r) = \int_0^\infty F^r f(F)\,dF$$

$$= \frac{(\nu_1/\nu_2)^{\nu_1/2}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)} \int_0^\infty F^r \frac{F^{(\nu_1/2)-1}}{\left[1+\frac{\nu_1}{\nu_2}F\right]^{(\nu_1+\nu_2)/2}}\,dF \qquad \ldots(*)$$

To evaluate the integral, put

$$\frac{\nu_1}{\nu_2}F = y \text{, so that } dF = \frac{\nu_2}{\nu_1}dy$$

$$\therefore \quad \mu_r' = \frac{[\nu_1/\nu_2]^{\nu_1/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty \frac{\left(\frac{\nu_2}{\nu_1} y\right)^{r + (\nu_1/2) - 1}}{[1 + y]^{(\nu_1 + \nu_2)/2}} \left(\frac{\nu_2}{\nu_1}\right) dy$$

$$= \frac{\left(\frac{\nu_2}{\nu_1}\right)^r}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \int_0^\infty \frac{y^{r + (\nu_1/2) - 1}}{[1 + y]^{(\nu_1/2) + r + \{(\nu_2/2) - r\}}} dy$$

$$= \left(\frac{\nu_2}{\nu_1}\right)^r \cdot \frac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot B\left(r + \frac{\nu_1}{2}, \frac{\nu_2}{2} - r\right), \nu_2 > 2r \qquad \ldots(14\cdot15)$$

Aliter for (14·15). (14·15) could also be obtained by substituting $\frac{\nu_1}{\nu_2} F = \tan^2\theta$ in (*) and using the Beta integral :

$$2\int_0^{\pi/2} \sin^p\theta \cos^q\theta \, d\theta = B\left(\frac{p+1}{2}, \frac{q+1}{2}\right)$$

$$\therefore \quad \mu_r' = \left(\frac{\nu_2}{\nu_1}\right)^r \cdot \frac{\Gamma[r + (\nu_1/2)]\,\Gamma[(\nu_2/2) - r]}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)}, \; r < \frac{\nu_2}{2}. \qquad \ldots(14\cdot16)$$

In particular

$$\mu_1' = \frac{\nu_2}{\nu_1} \cdot \frac{\Gamma[1 + (\nu_1/2)]\,\Gamma[(\nu_2/2) - 1]}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)}$$

$$= \frac{\nu_2}{\nu_2 - 2}, \nu_2 > 2 \qquad [\because \Gamma(r) = (r-1)\,\Gamma(r-1)] \qquad \ldots[14\cdot16(a)]$$

Thus the mean of F-distribution is independent of ν_1.

$$\mu_2' = \left(\frac{\nu_2}{\nu_1}\right)^2 \cdot \frac{\Gamma[(\nu_1/2) + 2]\,\Gamma[(\nu_2/2) - 2]}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)}$$

$$= \left(\frac{\nu_2}{\nu_1}\right)^2 \cdot \frac{[(\nu_1/2) + 1]\,(\nu_1/2)}{[(\nu_2/2) - 1]\,[(\nu_2/2) - 2]}$$

$$= \frac{\nu_2^2(\nu_1 + 2)}{\nu_1(\nu_2 - 2)\,(\nu_2 - 4)}, \nu_2 > 4.$$

$$\therefore \quad \mu_2 = \mu_2' - \mu_1'^2 = \frac{\nu_2^2(\nu_1 + 2)}{\nu_1(\nu_2 - 2)(\nu_2 - 4)} - \frac{\nu_2^2}{(\nu_2 - 2)^2}$$

$$= \frac{2\nu_2^2\,(\nu_2 + \nu_1 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}, \nu_2 > 4 \qquad \ldots[14\cdot16(b)]$$

Similarly, on putting $r = 3$ and 4 in μ_r', we get μ_3' and μ_4' respectively, from which the central moments μ_3 and μ_4 can be obtained.

Remark. It has been proved that for large degrees of freedom, ν_1 and ν_2, F tends to $N[1, 2\{(1/\nu_1) + (1/\nu_2)\}]$ variate.

14·5·3. Mode and Points of Inflexion of F-distribution. We have

$$\log f(F) = C + \{(\nu_1/2) - 1\} \log F - \left(\frac{\nu_1 + \nu_2}{2}\right) \log \{1 + (\nu_1/\nu_2) F\}$$

where C is a constant independent of F.

$$\frac{\partial}{\partial F} [\log f(F)] = \left(\frac{\nu_1}{2} - 1\right) \cdot \frac{1}{F} - \frac{(\nu_1 + \nu_2)}{2} \cdot \frac{1}{\left[1 + \frac{\nu_1}{\nu_2} F\right]} \cdot \frac{\nu_1}{\nu_2}$$

$$f'(F) = \frac{\partial}{\partial F} f(F) = 0 \quad \Rightarrow \quad \frac{\nu_1 - 2}{2F} - \frac{\nu_1 (\nu_1 + \nu_2)}{2(\nu_2 + \nu_1 F)} = 0$$

Hence
$$F = \frac{\nu_2 (\nu_1 - 2)}{\nu_1 (\nu_2 + 2)} \qquad \ldots(14\cdot17)$$

It can be easily verified that at this point $f''(F) < 0$. Hence

$$\text{Mode} = \frac{\nu_2 (\nu_1 - 2)}{\nu_1 (\nu_2 + 2)}$$

Remarks 1. Since $F > 0$, mode exists if and only if $\nu_1 > 2$.

2.
$$\text{Mode} = \left(\frac{\nu_2}{\nu_2 + 2}\right) \cdot \left(\frac{\nu_1 - 2}{\nu_1}\right)$$

Hence mode of F-distribution is always less than unity.

3. The points of inflexion of F-distribution exist for $\nu_1 > 4$ and are equidistant from mode.

Proof. We have $\frac{\nu_1}{\nu_2} F = \frac{X}{Y} \sim \beta_2 (l, m)$, (*)

where $l = \nu_1/2$ and $m = \nu_2/2$. We now find the points of inflexion of Beta distribution of second kind with parameters l and m.

If $X \sim \beta_2 (l, m)$, its *p.d.f.* is

$$f(x) = \frac{1}{\beta(l, m)} \cdot \frac{x^{l-1}}{(1 + x)^{l+m}}; \; 0 \le x < \infty \qquad \ldots(**)$$

Points of inflexion are the solution of

$$f''(x) = 0 \text{ and } f'''(x) \neq 0$$

From (**),

$$\log f(x) = -\log \beta(l, m) + (l - 1) \log x - (l + m) \log (1 + x)$$

Differentiating twice w.r.t x, we get

$$\frac{f'(x)}{f(x)} = \frac{l - 1}{x} - \frac{l + m}{1 + x} \qquad \ldots(***)$$

$$\frac{f(x) f''(x) - [f'(x)]^2}{[f(x)]^2} = -\left(\frac{l - 1}{x^2}\right) + \frac{l + m}{(1 + x)^2}$$

If $f''(x) = 0$, then we get

$$-\left[\frac{f'(x)}{f(x)}\right]^2 = -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2}$$

$$\Rightarrow \quad -\left[\frac{l-1}{x} - \frac{l+m}{1+x}\right]^2 = -\left(\frac{l-1}{x^2}\right) + \frac{l+m}{(1+x)^2} \qquad \text{[On using (***)]}$$

$$\Rightarrow \quad \frac{l-1}{x^2}(l-1-1) - 2\frac{(l-1)(l+m)}{x(1+x)} + \frac{l+m}{(1+x)^2} \times (l+m+1) = 0$$

$$\Rightarrow (l-1)(l-2)(1+x)^2 - 2x(1+x)(l-1)(l+m) + x^2(l+m)(l+m+1) = 0 \qquad \ldots(****)$$

which is a quadratic in x. It can be easily verified that at these values of x, $f'''(x) \neq 0$, if $l > 2$.

The roots of (****) give the points of inflexion of $\beta_2(l, m)$ distribution. The sum of the points of inflexion is equal to the sum of roots of (****) and is given by :

$$-\left[\frac{\text{Coefficient of } x \text{ in (****)}}{\text{Coefficient of } x^2 \text{ in (****)}}\right]$$

$$= -\left[\frac{2(l-1)(l-2) - 2(l-1)(l+m)}{(l-1)(l-2) - 2(l-1)(l+m) + (l+m)(l+m+1)}\right]$$

$$= \frac{2(l-1)[(l+m) - (l-2)]}{(l-1)(l-2) - (l-1)(l+m) - (l-1)(l+m) + (l+m)(l+m+1)}$$

$$= \frac{2(l-1)(m+2)}{(l-1)[(l-2-l-m] + (l+m)[l+m+1-l+1)}$$

$$= \frac{2(l-1)(m+2)}{-(l-1)(m+2) + (l+m)(m+2)}$$

$$= \frac{2(l-1)}{l+m-l+1} = \frac{2(l-1)}{(m+1)}$$

$\therefore$ Sum of points of inflexion of $\left(\frac{\nu_1}{\nu_2}F\right)$ distribution

$$= \frac{2(l-1)}{(m+1)} = \frac{2\left(\frac{\nu_1}{2} - 1\right)}{\left(\frac{\nu_2}{2} + 1\right)} = \frac{2(\nu_1 - 2)}{(\nu_2 + 2)}$$

$\Rightarrow$ Sum of points of inflexion of $F(\nu_1, \nu_2)$ distribution

$$= \frac{\nu_2}{\nu_1} \cdot \frac{2(\nu_1 - 2)}{(\nu_2 + 2)}, \text{ provided } l = \frac{\nu_1}{2} > 2$$

$$= 2\frac{\nu_2(\nu_1 - 2)}{\nu_1(\nu_2 + 2)}$$

$$= 2 \text{ Mode, provided } \nu_1 > 4$$

Hence the points of inflexion of $F(\nu_1, \nu_2)$ distribution, when they exist, (*i.e.*, when $\nu_1 > 4$), are equidistant from the mode.

4. Karl Pearson's coefficient of skewness is given by

$$S_k = \frac{\text{Mean} - \text{Mode}}{\sigma} > 0,$$

since mean > 1 and mode < 1. Hence F-distribution is highly positively skewed.

5. The probability $p(F)$ increases steadily at first until it reaches its peak (corresponding to the modal value which is less than 1) and then decreases slowly so as to become tangential at $F = \infty$, *i.e.*, F-axis is an asymptote to the right tail.

Example 14·16. *When $v_1 = 2$, show that the significance level of F corresponding to a significant probability p is*

$$F = \frac{v_2}{2}\left(p^{-(2/v_2)} - 1\right)$$

where v_1 an v_2 have their usual meanings.

Solution. When $v_1 = 2$,

$$dP(F) = \frac{1}{B\left(1, \frac{v_2}{2}\right)} \cdot \frac{2}{v_2} \cdot \frac{dF}{\left[1 + \frac{2}{v_2}F\right]^{(v_2/2)+1}} \qquad \text{(c.f. §14·14}a\text{)}$$

$$= \frac{\Gamma\left(\frac{v_2}{2} + 1\right)}{\Gamma(1)\Gamma(v_2/2)} \times \frac{\frac{2}{v_2}}{\left(\frac{2}{v_2}\right)^{(v_2/2)+1}\left[F + \frac{v_2}{2}\right]^{(v_2/2)+1}} dF$$

$$= \frac{\left[\frac{v_2}{2}\right]^{(v_2/2)+1}}{\left[F + \frac{v_2}{2}\right]^{(v_2/2)+1}} dF.$$

Hence $$p = \int_F^\infty f(F)\, dF$$

$$= \left[\frac{v_2}{2}\right]^{(v_2/2)+1} \times \int_F^\infty \frac{dF}{\left[F + \frac{v_2}{2}\right]^{(v_2/2)+1}}$$

$$= \left[\frac{v_2}{2}\right]^{(v_2/2)+1} \times \left| \frac{\left[F + \frac{v_2}{2}\right]^{-(v_2/2)}}{-\frac{v_2}{2}} \right|_F^\infty$$

$$= \left[\frac{\left(\frac{v_2}{2}\right)}{F + \frac{v_2}{2}}\right]^{v_2/2} = \frac{1}{\left[1 + \frac{2}{v_2}F\right]^{v_2/2}}$$

$$\Rightarrow \quad p^{-(2/\nu_2)} = 1 + \frac{2F}{\nu_2} \quad \Rightarrow \quad F = \frac{\nu_2}{2}\left[p^{-(2/\nu_2)} - 1\right]$$

Example 14·17. *If $F(n_1, n_2)$ represent an F-variate with n_1 and n_2 d.f., prove that $F(n_2, n_1)$ is distributed as $1/F$ (n_1, n_2) variate. Deduce that*

$$P[F(n_1, n_2) \geq c] = P\left[F(n_2, n_1) \leq \frac{1}{c}\right]$$

Or

Show how the probability points of $F(n_2, n_1)$ can be obtained from those of $F(n_1, n_2)$.

Solution. Let X and Y be independent chi-square variates with n_1 and n_2 d.f. respectively. Then by definition, we have

$$F = \frac{(X/n_1)}{(Y/n_2)} \sim F(n_1, n_2)$$

$$\therefore \quad \frac{1}{F} = \frac{(Y/n_2)}{(X/n_1)} \sim F(n_2, n_1) \qquad \ldots(*)$$

Hence the result.

We have :

$$P[F(n_1, n_2) \geq c] = P\left[\frac{1}{F(n_1, n_2)} \leq \frac{1}{c}\right]$$

$$= P\left[F(n_2, n_1) \leq \frac{1}{c}\right] \qquad \text{[From (*)]}$$

Remark. $P[F(n_1, n_2) = c] = P\left[F(n_2, n_1) = \frac{1}{c}\right]$

Let $P[F(n_1, n_2) \geq c] = \alpha$

i.e., let c be the upper α-significant point of $F(n_1, n_2)$ distribution.

$$\therefore \quad 1 - \alpha = 1 - P[F(n_1, n_2) \geq c] = 1 - P\left[\frac{1}{F(n_1, n_2)} \leq \frac{1}{c}\right]$$

$$\Rightarrow \quad \alpha = P\left[F(n_2, n_1) \leq \frac{1}{c}\right] = 1 - P\left[F(n_2, n_1) \geq \frac{1}{c}\right]$$

$$\Rightarrow \quad P\left[F(n_2, n_1) \geq \frac{1}{c}\right] = 1 - \alpha$$

Thus $(1 - \alpha)$ significant points of $F(n_2, n_1)$ distribution are the reciprocal of α-significant points of $F(n_1, n_2)$ distribution, *e.g.*,

$$F_{8,4}(0{\cdot}05) = 6{\cdot}04 \Rightarrow F_{4,8}(0{\cdot}95) = \frac{1}{6{\cdot}04}$$

Example 14·18. *Prove that if $n_1 = n_2$, the median of F-distribution is at $F = 1$ and that the quartiles Q_1 and Q_3 satisfy the condition $Q_1 Q_3 = 1$.*

[*Delhi Univ. B.Sc. (Stat.Hons.), 1989*]

Solution. Since $n_1 = n_2 = n$, (say), the median (M) of $F(n_1, n_2) = F(n, n)$ distribution is given by :

$$P[F(n, n) \le M] = 0{\cdot}5 \qquad \ldots(*)$$

$$\Rightarrow \quad P\left[\frac{1}{F(n, n)} \ge \frac{1}{M}\right] = 0{\cdot}5$$

$$\Rightarrow \quad P\left[F(n, n) \ge \frac{1}{M}\right] = 0{\cdot}5 \qquad \left[\because \frac{1}{F(m, n)} = F(n, m)\right]$$

$$\Rightarrow \quad P\left[F(n, n) \le \frac{1}{M}\right] = 1 - P\left[F(n, n) \ge \frac{1}{M}\right]$$

$$= 1 - 0{\cdot}5$$

$$= 0{\cdot}5 \qquad \ldots(**)$$

From (*) and (**), we get

$$M = \frac{1}{M} \quad \Rightarrow \quad M^2 = 1 \quad \Rightarrow \quad M = 1$$

the negative value $M = -1$, is discarded since $F > 0$.

Hence the median of $F(n, n)$ distribution is at $F = 1$.

Similarly, by definition of Q_1 and Q_3, we have :

$$P[F(n, n) \le Q_1] = 0{\cdot}25 \qquad \ldots(****)$$

and $$P[F(n, n) \ge Q_3] = 0{\cdot}25$$

$$\Rightarrow \quad P\left[\frac{1}{F(n, n)} \le \frac{1}{Q_3}\right] = 0{\cdot}25$$

$$\Rightarrow \quad P\left[F(n, n) \le \frac{1}{Q_3}\right] = 0{\cdot}25 \quad \left[\because \frac{1}{F(m, n)} = F(n, m)\right] \qquad \ldots(***)$$

From (***) and (****), we get

$$Q_1 = \frac{1}{Q_3} \quad \Rightarrow \quad Q_1 Q_3 = 1$$

Example 14·19. *Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(0, 1)$.*

Define $$\bar{X}_k = \frac{1}{k}\sum_1^k X_i \quad \text{and} \quad \bar{X}_{n-k} = \frac{1}{n-k}\sum_{k+1}^{n} X_i$$

Find the distribution of :

(*a*) $\frac{1}{2}\left(\bar{X}_k + \bar{X}_{n-k}\right)$, (*b*) $k\bar{X}_k^2 + (n-k)\bar{X}_{n-k}^2$

(*c*) X_1^2/X_2^2, (*d*) X_1/X_2

[***Delhi Univ. B.A, (Stat. Hons. Spl. Course), 1989***]

Solution. (*a*) Since $X_1, X_2, \ldots, X_n$ is a random sample from $N(0, 1)$,

$$\bar{X}_k \sim N\left(0, \frac{1}{k}\right) \quad \text{and} \quad \bar{X}_{n-k} \sim N\left(0, \frac{1}{n-k}\right) \qquad \ldots(*)$$

Further, since $(X_1, X_2, \ldots, X_k)$ and $(X_{k+1}, X_{k+2}, \ldots, X_n)$ are independent, $\bar{X}_k$ and $\bar{X}_{n-k}$ are independent. Hence,

$$\frac{1}{2}(\bar{X}_k + \bar{X}_{n-k}) = \frac{1}{2}\bar{X}_k + \frac{1}{2}\bar{X}_{n-k} \sim N\left(0, \frac{1}{4k} + \frac{1}{4(n-k)}\right)$$

$$\Rightarrow \qquad \frac{1}{2}(\bar{X}_k + \bar{X}_{n-k}) \sim N\left(0, \frac{n}{4k\,(n-k)}\right)$$

(*b*) From (*), we get

$$\frac{\bar{X}_k}{\sqrt{1/k}} \sim N(0, 1) \quad \text{and} \quad \frac{\bar{X}_{n-k}}{\sqrt{1/(n-k)}} \sim N(0, 1)$$

$$\Rightarrow \qquad k\,\bar{X}_k^2 \sim \chi^2_{(1)} \qquad \text{and} \quad (n-k)\,\bar{X}^2_{n-k} \sim \chi^2_{(1)}$$

Since $\bar{X}_k$ and $\bar{X}_{n-k}$ are independent, by additive property of chi-square distribution,

$$k\,\bar{X}_k^2 + (n-k)\,\bar{X}^2_{n-k} \sim \chi^2_{(1+1)} = \chi^2_{(2)}$$

(*c*) Since $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 1)$ are independent,

$$X_1^2 \sim \chi^2_{(1)} \text{ and } X_2^2 \sim \chi^2_{(1)},$$

are also independent. Hence by definition of F-statistic,

$$\frac{X_1^2/1}{X_2^2/1} \sim F_{(1,1)} \Rightarrow \frac{X_1^2}{X_2^2} \sim F_{(1,1)}$$

(*d*) X_1/X_2, being the ratio of two independent standard normal variates is a standard Cauchy variate. [See Example 8·43].

EXERCISE 14(e)

1. (*a*) Derive the distribution of $F = S_1^2/S_2^2$, where S_1^2 and S_2^2 are two independent unbiased estimates of the common population variance σ^2, defined by

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2; \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2$$

(*b*) Find the limiting form when the degrees of freedom of the χ^2 in the denominator tend to infinity and give an intuitive justification of the result.

2. (*a*) If $X_1, X_2, \ldots, X_m, X_{m+1}, \ldots\ldots, X_{m+n}$ are independent normal variates with zero mean and standard deviation σ, obtain the distribution of

$$\sum_{i=1}^{m} X_i^2 \Big/ \sum_{i=m+1}^{m+n} X_i^2$$

Ans. $F(m, n)$.

(*b*) If X has an F distribution with n_1 and n_2 d.f., find the distribution of $1/X$ and give one use of this result.

(*c*) If X is t-distributed, show that X^2 is F-distributed.

[***Delhi Univ. B.Sc. (Maths. Hons.), 1990***]

Hint. See § 14·5·6.

3. (*a*) Derive the distribution of the F-statistic on (n_1, n_2) degrees of freedom and show that the statistic $\left(1 + \frac{n_1}{n_2} F\right)^{-1}$ has a Beta distribution.

(*b*) Show that the probability curve of the distribution of F is positively skewed.

4. Prove the following :

(*i*) $$F_{n_1, n_2} = \frac{1}{F_{n_2, n_1}}$$

(*ii*) $$F_{n_1, n_2} = \frac{n_2}{n_1} \cdot \frac{x}{1 - x},$$ where x has Beta-distribution.

5. If X and Y are independent chi-square variates with ν_1 and ν_2 d.f. respectively, show that $U = X + Y$ and $V = \frac{\nu_2 X}{\nu_1 Y}$ are independently distributed. Find the distribution of V.

6. Prove that if X has the F-distribution with (m, n) d.f. and Y has the F-distribution with (n, m) d.f., then for every $a > 0$,

$$P(X \leq a) + P\left\{Y \leq \frac{1}{a}\right\} = 1$$

7. Show that the mode of the F-distribution with ν_1 (≥ 2), ν_2 d.f. is given by $\frac{\nu_2 (\nu_1 - 2)}{\nu_1 (\nu_2 + 2)}$ and is always less than unity.

8. X is F-variate with 2 and n ($n \geq 2$) degrees of freedom. Show that

$$P(F \geq k) = \left(1 + \frac{2k}{n}\right)^{-n/2}$$

[Gujarat Univ. B.Sc., 1992]

Deduce the significance level of F corresponding to the significance level of probability P.

9. Let X_1, X_2 be independent random variables following the density law $f(x) = e^{-x}$, $0 < x < \infty$. Show that

$Z = X_1/X_2$, has an F-distribution.

10. (*a*) If $X \sim F(n_1, n_2)$, show that its mean is independent of n_1.

(*b*) Obtain the mode of F-distribution with (n_1, n_2) d.f. and show that it lies between 0 and 1.

(*c*) Show that for F-distribution with n_1 and n_2 d.f., the points of inflexion exist if $n_1 > 4$ and are equidistant from the mode.

11. X is a binomial variate with parameters n and p and F_{ν_1, ν_2} is an F-statistic with ν_1 and ν_2 d.f. Prove that

$$P(X \leq k - 1) = P\left[F_{2k,\, 2(n - k + 1)} > \frac{n - k + 1}{k} \cdot \frac{p}{1 - p}\right]$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1985]

Hint. If $X \sim B(n, p)$, then we have [c.f. Example 7·23]

$$P(X \leq k-1) = (n-k+1).\binom{n}{k-1}\int_0^q t^{n-k}(1-t)^{k-1}\,dt$$

$$= \frac{1}{B(k,\ n-k+1)}\int_0^q t^{n-k}(1-t)^{k-1}\,dt$$

$$P\left[F_{2k,2(n-k+1)} > \frac{n-k+1}{k}\left(\frac{p}{1-p}\right)\right] = \int_{\frac{n-k+1}{k}\cdot\frac{p}{q}}^{\infty} p[F_{2k,2(n-k+1)}]\,dF$$

$$= \frac{1}{B(k,\ n-k+1)}\int_{\frac{n-k+1}{k}\cdot\frac{p}{q}}^{\infty} \frac{[k/(n-k+1)]^k \cdot F^{k-1}\,dF}{\left[1+\frac{kF}{n-k+1}\right]^{n+1}}$$

$$= \frac{1}{B(k,\ n-k+1)}\int_0^q y^{n-k}(1-y)^{k-1}\,dy$$

where $1+\frac{kF}{n-k+1} = \frac{1}{y}$.

12. (*a*) If $X \sim F(n_1, n_2)$ distribution, show that

$$U = \frac{n_1 X}{n_2 + n_1 X} \sim \beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$$

[Delhi Univ. B.Sc. (Maths. Hons.), 1992]

Hence obtain the distribution function of X.

Hint. The distribution function of $X \sim F(n_1, n_2)$ is given by

$$G_X(x) = \int_0^x f(F)\,dF = \int_0^y h(u)du, \quad \left[y = \frac{n_1 x}{n_2 + n_1 x}\right]$$

$$= \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)}\int_0^y u^{\frac{n_1}{2}-1}(1-u)^{\frac{n_2}{2}-1}\,du$$

$$= I_y\left(\frac{n_1}{2}, \frac{n_2}{2}\right),$$

where $I_x(p, q) = \frac{1}{B(p, q)}\int_0^x t^{p-1}(1-t)^{q-1}\,dt,$

is the incomplete Beta function. Hence the distribution function of F distribution can be obtained from the tables of incomplete Beta function.

(*b*) $X \sim F(m, n)$, show that

$$W = \frac{m X/n}{1 + (m X/n)} \sim \beta_1\left(\tfrac{1}{2}m, \tfrac{1}{2}n\right)$$

Deduce the variance of X from p.d.f. of W.

[Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1989]

13. Let X_1 and X_2 be a random sample of size 2 form $N(0, 1)$ and Y_1 and Y_2 be a random sample of size 2 from $N(1, 1)$, and let the Y_i's be independent of the X_i's. Find the distribution of the following :

(*i*) $(X_1 - X_2)/\sqrt{2}$ (*ii*) $(X_1 + X_2)^2/(X_2 - X_1)^2$

(*iii*) $\bar{X} + \bar{Y}$ (*iv*) $(Y_1 + Y_2 - 2)^2/(X_2 - X_1)^2$

(*v*) $(X_1 + X_2)/\sqrt{[(X_2 - X_1)^2 + (Y_2 - Y_1)^2]/2}$

(*vi*) $[(Y_1 - Y_2)^2 + (X_1 - X_2)^2 + (X_1 + X_2)^2]/2$

[*Delhi Univ. B.Sc. (Maths. Hons.), 1988, 1987*]

Ans. (*i*) $N(0, 1)$, (*ii*) $F(1, 1)$, (*iii*) $N(1, 1)$

(*iv*) $F(1, 1)$, (*v*) $t_{(2)}$, (*vi*) $\chi^2_{(3)}$.

14. Let $X_i \sim N(i, i^2)$, $i = 1, 2, 3$ be independent random variables. Using only the three random variables X_1, X_2, and X_3, give an example of a statistic that has :

(*i*) A chi-square distribution with 3 d.f.

(*ii*) An F-distribution with (1, 2) d.f.

(*iii*) A t-distribution with 2 d.f.

[*Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1986*]

Ans. Hint. $Z_i = (X_i - i)/i$, $i = 1, 2, 3$, are i.i.d. $N(0, 1)$.

(*i*) $\sum_{i=1}^{3} Z_i^2 \sim \chi^2_{(3)}$; (*ii*) $\frac{2Z_1^2}{Z_2^2 + Z_3^2} \sim F(1, 2)$; (*iii*) $\frac{Z_1}{[(Z_2^2 + Z_3^2)/2]^{1/2}} \sim t_{(2)}$

15. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Define :

$$\bar{X}_k = \frac{1}{k}\sum_{1}^{k} X_i\,,\quad \bar{X}_{n-k} = \frac{1}{n-k}\sum_{k+1}^{n} X_i\,,\quad \bar{X} = \frac{1}{n}\sum_{1}^{n} X_i$$

$$S_k^2 = \frac{1}{k-1}\sum_{1}^{k} X_i - \bar{X}_k)^2\,,\quad S^2_{n-k} = \frac{1}{n-k-1}\sum_{k+1}^{n} (X_i - \bar{X}_{n-k})^2$$

and $$S^2 = \frac{1}{n-1}\sum_{1}^{n} (X_i - \bar{X})^2.$$

Answer the following questions :

(*i*) What is the distribution of

$$\sigma^{-2}[(k-1)S^2_k + (n-k-1)S^2_{n-k}]\ ?$$

(*ii*) What is the distribution of S^2_k/S^2_{n-k} ?

(*iii*) What is the distribution of $(\bar{X} - \mu)\sqrt{n}/S$?

[*Delhi Univ. B.Sc. (Maths Hons.), 1989*]

(*iv*) What is the distribution of $\frac{1}{2}(\bar{X}_k + \bar{X}_{n-k})$?

(*v*) What is the distribution of $(X_i - \mu)^2/\sigma^2$?

Ans. (*i*) $\chi^2_{(k-1)+(n-k-1)} = \chi^2_{(n-2)}$; (*ii*) $F_{(k-1, n-k-1)}$

(*iii*) $t_{(n-1)}$; (*iv*) $N\left(\mu, \frac{n\sigma^2}{4k(n-k)}\right)$, (*v*) $\chi^2_{(1)}$

16. If $X \sim F(1, n)$, show that

$$\left(n - \frac{1}{2}\right) \log\ [1 + (X/n)] \sim \chi^2_{(1)},$$

for large n.

17. If X_1, X_2, X_3 and X_4 are independent observations from N (0, 1) population, state giving reasons, the sampling distributions of

(*i*) $U = \dfrac{\sqrt{2}\, X_3}{\sqrt{X_1^2 + X_2^2}}$ and (*ii*) $V = \dfrac{3X_4^2}{X_1^2 + X_2^2 + X_3^2}$

Ans. (*i*) $U \sim t_{(2)}$; (*ii*) $V \sim F(1, 3)$.

18. Let (X_1, X_2) be a random sample from $N(0, 1)$. Answer the following, giving reasons :

(*i*) What is the distribution of $(X_2 - X_1)^2/2$?

(*ii*) What is the distribution of $(X_1 + X_2)^2/(X_2 - X_1)^2$?

(*iii*) What is the distribution of $(X_2 + X_1)/\sqrt{(X_1 - X_2)^2}$?

(*iv*) What is the distribution of $1/Z$, if $Z = X_1^2/X_2^2$?

[*Delhi Univ. B.Sc. (Maths. Hons.), 1992*]

Ans. (*i*) $\chi^2_{(1)}$; (*ii*) $F(1, 1)$; (*iii*) Standard Cauchy ; (*iv*) $F(1, 1)$

14·5·4. Applications of F-distribution. F-distribution has the following applications in Statistical theory.

14·5·5. F-test for Equality of Population Variances. Suppose we want to test (*i*) whether two independent samples x_i, $(i = 1, 2, \ldots, n_1)$ and y_j, $(j = 1, 2, \ldots, n_2)$ have been drawn from the normal populations with the same variance σ^2, (say), or (*ii*) whether the two independent estimates of the population variance are homogeneous or not.

Under the null hypothesis (H_0) that (*i*) $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, *i.e., the population variances are equal or (ii) Two independent estimates of the population variance are homogeneous,* the statistic F is given by

$$F = \frac{S_X^2}{S_Y^2} \qquad \ldots(14\cdot18)$$

where $S_X^2 = \dfrac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$ and $S_Y^2 = \dfrac{1}{n_2 - 1} \sum_{j=1}^{n_2} (y_j - \bar{y})^2$ ...(14·18*a*)

are unbiased estimates of the common population variance σ^2 obtained from two independent samples and it follows Snedecor's F-distribution with (ν_1, ν_2) d.f. [where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$].

Proof.
$$F = \frac{S_X^2}{S_Y^2} = \left[\frac{n_1}{n_1 - 1} s_X^2\right] \Big/ \left[\frac{n_2}{n_2 - 1} \cdot s_Y^2\right]$$

$$= \left[\frac{n_1 s_X^2}{\sigma_X^2} \cdot \frac{1}{(n_1 - 1)}\right] \Big/ \left[\frac{n_2 s_Y^2}{\sigma_Y^2} \cdot \frac{1}{(n_2 - 1)}\right]$$

$(\because \sigma_X^2 = \sigma_Y^2 = \sigma^2$ under $H_0)$

Since $\frac{n_1 s_X^2}{\sigma_X^2}$ and $\frac{n_2 s_Y^2}{\sigma_Y^2}$ are independent chi-square variates with $(n_1 - 1)$ and $(n_2 - 1)$ d.f. respectively, F follows Snedecor's F-distribution with $(n_1 - 1, n_2 - 1)$ d.f. (c.f. § 14·5).

Remarks 1. In (14·18), greater of the two variances S_X^2 and S_Y^2 is to be taken in the numerator and n_1 corresponds to the greater variance.

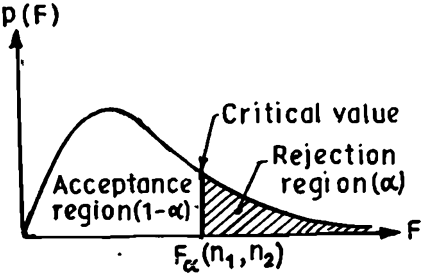
By comparing the calculated value of F obtained by using (14·18) for the two given samples with the tabulated value of F for (n_1, n_2) d.f. at certain level of significance (5% or 1%), H_0 is either rejected or accepted.

2. *Critical values of F-distribution.* The available F-tables (given in the Appendix at the end of the book) give the critical values of F for the right-tailed test, *i.e.*, the critical region is determined by the right-tail areas. Thus the significant value $F_\alpha (n_1, n_2)$ at level of significance α and (n_1, n_2) *d.f.* is determined by

$$P[F > F_\alpha (n_1, n_2)] = \alpha, \qquad \ldots(*)$$

as shown in the following diagram.

CRITICAL VALUES OF F-DISTRIBUTION



From Remark to Example 14·17, we have the following reciprocal relation between the upper and lower α-significant points of F-distribution :

$$F_\alpha (n_1, n_2) = \frac{1}{F_{1-\alpha} (n_2, n_1)}$$

$$\Rightarrow \qquad F_\alpha (n_1, n_2) \times F_{1-\alpha} (n_2, n_1) = 1 \qquad \ldots(**)$$

The critical values of F for left tail test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 < \sigma_2^2$ are given by $F < F_{n_1-1, n_2-1}(1-\alpha)$, and for the two tailed test, $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$ are given by $F > F_{n_1-1, n_2-1}(\alpha/2)$ and $F < F_{n_1-1, n_2-1}(1 - \alpha/2)$ [For details, see § 16·7·5].

Example 14·20. *Pumpkins were grown under two experimental conditions. Two random samples of 11 and 9 pumpkins show the sample standard deviations of their weights as 0·8 and 0·5 respectively. Assuming that the weight distributions are normal, test the hypothesis that the true variances are equal, against the alternative that they are not, at the 10% level.* [*Assume that* $P(F_{10,8} \geq 3{\cdot}35) = 0{\cdot}05$ *and* $P(F_{8,10} \geq 3{\cdot}07) = 0{\cdot}05$].

Solution. We want to test *Null Hypothesis*, $H_0 : \sigma_X^2 = \sigma_Y^2$.

against the *Alternative Hypothesis*, $H_1 : \sigma_X^2 \neq \sigma_Y^2$ (Two-tailed).

We are given :

$$n_1 = 11, n_2 = 9, s_X = 0{\cdot}8 \text{ and } s_Y = 0{\cdot}5.$$

Under the null hypothesis, $H_0 : \sigma_X = \sigma_Y$, the statistic

$$F = \frac{s_X^2}{s_Y^2}$$

follows *F*-distribution with $(n_1 - 1, n_2 - 1)$ d.f.

Now $$n_1 s_X^2 = (n_1 - 1) S_X^2$$

$$\therefore \quad S_X^2 = \left(\frac{n_1}{n_1 - 1}\right) s_X^2 = \left(\frac{11}{10}\right) \times (0{\cdot}8)^2 = 0{\cdot}704$$

Similarly, $$S_Y^2 = \left(\frac{n_2}{n_2 - 1}\right) s_Y^2 = \left(\frac{9}{8}\right) \times (0{\cdot}5)^2 = 0{\cdot}28125$$

$$\therefore \quad F = \frac{0{\cdot}704}{0{\cdot}28125} = 2{\cdot}5$$

The significant values of *F* for two tailed test at level of significance $\alpha = 0{\cdot}10$ are :

$$\left.\begin{aligned} F > F_{10,8}(\alpha/2) = F_{10,8}(0{\cdot}05) \\ \text{and } F < F_{10,8}(1 - \alpha/2) = F_{10,8}(0{\cdot}95) \end{aligned}\right\} \quad \ldots(*)$$

We are given the tabulated (significant) values :

$$P[F_{10,8} \geq 3{\cdot}35] = 0{\cdot}05 \quad \Rightarrow \quad F_{10,8}(0{\cdot}05) = 3{\cdot}35 \quad \ldots(**)$$

Also $$P[F_{8,10} \geq 3{\cdot}07] = 0{\cdot}05 \quad \Rightarrow \quad P\left[\frac{1}{F_{8,10}} \leq \frac{1}{3{\cdot}07}\right] = 0{\cdot}05$$

$$\Rightarrow \quad P[F_{10,8} \leq 0{\cdot}326] = 0{\cdot}05 \quad \Rightarrow \quad P[F_{10,8} \geq 0{\cdot}326] = 0{\cdot}95 \ldots(***)$$

Hence from (*), (**) and (***), the critical values for testing $H_0 : \sigma_X^2 = \sigma_Y^2$, against $H_1 : \sigma_X^2 \neq \sigma_Y^2$ at level of significance $\alpha = 0{\cdot}10$ are given by :

$$F > 3{\cdot}35 \quad \text{and} \quad F < 0{\cdot}326 = 0{\cdot}33$$

Since, the calculated value of *F* (=2·5) lies between 0·33 and 3·35, it is not significant and hence null hypothesis of equality of population variances may be accepted at level of significance $\alpha = 0{\cdot}10$.

Example 14·21. *In one sample of 8 observations, the sum of the squares of deviations of the sample values from the sample mean was 84·4 and in the other sample of 10 observations it was 102·6. Test whether this difference is significant at 5 per cent level, given that the 5 per cent point of F for $n_1 = 7$ and $n_2 = 9$ degrees of freedom is 3·29.* **[Delhi Univ. B.Sc. (Maths Hons.), 1986]**

Solution. Here $n_1 = 8, n_2 = 10$

and $$\Sigma(x - \bar{x})^2 = 84{\cdot}4, \quad \Sigma(y - \bar{y})^2 = 102{\cdot}6$$

$$\therefore \quad S_X^2 = \frac{1}{n_1 - 1}\Sigma(x - \bar{x})^2 = \frac{84{\cdot}4}{7} = 12{\cdot}057$$

$$S_Y^2 = \frac{1}{n_2 - 1}\Sigma(y - \bar{y})^2 = \frac{102{\cdot}6}{9} = 11{\cdot}4$$

Under $H_0 : \sigma_X^2 = \sigma_Y^2 = \sigma^2$, *i.e.*, the estimates of σ^2 given by the samples are homogeneous, the test statistic is

$$F = \frac{S_X^2}{S_Y^2} = \frac{12{\cdot}057}{11{\cdot}4} = 1{\cdot}057$$

Tabulated $F_{0{\cdot}05}$ for (7, 9) d.f. is 3·29.

Since calculated $F < F_{0{\cdot}05}$, H_0 may be accepted at 5% level of significance.

Example 14·22. *Two random samples gave the following results :*

Sample	*Size*	*Sample mean*	*Sum of squares of deviations from the mean*
1	*10*	*15*	*90*
2	*12*	*14*	*108*

Test whether the samples come from the same normal population at 5% level of significance.

[*Given* : $F_{0{\cdot}05}(9, 11) = 2{\cdot}90$, $F_{0{\cdot}05}(11,9) = 3{\cdot}10$ (approx.)

and $t_{0{\cdot}05}(20) = 2{\cdot}086$, $t_{0{\cdot}05}(22) = 2{\cdot}07$]

[Delhi Univ. MCA, 1987]

Solution. A normal population has two parameters, *viz.*, mean μ and variance σ^2. To test if two independent samples have been drawn from the same normal population we have to test (*i*) the equality of population means, and (*ii*) the equality of population variances.

Null Hypothesis : The two samples have been drawn from the same normal population, *i.e.*, $H_0 : \mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$.

Equality of means will be tested by applying *t*-test and equality of variances will be tested by applying *F*-test. Since *t*-test assumes $\sigma_1^2 = \sigma_2^2$, we shall first apply *F*-test and then *t*-test.

We are given
$$\left.\begin{aligned} n_1 = 10,\ n_2 = 12;\ \bar{x}_1 = 15,\ \bar{x}_2 = 14 \\ \Sigma(x_1 - \bar{x}_1)^2 = 90,\ \Sigma(x_2 - \bar{x}_2)^2 = 108 \end{aligned}\right\}$$

F-test

Here
$$S_1^2 = \frac{1}{n_1 - 1}\Sigma(x_1 - \bar{x}_1)^2 = \frac{90}{9} = 10$$

$$S_2^2 = \frac{1}{n_2 - 1}\Sigma(x_2 - \bar{x}_2)^2 = \frac{108}{11} = 9{\cdot}82$$

Since $S_1^2 > S_2^2$, under $H_0 : \sigma_1^2 = \sigma_2^2$, the test statistic is

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) = F(9, 11)$$

Now $F = \frac{10}{9 \cdot 82} = 1 \cdot 018$

Tabulated $F_{0 \cdot 05}(9, 11) = 2 \cdot 90$

Since calculated F is less than tabulated F it is not significant. Hence null hypothesis of equality of population variances may be accepted.

Since $\sigma_1^2 = \sigma_2^2$, we can now apply t test for testing $H_0 : \mu_1 = \mu_2$.

t-test. Under $H_0' : \mu_1 = \mu_2$, against alternative hypothesis, $H_1' : \mu_1 \neq \mu_2$, the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2} = t_{20}$$

where $$S^2 = \frac{1}{n_1 + n_2 - 2} [\Sigma (x_1 - \bar{x}_1)^2 + \Sigma (x_2 - \bar{x}_2)^2]$$

$$= \frac{1}{20} [90 + 108] = 9 \cdot 9$$

$$\therefore \quad t = \frac{15 - 14}{\sqrt{9 \cdot 9 \left(\frac{1}{10} + \frac{1}{12}\right)}} = \frac{1}{\sqrt{9 \cdot 9 \times \frac{11}{60}}}$$

$$= \frac{1}{\sqrt{1 \cdot 815}} = 0 \cdot 742$$

Now $t_{0 \cdot 05}$ for 20 *d.f.* $= 2 \cdot 086$

Since $| t | < t_{0 \cdot 05}$, it is not significant. Hence the hypothesis $H_0' : \mu_1 = \mu_2$ may be accepted. Since both the hypotheses, *i.e.*, $H_0' : \mu_1 = \mu_2$ and $H_0 : \sigma_1^2 = \sigma_2^2$ are accepted, we may regard that the given samples have been drawn from the same normal population.

EXERCISE 14(f)

1. (*a*) If χ_1^2 and χ_2^2 are independent chi-square variates with n_1 and n_2 d.f., obtain the probability density function of F-statistic defined by

$$F = \frac{(\chi_1^2 / n_1)}{(\chi_2^2 / n_2)}$$

Mention the types of hypotheses which are tested with the help of this statistic.

(*b*) Explain why the larger variance is placed in the numerator of the statistic F. Discuss the application of F-test in testing if two variances are homogeneous.

2. An investigator, newly appointed, was made to take ten independent measurements on the maximum internal diameter of a pot at specified equal intervals of time and the standard deviation of these ten observations was found

to be 0·0345 mm. After he had been some time on similar jobs, he was asked to repeat this experiment an equal number of times and the standard deviation of the new set of ten observations was found to be 0·0285 mm. Can it be concluded that the investigator has become more consistent (*i.e.* less variable) with practice ?

3. (*a*) Two independent samples of 8 and 7 items respectively had the following values of the variables.

Sample I :	9	11	13	11	15	9	12	14
Sample II :	10	12	10	14	9	8	10	

Do the estimates of population variance differ significantly ?

[*Delhi Univ. B.Sc., 1992*]

(*b*) Five measurements of the output of two units have given the following results (in kilograms of material per one hour of operation).

Unit A :	14·1	10·1	14·7	13·7	14·0
Unit B :	14·0	14·5	13·7	12·7	14·1

Assuming that both samples have been obtained from normal populations, test at 10% significance level if the two populations have the same variance, it being given that $F_{0.95}(4, 4) = 6.39$

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1991*]

(*c*) In one sample of 10 observations from a normal population, the sum of the squares of the deviations of the sample values from the sample mean is 102·4 and in another sample of 12 observations from another normal population, the sum of the squares of the deviations of the sample values from the sample mean is 120·5. Examine whether the two normal populations have the same variance.

4. (*a*) Two random samples of sizes 8 and 11, drawn from two normal populations, are characterised as follows :

Population from which the sample is drawn	*Size of sample*	*Sum of observations*	*Sum of squares of observations*
I	8	9·6	61·52
II	11	16·5	73·26

You are to decide if the two populations can be taken to have the same variance. What test function would you use ? How is it distributed and what value it has in this sampling experiment ?

(*b*) The following are the values in thousands of an inch obtained by two engineers in 10 successive measurements with the same micrometer. Is one engineer significantly more consistent than the other ?

Engineer A : 503, 505, 497, 505, 495, 502, 499, 493, 510, 501

Engineer B : 502, 497, 492, 498, 499, 495, 497, 496, 498,

Ans. $H_0 : \sigma_1^2 = \sigma_2^2$ (both engineers are equally consistent). $F = 2.4$. Not significant.

(*c*) The nicotine content (in miligrams) of two samples of tobacco were found to be as follows :

Sample A :	24	27	26	21	25	
Sample B :	27	30	28	31	22	36

Can it be said that the two samples come from the same normal population ?

Ans. H_0: $\mu_1 = \mu_2$; $t = 1{\cdot}9$, Not significant.

H_0': $\sigma_1^2 = \sigma_2^2$, $F = 4{\cdot}08 < 6{\cdot}26$ $[F_{0{\cdot}05}(5, 4)]$. Not significant.

Hence the two samples have come from the same normal population.

5. (*a*) Two random samples drawn from two normal populations are :

Sample I : 20, 16, 26, 27, 23, 22, 18, 24, 25, 19

Sample II : 27, 33, 42, 35, 32, 34, 38, 28, 41, 43, 30, 37

Obtain estimates of the variances of the populations and test whether the populations have same variances.

[Given $F_{0{\cdot}05} = 3{\cdot}11$ for 11 and 9 degrees of freedom.]

(*b*) Test H_0: $\sigma_1^2 = \sigma_2^2$ against H_1: $\sigma_1^2 \neq \sigma_2^2$

given $n_1 = 25$, $\Sigma (x_i - \bar{x})^2 = 164 \times 24$,

$n_2 = 21$, $\Sigma (y_j - \bar{y})^2 = 190 \times 21$.

Make necessary assumptions, stating them.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1987*]

(*c*) The diameters of two random samples, each of size 10, of bullets produced by two machines have standard deviations $s_1 = 0{\cdot}01$ and $s_2 = 0{\cdot}015$. Assuming that the diameters have independent distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, test the hypothesis that, the two machines are equally good by testing :

$$H_0 : \sigma_1 = \sigma_2 \text{ against } H_1 : \sigma_1 \neq \sigma_2.$$

6. The following table shows the yield of corn in bushels per plot in 20 plots, half of which are treated with phosphate as fertiliser.

Treated	: 5	0	8	3	6	1	0	3	3	1
Untreated	: 1	4	1	2	3	2	5	0	2	0

Test whether the treatment by phosphate has

(*i*) changed the variability of the plot yields,

(*ii*) improved the average yield of corn.

7. (*a*) The following figures give the prices in rupees of a certain commodity in a sample of 15 shops selected at random from a city *A* and those in a sample of 13 shops from another city *B*.

City *A* :	7·41	7·77	7·44	7·40	7·38	7·93	7·58	
	8·28	7·23	7·52	7·82	7·71	7·84	7·63	7·68
City *B*	7·08	7·49	7·42	7·04	6·92	7·22	7·68	
	7·24	7·74	7·81	7·28	7·43	7·47		

Assuming that the distribution of prices in the two cities is normal, answer the following :

(*i*) Is it possible that the average price of city *B* is Rs. 7·20 ?

(*ii*) Is the observed variance in the first sample consistent with the hypothesis that the standard deviation of prices in city *A* is Rs. 0·30 ?

(*iii*) Is it reasonable to say that the variability of prices in the two cities is the same ?

(*iv*) Is it reasonable to say that the average prices are the same in two cities ?

14·5·6. Relation between t and F distributions. In F-distribution with (ν_1, ν_2) d.f. [c.f. 14·5 (*a*)], take $\nu_1 = 1$, $\nu_2 = \nu$ and $t^2 = F$, *i.e.*, $dF = 2t\,dt$. Thus the probability differential of F transforms to

$$dG(t) = \frac{(1/\nu)^{1/2}}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{(t^2)^{\frac{1}{2}-1}}{\left[1 + \frac{t^2}{\nu}\right]^{(\nu+1)/2}} 2t\,dt, \quad 0 \le t^2 < \infty$$

$$= \frac{1}{\sqrt{\nu}\, B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \cdot \frac{1}{\left[1 + \frac{t^2}{\nu}\right]^{(\nu+1)/2}} dt, \quad -\infty < t < \infty$$

the factor 2 disappearing since the total probability in the range $(-\infty, \infty)$ is unity. This is the probability function of Student's t-distribution with ν d.f. Hence we have the following relation between t and F distributions.

'If a statistic t follows Student's t distribution with n d.f., then t^2 follows Snedecor's F-distribution with (1, n) d.f. Symbolically,

$$\left.\begin{array}{ll} \text{if} & t \sim t_{(n)} \\ \text{then} & t^2 \sim F_{(1,\,n)} \end{array}\right\} \quad \ldots(14\cdot19)$$

Aliter Proof of (14·19). If $\xi \sim N(0, 1)$ and $X \sim \chi^2_{(n)}$ are independent *r.v.*'s then :

$$U = \xi^2 \sim \chi^2_{(1)} \qquad \text{[Square of a S.N.V.]}$$

and
$$t = \frac{\xi}{\sqrt{X/n}} \sim t_{(n)}$$

$$\Rightarrow \qquad t^2 = \frac{\xi^2}{(X/n)} = \frac{(\xi^2/1)}{(X/n)},$$

being the ratio of two independent chi-square variates divided by their respective degrees of freedom is $F(1, n)$ variate.

Hence $t^2 \sim F(1, n)$

With the help of relation (14·19), all the uses of t-distribution can be regarded as the applications of F-distribution also, *e.g.*, for test for a single mean, instead of computing

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}},$$

we may compute

$$F = t^2 = \frac{n(\bar{x} - \mu)^2}{S^2}$$

and then apply F-test with $(1, n)$ d.f. and so on.

Similarly, we can write the test statistic F from § 14·2·9, § 14·2·10 and § 14·2·11 for testing the significance of an observed sample correlation coefficient, regression coefficient and partial correlation coefficient respectively.

Example 14·23. *Given* : $P[F(10, 12) > 2{\cdot}753] = 0{\cdot}05$

$$= P[F(1, 12) > 4{\cdot}747]$$

find $P[F(12, 10) > (2{\cdot}753)^{-1}]$, *and* $P[-\sqrt{4{\cdot}747} < t_{12} < \sqrt{4{\cdot}747}\,]$

Solution.

$$P[F(12, 10) > (2{\cdot}753)^{-1}] = P\left[\frac{1}{F(12, 10)} < 2{\cdot}753\right]$$

$$= P[F(10, 12) < 2{\cdot}753]$$

$$= 1 - P[F(10, 12) > 2{\cdot}753]$$

$$= 1 - 0{\cdot}05 = 0{\cdot}95$$

$$P[-\sqrt{4{\cdot}747} < t_{12} < \sqrt{4{\cdot}747}\,] = P(t^2_{12} < 4{\cdot}747)$$

$$= P[F(1, 12) < 4{\cdot}747)]$$

$$= 1 - P[F(1, 12) > 4{\cdot}747]$$

$$= 1 - 0{\cdot}05 = 0{\cdot}95$$

14·5·7. Relation between F and χ^2. In $F(n_1, n_2)$ distribution if we let $n_2 \to \infty$, then $\chi^2 = n_1 F$ follows χ^2-distribution with n_1 d.f.

Proof. We have

$$p(F) = \frac{(n_1/n_2)^{n_1/2} F^{(n_1/2)-1}}{\Gamma(n_1/2)\,\Gamma(n_2/2)} \cdot \frac{\Gamma(n_1 + n_2)/2]}{\left[1 + \frac{n_1}{n_2} F\right]^{(n_1 + n_2)/2}}, \quad 0 < F < \infty$$

In the limit as $n_2 \to \infty$, we have

$$\frac{\Gamma(n_1 + n_2)/2]}{n_2^{n_1/2}\,\Gamma(n_2/2)} \to \frac{(n_2/2)^{n_1/2}}{n_2^{n_1/2}} = \frac{1}{2^{n_1/2}}$$

$$\left[\because \frac{\Gamma(n + k)}{\Gamma(n)} \to n^k \text{ as } n \to \infty. (c.f. \text{ Remark below.})\right]$$

$$\text{Also } \lim_{n_2 \to \infty} \left[1 + \frac{n_1}{n_2} F\right]^{(n_1 + n_2)/2} = \lim_{n_2 \to \infty} \left[\left(1 + \frac{n_1}{n_2} F\right)^{n_2}\right]^{1/2} \times \lim_{n_2 \to \infty} \left(1 + \frac{n_1}{n_2} F\right)^{n_1/2}$$

$$= \exp(n_1 F/2) = \exp(\chi^2/2) \quad (\because n_1 F = \chi^2)$$

Hence in the limit, the p.d.f. of $\chi^2 = n_1 F$ becomes

$$dP(\chi^2) = \frac{(n_1/2)^{n_1/2}\, e^{-\chi^2/2}}{\Gamma(n_1/2)} \cdot \left(\frac{\chi^2}{n_1}\right)^{(n_1/2)-1} d\left(\frac{\chi^2}{n_1}\right)$$

$$= \frac{1}{2^{n_1/2}\,\Gamma(n_1/2)} \cdot e^{-\chi^2/2} (\chi^2)^{(n_1/2)-1}\, d\chi^2, \; 0 < \chi^2 < \infty$$

which is the p.d.f. of chi-square distribution with n_1 d.f.

Remark. $\lim_{n \to \infty} \frac{\Gamma(n+k)}{\Gamma(n)} = \lim_{n \to \infty} \frac{(n+k-1)!}{(n-1)!}$, (for large n)

$$\simeq \lim \frac{\sqrt{2\pi}\, e^{-(n+k-1)} (n+k-1)^{n+k-(1/2)}}{\sqrt{2\pi}\, e^{-(n-1)} (n-1)^{n-(1/2)}}$$

(On using Stirling's approximation for $n!$ as $n \to \infty$.)

$$= e^{-k} \lim_{n \to \infty} \frac{n^{n+k-\frac{1}{2}} \left(1 + \frac{k-1}{n}\right)^{n+k-\frac{1}{2}}}{n^{n-\frac{1}{2}} \left(1 - \frac{1}{n}\right)^{n-\frac{1}{2}}}$$

$$= e^{-k} n^k \frac{\lim_{n \to \infty} \left(1 + \frac{k-1}{n}\right)^{n} \lim_{n \to \infty} \left(1 + \frac{k-1}{n}\right)^{k-\frac{1}{2}}}{\lim_{n \to \infty} \left(1 - \frac{1}{n}\right)^{n} \lim_{n \to \infty} \left(1 - \frac{1}{n}\right)^{-\frac{1}{2}}}$$

$$= e^{-k} n^k \left[\frac{e^{(k-1)} \cdot 1}{e^{-1} \cdot 1}\right] = n^k$$

$$\lim_{n \to \infty} \frac{\Gamma(n+k)}{\Gamma n} = n^k$$

14·5·8. F-test for Testing the Significance of an Observed Multiple Correlation Coefficient. If R is the observed multiple correlation coefficient of a variate with k other variates in a random sample of size n from a $(k+1)$ variate population, then Prof. R.A. Fisher proved that under the null hypothesis (H_0) *that the multiple correlation coefficient in the population is zero, the statistic*

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n-k-1}{k}. \qquad \ldots(14\cdot20)$$

conforms to F-distribution with $(k, n-k-1)$ d.f.

14·5·9. F-test for significance of an observed sample correlation Ratio η_{YX}. Under the null hypothesis that *population correlation ratio is zero,* the test statistic is

$$F = \frac{\eta^2}{1 - \eta^2} \frac{N-h}{h-1} \sim F(h-1, N-h) \qquad \ldots(14\cdot21)$$

where N is the size of the sample (from a bi-variate normal population) arranged in h arrays.

14·5·10. F-test for Testing the Linearity of Regression. For a sample of size N arranged in h arrays, from a bi-variate normal population, the test statistic for testing the hypothesis of linearity of regression is

$$F = \frac{\eta^2 - r^2}{1 - \eta^2} \frac{N-h}{h-2} \sim F(h-2, N-h) \qquad \ldots(14\cdot22)$$

14·5·11. F-test for Equality of Several Means. This test is carried out by the technique of Analysis of Variance, which plays a very important and fundamental role in Design of Experiments in Agricultural Statistics.

14·5. Non-Central F-distribution. The ratio of two independent χ^2 variates each divided by the corresponding d.f. has a non-central F-distribution if the numerator has a non-central χ^2-distribution and the denominator has a central χ^2-distribution. Thus, if X has a non-central χ^2-distribution with n_1 d.f. and non-centrality parameter λ, *i.e.*, if $X \sim \chi'^2_{n_1}$ and Y is an independent (central) χ^2-variate with n_2 d.f. *i.e.*, if $Y \sim \chi^2_{n_2}$, then the non-central F-statistic is defined as:

$$F' = \frac{X/n_1}{Y/n_2} = \frac{n_2 X}{n_1 Y}$$

p.d.f. of F'. Since X and Y are independent, their joint p.d.f. is given by

$$p(x, y) = p_1(x).\, p_2(y) = \left[\sum_{i=0}^{\infty} \frac{e^{-\lambda} \lambda^i}{i!} \cdot \frac{e^{-x/2}\, x^{(n_1/2) + i - 1}}{2^{(n_1 + 2i)/2}\, \Gamma[(n_1 + 2i)/2]} \right]$$

$$\times \frac{e^{-y/2}\, y^{(n_2/2) - 1}}{2^{n_2/2}\, \Gamma(n_2/2)}; \quad 0 \le (x, y) < \infty.$$

Let us transform to the new set of r.v.'s F' and U defined by the transformation:

$$F' = \frac{n_2 x}{n_1 y} \text{ and } u = y \Rightarrow y = u,\ x = \frac{n_1 u F'}{n_2}$$

$$J = \frac{\partial(x, y)}{\partial(F', u)} = \begin{vmatrix} \frac{n_1}{n_2} u & \frac{n_1}{n_2} F' \\ 0 & 1 \end{vmatrix} = \frac{n_1}{n_2} u$$

The joint p.d.f. of F' and U is given by

$$g(F', u) = \left\{ \sum_{i=0}^{\infty} \frac{e^{-\lambda} \lambda^i}{i!} \frac{\exp\left[-\frac{n_1 u F'}{2 n_2} \right] \cdot \left(\frac{n u F'}{n_2} \right)^{(n_1/2) + i - 1}}{2^{i + (n_1 + n_2)/2}\, \Gamma(n_2/2)\, \Gamma[(n_1 + 2i)/2]} \right\}$$

$$\times e^{-u/2} u^{(n_2/2) - 1} \cdot \left(\frac{n_1}{n_2} \right) u$$

$$= \frac{n_1}{n_2} \sum_{i=0}^{\infty} \left\{ \frac{e^{-\lambda} \lambda^i}{i!} \frac{\left(\frac{n_1}{n_2} F' \right)^{(n_1/2) + i - 1}}{2^{i + (n_1 + n_2)/2}\, \Gamma(n_2/2)\, \Gamma[(n_1 + 2i)/2]} \right.$$

$$\left. \times \exp\left[-\frac{u}{2} \left(1 + \frac{n_1}{n_2} F' \right) \right] \cdot u^{\frac{n_1 + n_2}{2} + i - 1} \right\}$$

$$0 \le F' < \infty,\ 0 < u < \infty$$

Integrating it w.r.t. u between the limits 0 to ∞ and using Gamma Integral, we obtain the marginal p.d.f. of F' as

$$g(F') = \frac{n_1}{n_2} \sum_{i=0}^{\infty} \left\{ \frac{e^{-\lambda}\lambda^i}{i!} \cdot \frac{\left(\frac{n_1}{n_2}F'\right)^{(n_1/2)+i-1}}{2^{i+(n_1+n_2)/2}\,\Gamma(n_2/2)\,\Gamma[(n_1+2i)/2]} \times \frac{\Gamma\left(\frac{n_1+n_2}{2}+i\right)}{\left[\frac{1}{2}\left(1+\frac{n_1}{n_2}F'\right)\right]^{i+(n_1+n_2)/2}} \right\}$$

$$= \frac{n_1}{n_2} \sum_{i=0}^{\infty} \left\{ \frac{e^{-\lambda}\lambda^i}{i!} \cdot \frac{\left(\frac{n_1}{n_2}F'\right)^{(n_1/2)+i-1}}{B\left(\frac{n_1}{2}+i, \frac{n_2}{2}\right)} \times \frac{1}{\left(1+\frac{n_1}{n_2}F'\right)^{i+(n_1+n_2)/2}} \right\}; \; 0 \le F' < \infty \quad \ldots(14\cdot23)$$

Remarks. 1. For $\lambda = 0$, we get

$$g(F') = \frac{n_1}{n_2} \cdot \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{\left(\frac{n_1}{n_2}F'\right)^{(n_1/2)-1}}{\left(1+\frac{n_1}{n_2}F'\right)^{(n_1+n_2)/2}}; \; 0 \le F' < \infty,$$

since for $\lambda = 0$, we get the contribution from the sum only when $i = 0$ and all other terms vanish. Thus for $\lambda = 0$, $g(F')$ reduces to the p.d.f. of central F-distribution with (n_1, n_2) d.f.

2. The hyper-geometric function of first kind is defined by

$${}_1F_1(a, b, y) = \sum_{i=0}^{\infty} \frac{\Gamma(a+i)\,\Gamma b}{\Gamma a\,\Gamma(b+i)} \cdot \frac{y^i}{i!} \qquad \ldots(14\cdot23b)$$

$$\therefore \quad {}_1F_1\left(\frac{n_1+n_2}{2}, \frac{n_1}{2}, \frac{\lambda n_1 F'}{n_2\left(1+\frac{n_1}{n_2}F'\right)}\right)$$

$$= \sum_{i=0}^{\infty} \frac{\Gamma\left(\frac{n_1+n_2}{2}+i\right)\Gamma\left(\frac{n_1}{2}\right)}{\Gamma\left(\frac{n_1+n_2}{2}\right)\Gamma\left(\frac{n_1}{2}+i\right)} \times \frac{(\lambda n_1 F')^i}{\left[n_2\left(1+\frac{n_1}{n_2}F'\right)\right]^i} \times \frac{1}{i!}$$

$$\begin{vmatrix} y & x^2 & x & 1 \\ \Sigma y_i & \frac{n(n+1)(2n+1)}{3} & 0 & an+1 \\ \Sigma x_i y_i & 0 & \frac{n(n+1)(2n+1)}{3} & 0 \\ \Sigma x_i^2 y_i & \frac{n^2(n+1)^2}{2} & 0 & \frac{n(n+1)(2n+1)}{3} \end{vmatrix} = 0$$

[Delhi Univ. B.A. (Pass), 1984]

Hint. Use (9·4*b*), with

$$x_i = a + i\,;\ i = 1, 2, ..., (2n+1).\ \text{Since } \bar{x} = 0,$$

$$\Sigma x_i = (2n+1)a + \Sigma i \quad \Rightarrow \quad 0 = (2n+1)a + \frac{(2n+1)(2n+2)}{2}$$

$$\Rightarrow \qquad a = -(n+1)$$

$$\Sigma x_i^2 = \Sigma (a+i)^2 = (2n+1)a^2 + \Sigma i^2 + 2a \Sigma i$$

and so on, for Σx_i^3 and Σx_i^4.

$$\left[\sum_{i=1}^{m} i^2 = \frac{m(m+1)(2m+1)}{6}\,;\ \sum_{i=1}^{m} i^3 = \left[\frac{m(m+1)}{2} \right]^2 \right.$$

and $$\left. \sum_{i=1}^{m} i^4 = \frac{1}{30} m(m+1)(2m+1)(3m^2+3m-1) \right]$$

(b) When do we prefer logarithmic curve to ordinary curve ?

9·5. Curve Fitting by Orthogonal Polynomials. Suppose that the polynomial of *p*th degree of Y on X is

$$Y = a_0 + a_1 X + a_2 X^2 + \ldots + a_p X^p \qquad \ldots(9\cdot13)$$

The normal equations for determining the constants a_i's are obtained by the principle of least squares by minimising the residual or error sum of squares

$$E = \Sigma (y - a_0 - a_1 x - a_2 x^2 - \ldots - a_p x^p)^2 \qquad \ldots(9\cdot14)$$

summation being extended over the given set of observations. The normal equations are :

$$\frac{\partial E}{\partial a_j} = 0,\ (j = 0, 1, 2, ..., p)$$

i.e., $$\Sigma x^j (y - a_0 - a_1 x - a_2 x^2 - \ldots - a_p x^p) = 0,\ [j = 0, 1, 2, ..., p] \qquad \ldots(9\cdot15)$$

Assume that X and Y are measured from their means (and this we can do without any loss of generality) so that

$$\mu_r = \mu_r' = E(X^r) = \frac{1}{N} \Sigma x^r$$

and write,

$$\mu_{j1} = \frac{1}{N} \Sigma x^j . y\,,$$

where N is number of observations taken on each of the variables X and Y. Hence (9·15) gives

$$\mu_{j1} - a_0\mu_j - a_1\mu_{j+1} - a_2\mu_{j+2} - \ldots - a_p\mu_{j+p} = 0\ ;\quad j = 0, 1, 2, \ldots, p$$

$$\Rightarrow \quad a_0\mu_j + a_1\mu_{j+1} + a_2\mu_{j+2} + \ldots + a_p\mu_{j+p} = \mu_{j1}\ ;\ j = 0, 1, 2, \ldots, p$$

Putting $j = 0, 1, 2, \ldots, p$, we get respectively

$$\left.\begin{array}{llllll} a_0\mu_0 + a_1\mu_1 & + a_2\mu_2 & + \ldots & + a_p\mu_p & = \mu_{01} \\ a_0\mu_1 + a_1\mu_2 & + a_2\mu_3 & + \ldots & + a_p\mu_{p+1} & = \mu_{11} \\ \vdots \quad\quad \vdots & \vdots & & \vdots & \vdots \\ a_0\mu_p + a_1\mu_{p+1} & + a_2\mu_{p+2} & + \ldots & + a_p\mu_{2p} & = \mu_{p1} \end{array}\right\} \quad \ldots(9\cdot16)$$

Solving (9·16) for $a_0, a_1, \ldots, a_p$ in terms of the moments μ_j's and μ_{j1}'s, $j = 0, 1, 2, \ldots, p$ and substituting in (9·13) we get the required curve of best fit.

Let

$$\Delta^{(p)} = \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \ldots & \mu_p \\ \mu_1 & \mu_2 & \mu_3 & \ldots & \mu_{p+1} \\ \vdots & \vdots & \vdots & & \vdots \\ \mu_p & \mu_{p+1} & \mu_{p+2} & \ldots & \mu_{2p} \end{vmatrix} \quad \ldots(9\cdot17)$$

and $\Delta_j^{(p)}$ be the determinant obtained on replacing $(j+1)$th column of $\Delta^{(p)}$ by the column

$$\begin{pmatrix} \mu_{01} \\ \mu_{11} \\ \vdots \\ \mu_{p1} \end{pmatrix} \quad \text{then} \qquad a_j = \frac{\Delta_j^{(p)}}{\Delta^{(p)}} \qquad \ldots(9\cdot18)$$

The required curve of best fit is the eliminant of a_i's in (9·13) and (9·16) and is given by

$$\begin{vmatrix} Y & 1 & X & X^2 & \ldots & X^p \\ \mu_{01} & \mu_0 & \mu_1 & \mu_2 & \ldots & \mu_p \\ \mu_{11} & \mu_1 & \mu_2 & \mu_3 & \ldots & \mu_{p+1} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \mu_{p1} & \mu_p & \mu_{p+1} & \mu_{p+2} & \ldots & \mu_{2p} \end{vmatrix} = 0 \quad \ldots(9\cdot19)$$

The use of equation (9·19) is subject to one serious drawback. If we have a set of data and apart from inspection if there is no guide regarding the order of the polynomial to be fitted, the only way left to us is to try curves of order 1, 2, 3, ... until we reach the point where further terms do not improve the fit. Every time we add a new term, the a_j's given by (9·18) change and accordingly the determinantal arithmetic has to be done afresh. For example, if we want to fit a polynomial curve of third or higher degree to the same data then we cannot use the coefficients which we computed while fitting a second degree parabola. To overcome this drawback Prof. R.A. Fisher suggested a method which involved the fitting of *Orthogonal Polynomials* by the principle of least squares, so that each term is independent of the other, *i.e.*, each of the coefficients in

the polynomial is independent of the other so that each of them can be calculated independently. In this method, the coefficients computed earlier remain the same and we have to compute the coefficient only for the added term.

9·5·1. Orthogonal Polynomials *(Def.).* Two polynomials $P_1(x)$ and $P_2(x)$ are said to be *orthogonal* to each other if

$$\Sigma P_1(x)\, P_2(x) = 0, \qquad \ldots(9\cdot20)$$

where summation is taken over a specified set of values of x. If x were a continuous variable in the range from a to b, the condition for orthogonality gives

$$\int_a^b P_1(x)\, P_2(x)\, dx = 0 \qquad \ldots(9\cdot20a)$$

For example, if we take

$$P_0 = 1, P_1(x) = x - 4, P_2(x) = x^2 - 8x + 12, P_3(x) = x^3 - 12x^2 + 41x - 36 \qquad \ldots(9\cdot20b)$$

then these are orthogonal to each other for a set of integral values of x from 1 to 7 as explained in the following table. Other examples of orthogonal polynomials are Hermite polynomials, Gram Charlier's polynomials, Legender's polynomials, etc.

ORTHOGONALITY OF POLYNOMIALS DEFINED IN (9·20*b*)

x	P_0P_1	P_0P_2	P_0P_3	P_1P_2	P_1P_3	P_2P_3
1	−3	5	−6	−15	18	−30
2	−2	0	6	0	−12	0
3	−1	−3	6	3	−6	−18
4	0	−4	0	0	0	0
5	1	−3	−6	−3	−6	18
6	2	0	−6	0	−12	0
7	3	5	6	15	18	30
Total	0	0	0	0	0	0

9·5·2. Fitting of Orthogonal Polynomials. The pth degree polynomial (9·13) can be rewritten as

$$Y = b_0P_0 + b_1P_1 + b_2P_2 + \ldots + b_pP_p \qquad \ldots(9\cdot21)$$

where P's are polynomials in x, P_j being a polynomial of degree j, ($j = 0, 1, 2, \ldots, p$). We shall determine P's so that they satisfy the condition of orthogonality, *viz.*,

$$\Sigma P_j P_k = \sum_x P_j(x)\, P_k(x) = 0 \,;\, j \neq k \qquad \ldots(9\cdot22)$$

the summation being extended over the observed values of x. The normal equations for estimating the constants b_j's are obtained on minimising

$$E = \Sigma (y - b_0P_0 - b_1P_1 - \ldots - b_pP_p)^2 \qquad \ldots(9\cdot23)$$

and are given by

$$\frac{\partial E}{\partial b_j} = 0$$

$$\Rightarrow \quad \Sigma P_j (y - b_0 P_0 - b_1 P_1 - \ldots - b_p P_p) = 0 : j = 0, 1, 2, \ldots, p .$$

Simlifying and using (9·22), we get

$$\Sigma P_j . y - b_j \Sigma P_j^2 = 0$$

$$\Rightarrow \quad b_j = \frac{\Sigma y P_j}{\Sigma P_j^2}, j = 0, 1, 2, \ldots, p . \qquad \ldots(9·24)$$

Thus b_j is determined by P_j. If having fitted a curve of order p we wish to go a step further by adding a term $b_{p+1} P_{p+1}$, the coefficients already obtained in (9·24) remain unaltered.

Moreover, the use of orthogonal polynomials will give us a very convenient method of determining, step by step, the goodness of fit of the polynomial curve. For pth degree polynomial (9·21), the error sum of squares is [c.f. (9·23)]

$$\begin{aligned} E &= \Sigma (y - b_0 P_0 - b_1 P_1 - \ldots - b_p P_p)^2 \\ &= \Sigma y^2 + b_0^2 \Sigma P_0^2 + b_1^2 \Sigma P_1^2 + \ldots + b_p^2 \Sigma P_p^2 \\ &\qquad - 2 b_0 \Sigma y P_0 - 2 b_1 \Sigma y P_1 - \ldots - 2 b_p \Sigma y P_p , \end{aligned}$$

other terms vanish because of orthogonality conditions (9·22). Using (9·24) we finally obtain

$$E = \Sigma y^2 - b_0^2 \Sigma P_0^2 - b_1^2 \Sigma P_1^2 - \ldots - b_p^2 \Sigma P_p^2 \qquad \ldots(9·25)$$

Thus the effect of adding any term $b_j P_j$ is to reduce the error (residual) sum of squares E by $b_j^2 \Sigma P_j^2$ and we may examine the effect of this term on E separately. If we find that the addition of any term $b_p P_p$ does not reduce E significantly, we may conclude that it is not desired (as far as the representation of the given data by a polynomial curve is concerned).

9·5·3. Finding The Orthogonal Polynomial P_p, in Fitting a Polynomial of Degree p. Let P_p, the polynomial of degree p in x be given by

$$P_p = \sum_{j=0}^{p} c_{pj} x^j \qquad \ldots(9·26)$$

This contains $(p + 1)$ unknown constants $c_{p0}, c_{p1}, \ldots, c_{pp}$. Hence in all the polynomials in (9·21) up to and including those of pth order, there are

$$1 + 2 + 3 + \ldots + (p + 1) = \frac{(p + 1)(p + 2)}{2},$$

unknown constants. The orthogonality conditions

$$\Sigma P_i P_j = 0, i \neq j = 0, 1, 2, \ldots, p,$$

provide ${}^{p+1}C_2 = \frac{(p + 1)p}{2}$ conditions on the c's so that there are

$$\frac{(p + 1)(p + 2)}{2} - \frac{(p + 1)p}{2} = p + 1,$$

constants which can be assigned arbitrarily. We will take one for each polynomial P_j ($j = 0, 1, 2, ..., p$) and assign it such that the coefficient of x^j in P_j is unity *i.e.*,

$$c_{jj} = 1, j = 0, 1, 2, ..., p, \qquad ...(9.27)$$

In particular $c_{00} = P_0 = 1$. The orthogonality conditions give:

$$\sum_x P_p P_j = 0, j < p \ (j = 0, 1, 2, ..., p-1) \qquad ...(9·28)$$

$j = 0$, gives $\sum P_p P_0 = 0 \Rightarrow \sum P_p = 0$; $(\because P_0 = 1)$ ...(*)

$j = 0$, gives $\sum P_p P_1 = 0 \Rightarrow \sum P_p = 0, (x + k) = 0$

$\Rightarrow \sum P_p . x + k \sum P_p = 0$

$\Rightarrow \sum x P_p = 0$ ...(**)

$j = 2$, gives $\sum P_p P_2 = 0 \Rightarrow \sum P_p = (x^2 + k_1 x + k_2) = 0$ [Using (*)]

$\Rightarrow \sum x^2 . P_p = 0$ [Using (*) and (**)]

Similarly proceeding, we shall get in general

$$\sum_x P_p x^r = 0, \quad r = 0, 1, 2, ..., p-1 \qquad ...(9·29)$$

$$\Rightarrow \sum_x \left(\sum_{j=0}^{p} c_{pj} . x^j \right) x^r = 0$$

$$\Rightarrow \sum_{j=0}^{p} \left(c_{pj} \sum_x x^{j+r} \right) = 0$$

Dividing both sides by N, the number of observations on each of the variables X and Y, we get.

$$\sum_{j=0}^{P} c_{pj} \ \mu_{j+r} = 0; r = 0, 1, 2, ..., (p-1) \qquad ...(9·30)$$

where x is assumed to be measured from mean. Putting $r = 0, 1, 2, .., (p-1)$ in (9·30), we get respectively

$$c_{p0}\mu_0 + c_{p1}\mu_1 + \cdots + c_{pj}\mu_j + \cdots + c_{p,p-1}\mu_{p-1} + c_{pp}\mu_p = 0$$
$$c_{p0}\mu_1 + c_{p1}\mu_2 + \cdots + c_{pj}\mu_{j+1} + \cdots + c_{p,p-1}\mu_p + c_{pp}\mu_{p+1} = 0$$
$$c_{p0}\mu_{p-1} + c_{p1}\mu_p + \cdots + c_{pj}\mu_{j+p-1} + \cdots + c_{p,p-1}\mu_{2p-2} + c_{pp}\mu_{2p-1} = 0$$

Noting that $c_{pp} = 1$, solving the above equations for c's, we get

$$c_{pj} = \frac{\begin{vmatrix} \mu_0 & \mu_1 & \cdots & -\mu_p & \cdots & \mu_{p-1} \\ \mu_1 & \mu_2 & \cdots & -\mu_{p+1} & \cdots & \mu_p \\ \vdots & \vdots & & \vdots & & \vdots \\ \mu_{p-1} & \mu_p & \cdots & \mu_{2p-1} & \cdots & \mu_{2p-2} \end{vmatrix}}{\begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_j & \cdots & \mu_{p-1} \\ \mu_1 & \mu_2 & \cdots & \mu_{j+1} & \cdots & \mu_p \\ \vdots & \vdots & & \vdots & & \vdots \\ \mu_{p-1} & \mu_p & \cdots & \mu_{j+p-1} & \cdots & \mu_{2p-2} \end{vmatrix}} = \frac{\Delta^{(p)} pj}{\Delta^{(p-1)}} \qquad ...(9·31)$$

where $\Delta^{(p)}$ has been defined in (9.17) and $\Delta^{(p)}_{pj}$ is the minor of the element in the last row and $(j+1)$th column in $\Delta^{(p)}$. Substituting this value of c_{pj} in (9·26), we get

$$P_p = \sum_{j=0}^{p} c_{pj} x^j = \sum_{j=0}^{p} \frac{\Delta^{(p)}_{pj}}{\Delta^{(p-1)}} . x^j$$

$$= \frac{1}{\Delta^{(p-1)}} \sum_{j=0}^{p} \Delta^{(p)}_{pj} x^j$$

$$= \frac{1}{\Delta^{(p-1)}} \left[\Delta^{(p)}_{p0} + x \Delta^{(p)}_{p1} + \ldots + x^p . \Delta^{(p)}_{pp} \right]$$

$$\Rightarrow \quad P_p = \frac{1}{\Delta^{(p-1)}} \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \mu_p \\ \mu_1 & \mu_2 & \mu_3 & \mu_{p+1} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{p-1} & \mu_p & \mu_{p+1} & \mu_{2p-1} \\ 1 & x & x^2 & x^p \end{vmatrix} \qquad \ldots(9\cdot32)$$

In particular if $\mu_0 = 1, \mu_1 = 0$ and $\mu_2 = 1$, *i.e.*, if x is a standardised variate then the orthogonal polynomials are given by

$$P_0 = 1 \qquad \ldots(9\cdot33)$$

$$P_1(x) = \frac{\begin{vmatrix} \mu_0 & \mu_1 \\ 1 & x \end{vmatrix}}{\mu_0} = x \qquad \ldots(9\cdot33a)$$

$$P_2(x) = \frac{\begin{vmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ 1 & x & x^2 \end{vmatrix}}{\begin{vmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{vmatrix}} = x^2 - \mu_3 x - 1 \qquad \ldots(9\cdot33b)$$

$$(\because \mu_0 = 1, \mu_1 = 0, \mu_2 = 1)$$

$$P_3(x) = \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 & \mu_3 \\ \mu_1 & \mu_2 & \mu_3 & \mu_4 \\ \mu_2 & \mu_3 & \mu_4 & \mu_5 \\ 1 & x & x^2 & x^3 \end{vmatrix} \div \begin{vmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{vmatrix} \qquad \ldots(9\cdot33c)$$

and so on.

If we further assume that x is a standard *normal* variate so that $\mu_3 = \mu_5 = \ldots = \mu_{2r+1} = 0$, then the above orthogonal polynomials are called *Hermite Polynomials* and are given by

$P_0 = 1 ; P_1(x) = x ; P_2(x) = x^2 - 1 ; P_3(x) = x^3 - 3x ; P_4(x) = x^4 - 6x^2 + 3 ;$ and so on, where x is a continuous r.v. taking values from $-\infty$ to ∞. ...(9·34)

Remark. Hermite Polynomials defined in (9·34) are orthogonal w.r.t. the weight function

$$\alpha(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right); \quad -\infty < x < \infty$$

i.e.,
$$\int_{-\infty}^{\infty} P_i(x)\, P_j(x)\, \alpha(x) = 0\,; \quad i \neq j \qquad \ldots(9{\cdot}35)$$

where $P_1(x), P_2(x), P_3(x), P_4(x)$ are defined in (9·34).

9·5·4. Determination of the Coefficients b_j's in (9·21). From (9·24), we get

$$b_p = \Sigma\, y\, P_p \,/\, \Sigma\, P_p^{\,2} \qquad \ldots(9{\cdot}36)$$

Now
$$\Sigma P_p^{\,2} = \Sigma P_p\, P_p$$
$$= \sum_x P_p\, [c_{p0} + c_{p1}x + c_{p2}x^2 + \ldots + c_{pp}x^p\,]$$
$$= \sum_x P_p \,.\, x^p,$$

on using (9·29) and the fact that $c_{pp} = 1$.

$$\Sigma P_p^{\,2} = \sum_x \left(\sum_{j=0}^{p} c_{pj}\, x^j \right) x^p = \sum_{j=0}^{p} \left(c_{pj} \sum_x x^{p+j} \right)$$

$$= N \sum_{j=0}^{p} c_{pj}\, \mu_{p+j} = N \sum_{j=0}^{p} \frac{\Delta^{(p)}_{pj}}{\Delta^{(p-1)}} \cdot \mu_{p+j} \qquad \text{[From (9·31)]}$$

$$= \frac{N}{\Delta^{(p-1)}} \begin{vmatrix} \mu_0 & \mu_1 & \ldots & \mu_p \\ \mu_1 & \mu_2 & \ldots & \mu_{p+1} \\ \vdots & & & \\ \mu_{p-1} & \mu_p & \ldots & \mu_{2p-1} \\ \mu_p & \mu_{p+1} & \ldots & \mu_{2p} \end{vmatrix}$$

[Proceeding exactly as we obtained (9·32)]

$$= \frac{N\, \Delta^{(p)}}{\Delta^{(p-1)}} \qquad \ldots(9{\cdot}37)$$

Similarly, $\Sigma\, y\, P_p = N \sum_{j=0}^{p} \frac{\Delta^{(p)}_{pj}}{\Delta^{(p-1)}} \cdot \mu_{j1}$

$$= \frac{N}{\Delta^{(p-1)}} \begin{vmatrix} \mu_0 & \mu_1 & \ldots & \mu_p \\ \mu_1 & \mu_2 & \ldots & \mu_{p+1} \\ \vdots & & & \\ \mu_{p-1} & \mu_p & \ldots & \mu_{2p-1} \\ \mu_{01} & \mu_{11} & \ldots & \mu_{p1} \end{vmatrix}$$

$$= \frac{N \,.\, \Delta^{(p)}}{\Delta^{(p-1)}} \qquad \ldots(9{\cdot}38)$$

where $\Delta^{(p)}$ and $\Delta^{(p)}_j$ are defined in (9·17). Substituting in (9·36) we get

...(9·39)

If the variable x takes the integral values 1, 2, ..., N, then the first seven of these orthogonal polynomials P_j's $j = 0, 1, 2, 3, ..., 6$ are given by :

$$P_0(x) = 1, P_1(x) = \lambda_1 . \xi$$

$$P_2(x) = \lambda_2 \left\{ \xi^2 - \frac{N^2 - 1}{12} \right\}$$

$$P_3(x) = \lambda_3 \left\{ \xi^3 - \frac{3N^2 - 7}{20} \xi \right\}$$

$$P_4(x) = \lambda_4 \left\{ \xi^4 - \frac{3N^2 - 13}{14} \xi^2 + \frac{3}{560} (N^2 - 1)(N^2 - 9) \right\}$$

$$P_5(x) = \lambda_5 \left\{ \xi^5 - \frac{5}{18} (N^2 - 7) \xi^3 + \frac{1}{1008} (15N^4 - 230N^2 + 407) \xi \right\}$$

$$P_6(x) = \lambda_6 \left\{ \xi^6 - \frac{5}{44} (3N^2 - 31) \xi^4 + \frac{1}{176} (5N^4 - 110N^2 + 329) \xi^2 \right\}$$
$$- \frac{5}{14784} (N^2 - 1)(N^2 - 9)(N^2 - 25) \Bigg\}$$

and so on, where $\xi = x - \bar{x}$ so that $\Sigma \xi = 0$ and λ_i's are arbitrary constants.

If $y = b_0 + b_1 P_1(x) + b_2 P_2(x) + ... + b_p P_p(x)$; is the orthogonal polynomial fitted to the given data then, using (9·24), we get

$$\left.\begin{aligned} b_0 &= \frac{\Sigma y P_0}{\Sigma P_0^2} = \frac{\Sigma y}{N} \ ; \ (\because P_0 = 1) \\ b_i &= \frac{\sum\limits_x y P_i}{\sum\limits_x P_i^2} \ , (i = 1, 2, ..., p) \end{aligned}\right\} \qquad ...(9·40)$$

The origin of P_i's is so chosen that $\Sigma P_i = 0$.

If N, the number of observations is odd, then we take

$$\xi = \frac{x_j - A}{h}$$

and if N is even then we take

$$\xi = \frac{x_i - A_1}{(h/2)}$$

where h = length of the interval (for values of x)
A = middle value (item) of the data
and A_1 = Arithmetic mean of two middle values of the data.

The values of P_i's and λ_i's are obtained from 'Statistical Tables' by

R.A. Fisher for the values of N from 3 to 75. In these tables the orthogonal polynomials P_i's are denoted by ϕ_i's. We reproduce below these tables for $N = 3$ to $N = 6$.

TABLES OF ORTHOGONAL POLYNOMIALS

	$N=3$		$N=4$			$N=5$			
	φ_1	φ_2	φ_1	φ_2	φ_3	φ_1	φ_2	φ_3	φ_4
	−1	1	−3	1	−1	−2	2	−1	1
	0	−2	−1	−1	−3	−1	−1	2	−4
	1	1	1	−1	−3	0	−2	0	6
			3	1	1	1	−1	−2	−4
						2	2	1	1
	2	6	20	4	20	10	14	10	70
$\sum_x \phi_i^2$					10			5	35
λ_i	1	3	2	1	3	1	1	6	22

	$N=6$				
	φ_1	φ_2	φ_3	φ_4	φ_5
	−5	5	−5	1	−1
	−3	−1	7	−3	5
	−1	−4	4	2	−10
	1	−4	−4	−2	10
	3	−1	−7	−3	−5
	5	5	5	1	1
$\sum_x \varphi_i^2$	70	84	180	28	252
λ_i	2	3/2	5/3	7/12	21/10

Example 9·8. *Fit a straight line* $y = a + bx$...(*) *to the following data by using orthogonal polynomials.*

x	0	1	2	3	4
y	1	1·8	3·3	4·5	6·3

Solution. Here $N = 5$. Let us transform to the variable

$$\xi = \frac{x-2}{1} = x - 2 \text{ so that } \Sigma\xi = 0$$

Let the orthogonal polynomial form of straight line (*) be

$$y = b_0 + b_1 P_1(x) = b_0 + b_1 \phi_1(x) \qquad \text{.(**)}$$

x	$\xi = x-2$	y	ϕ_1	$y\phi_1$
0	−2	1	−2	−2
1	−1	1·8	−1	−1·8
2	0	3·3	0	0
3	1	4·5	1	4·5
4	2	6·3	2	12·6
Total		16·9	0	13·3

The values of ϕ_1 are noted from the tables for $N = 5$. From tables we also find

$$\Sigma \phi_1^2 = 10, \ \lambda_1 = 1$$

Now using (9·40), $b_0 = \dfrac{\Sigma y}{N} = \dfrac{16·9}{5} = 3·38$; $b_1 = \dfrac{\Sigma y \phi_1}{\Sigma \phi_1^2} = \dfrac{13·3}{10} = 1·33$

$$\phi_1(x) = \lambda_1 \xi = 1 \,.\, (x-2) = x-2$$

Substituting in (**), the required straight line is

$$y = 3·38 + 1·33\,(x-2)$$

$$\Rightarrow \qquad y = 1·33x + 0·72$$

Example 9·9. *Fit a second degree parabola to the following data, using the method of orthogonal polynomials.*

x	0·5	1·0	1·5	2·0	2·5	3·0
y	72	110	158	214	290	380

Solution. Let the second degree parabola be

$$y = a + bx + cx^2 \qquad \ldots(*)$$

and its orthogonal polynomial transform be :

$$y = b_0 + b_1 \phi_1(x) + b_2 \phi_2(x) \qquad \ldots(**)$$

Here we have $N = 6$. Let us transform to

$$\xi = \frac{x - \frac{1}{2}(1·5 + 2·0)}{\frac{1}{2}(0·5)} = 4\,(x - 1·75) = 4x - 7,$$

so that $\Sigma \xi = 0$. From Fisher's tables we note the values of ϕ_1 and ϕ_2 (as given in the following table) and also

$\Sigma \phi_1^2 = 70, \Sigma \phi_2^2 = 84$; $\lambda_1 = 2, \lambda_2 = 3/2$

x	$\xi = 4x-7$	y	ϕ_1	ϕ_2	$y\phi_1$	$y\phi_2$
0·5	−5	72	−5	5	−360	360
1·0	−3	110	−3	−1	−330	−110

1·5	−1	158	−1	−4	−158.	−632
2·0	1	214	1	−4	214	−856
2·5	3	290	3	−1	870	−290
3·0	5	380	5	5	1900	1900
Total		1224			2136	372

$$b_0 = \frac{\Sigma y}{N} = \frac{1224}{6} = 204\ ;\ b_1 = \frac{\Sigma y\,\phi_1}{\Sigma \phi_1^{\,2}} = \frac{2136}{70} = 30{\cdot}51$$

$$b_2 = \frac{\Sigma y\,\phi_2}{\Sigma \phi_2^{\,2}} = \frac{372}{84} = 4{\cdot}43\ ;\ \phi_1(x) = \lambda_1 \xi = 2[4x - 7] = 8x - 14$$

$$\phi_2(x) = \lambda_2 \left[\xi^2 - \frac{N^2 - 1}{12} \right] = \frac{3}{2} \left[(4x-7)^2 - \frac{36-1}{12} \right]$$

$$= \frac{3}{2} \left[16x^2 + 49 - 56x - \frac{35}{12} \right] = 24x^2 - 84x + 69{\cdot}125$$

Substituting in (**), we get

$$y = 204 + 30{\cdot}51(8x - 14) + 4{\cdot}43\,(24x^2 - 84x + 69{\cdot}125)$$
$$= 106{\cdot}32x^2 - 128{\cdot}04x + 83{\cdot}08$$

which is the required second degree parabola of best fit.

CHAPTER TEN

Correlation and Regression

10·1. Bivariate Distribution, Correlation. So far we have confined ourselves to unvariate distributions, *i.e.*, the distributions involving only one variable. We may, however, come across certain series where each term of the series may assume the values of two or more variables. For example, if we measure the heights and weights of a certain group of persons, we shall get what is known as *Bivariate distribution*—one variable relating to height and other variable relating to weight.

In a bivariate distribution we may be interested to find out if there is any correlation or covariation between the two variables under study. If the change in one variable affects a change in the other variable, the variables are said to be correlated. If the two variables deviate in the same direction, *i.e.*, if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be *direct* or *positive*. But if they constantly deviate in the opposite directions, *i.e.*, if increase (or decrease) in one results in corresponding decrease (or increase) in the other, correlation is said to be *diverse* or *negative*. For example, the correlation between (*i*) the heights and weights of a group of persons, (*ii*) the income and expenditure is positive and the correlation between (*i*) price and demand of a commodity, (*ii*) the volume and pressure of a perfect gas, is negative. Correlation is said to be *perfect* if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

10·2. Scatter Diagram. It is the simplest way of the diagrammatic representation of bivariate data. Thus for the bivariate distribution (x_i, y_i); $i = 1, 2, \ldots, n$, if the values of the variables X and Y be plotted along the x-axis and y-axis respectively in the xy plane, the diagram of dots so obtained is known as *scatter diagram*. From the scatter diagram, we can form a fairly good, though vague, idea whether the variables are correlated or not, *e.g.*, if the points are very dense, *i.e.*, very close to each other, we should expect a fairly good amount of correlation between the variables and if the points are widely scattered, a poor correlation is expected. This method, however, is not suitable if the number of observations is fairly large.

10·3. Karl Pearson Coefficient of Correlation. As a measure of itensity or degree of linear relationship between two variables, Karl Pearson (1867—1936), *a British Biometrician*, developed a formula called *Correlation Coefficient*.

Correlation coefficient between two random variables X and Y, usually denoted by $r(X, Y)$ or simply r_{XY}, is a numerical measure of *linear relationship* between them and is defined as

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \qquad (10\cdot1)$$

If (x_i, y_i) ; $i = 1, 2, ..., n$ is the bivariate distribution, then

$$\left.\begin{aligned} \text{Cov}(X, Y) &= E[\{X - E(X)\}\{(Y - E(Y)\}] \\ &= \frac{1}{n}\Sigma)x_i - \bar{x})(y_i - \bar{y}) = \mu_{11} \\ \sigma_X^2 &= E\{X - E(X)\}^2 = \frac{1}{n}\Sigma (x_i - \bar{x})^2 \\ \sigma_Y^2 &= E\{Y - E(Y)\}^2 = \frac{1}{n}\Sigma (y_i - \bar{y})^2 \end{aligned}\right\}, \quad \ldots (10{\cdot}2)$$

the summation extending over i from 1 to n.

Another convenient form of the formula (10·2) for computational work is as follows :

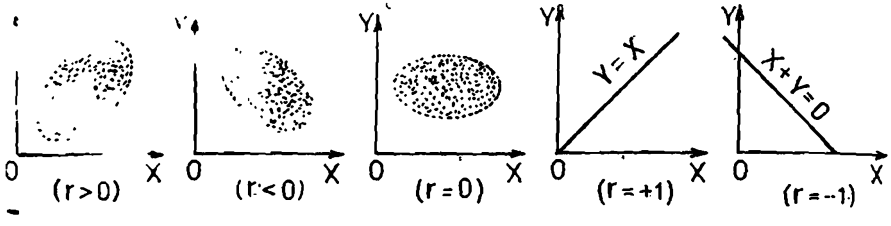
$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}(x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})$$

$$= \frac{1}{n}\Sigma x_i y_i - \bar{y}\frac{1}{n}\Sigma x_i - \bar{x}\frac{1}{n}\Sigma y_i + \bar{x}\bar{y}$$

$$\therefore \quad \text{Cov}(X, Y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y}, \; \sigma_X^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2$$

and $$\sigma_Y^2 = \frac{1}{n}\Sigma y_i^2 - \bar{y}^2 \qquad \ldots (10.2a)$$

Remarks 1. Following are the figures of the standard data for $r > 0, < 0, = 0$, and $r = \pm 1$.



2. It may be noted that $r(X, Y)$ provides a measure of *linear relationship* between X and Y. For nonlinear relationship, however, it is not very suitable.

3. Sometimes, we write : $\text{Cov}(X, Y) = \sigma_{XY}$

4. Karl Pearson's correlation coefficient is also called *product-moment correlation coefficient,* since

$\text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = \mu_{11}$.

10·3·1. Limits for Correlation Coefficient. We have

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\left[\frac{1}{n}\Sigma (x_i - \bar{x})^2 \cdot \frac{1}{n}\Sigma (y_i - \bar{y})^2\right]^{1/2}},$$

$$\therefore \quad r^2(X, Y) = \frac{(\sum a_i b_i)^2}{(\sum a_i^2)(\sum b_i^2)}, \text{ where } \begin{pmatrix} a_i = x_i - \bar{x} \\ b_i = y_i - \bar{y} \end{pmatrix} \qquad \ldots(*)$$

We have the Schwartz inequality which states that if $a_i, b_i; i = 1, 2, \ldots, n$ are real quantities then

$$(\sum_{i=1}^{n} a_i b_i)^2 \le (\sum_{i=1}^{n} a_i^2)(\sum_{i=1}^{n} b_i^2)$$

the sign of equality holding if and only if

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \ldots = \frac{a_n}{b_n}$$

Using Schwartz inequality, we get from (*)

$$r^2(X, Y) \le 1 \text{ i.e., } |r(X, Y)| \le 1 \Rightarrow -1 \le r(X, Y) \le 1 \qquad \ldots(10\text{·}3)$$

Hence correlation coefficient cannot exceed unity numerically. It always lies between –1 and +1. If $r = +1$, the correlation is perfect and positive and if $r = -1$, correlation is perfect and negative.

Aliter. If we write $E(X) = \mu_X$ and $E(Y) = \mu_Y$, then we have

$$E\left[\left(\frac{X - \mu_X}{\sigma_X}\right) \pm \left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right]^2 \ge 0$$

$$\Rightarrow \quad E\left(\frac{X - \mu_X}{\sigma_X}\right)^2 + E\left(\frac{Y - \mu_Y}{\sigma_Y}\right)^2 \pm 2\frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \ge 0$$

$$\Rightarrow \quad 1 + 1 \pm 2r(X, Y) \ge 0$$

$$\Rightarrow \quad -1 \le r(X, Y) \le 1.$$

Theorem 10·1. *Correlation coefficient is independent of change of origin and scale.*

Proof. Let $U = \frac{X - a}{h}$, $V = \frac{Y - b}{k}$, so that $X = a + hU$ and $Y = b + kV$, where a, b, h, k are constants; $h > 0, k > 0$.

We shall prove that $r(X, Y) = r(U, V)$

Since $X = a + hU$ and $Y = b + kV$, on taking expectations, we get

$$E(X) = a + hE(U) \quad \text{and} \quad E(Y) = b + kE(V)$$

$$\therefore \quad X - E(X) = h[U - E(U)] \quad \text{and} \quad Y - E(Y) = k[V - E(V)]$$

$$\Rightarrow \text{Cov } (X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$

$$= E[h\{U - E(U)\}\{k\{V - E(V)\}]$$

$$= hk\, E[\{U - E(U)\}\{V - E(V)\}] = hk \text{ Cov } (U, V) \qquad \ldots(10\text{·}4)$$

$$\sigma_X^2 = E[\{X - E(X)\}^2] = E[h^2\{U - E(U)\}^2] = h^2\sigma_U^2$$

$$\Rightarrow \quad \sigma_X = h\sigma_U, (h > 0) \qquad \ldots(10\text{·}4a)$$

$$\text{and} \quad \sigma_Y^2 = E[\{Y - E(Y)\}^2] = E[k^2\{V - E(V)\}^2] = k^2\sigma_V^2$$

$$\Rightarrow \quad \sigma_Y = k\sigma_V, (k > 0) \qquad \ldots(10\text{·}4b)$$

Substituting from (10·4), (10·4*a*) and (10·4*b*) in (10·1), we get

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{hk.\ \text{Cov}(U, V)}{hk.\sigma_U \sigma_V} = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = r(U, V)$$

This theorem is of fundamental importance in the numerical computation of the correlation coefficient.

Corollary. *If X and Y are random variables and a, b, c, d are any numbers provided only that $a \neq 0$, $c \neq 0$, then*

$$r(aX + b, cY + d) = \frac{ac}{|ac|} r(X, Y)$$

Proof. With usual notations, we have

$$\text{Var}(aX + b) = a^2\sigma_X^2; \quad \text{Var}(cY + d) = c^2\sigma_Y^2;$$

$$\text{Cov}(aX + b, cY + d) = ac\sigma_{XY}$$

$$\therefore \quad r(aX + b, cY + d) = \frac{\text{Cov}(aX + b, cY + d)}{[\text{Var}(aX + b)\ \text{Var}(cY + d)]^{1/2}}$$

$$= \frac{ac\,\sigma_{XY}}{|a|\,|c|\,\sigma_X \sigma_Y} = \frac{ac}{|ac|} r(X, Y)$$

If $ac > 0$, *i.e.*, if a and c are of same signs, then $ac/|ac| = +1$

If $ac < 0$, *i.e.*, if a and c are of opposite signs, then $ac/|ac| = -1$.

Theorem 10·2. *Two independent variables are uncorrelated.*

Proof. If X and Y are independent variables, then

$$\text{Cov}(X, Y) = 0 \qquad (cf.\ \S\ 6\text{·}4)$$

$$\therefore \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

Hence two independent variables are uncorrelated.

But the converse of the theorem is not true, *i.e.*, two uncorrelated variables may not be independent as the following example illustrates :

X	−3	−2	−1	1	2	3	Total $\Sigma X = 0$
Y	9	4	1	1	4	9	$\Sigma Y = 28$
XY	−27	− 8	−1	1	8	27	$\Sigma XY = 0$

$$\bar{X} = \frac{1}{n}\Sigma X = 0, \text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\,\bar{Y} = 0$$

$$\therefore \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

Thus in the above example, the variables X and Y are uncorrelated. But on careful examination we find that X and Y are not independent but they are connected by the relation $Y = X^2$. Hence two uncorrelated variables need not necessarily be independent. A simple reasoning for this strange conclusion is that $r(X, Y) = 0$, merely implies the absence of any linear relationship between

the variables X and Y. There may, however, exist some other form of relationship between them, *e.g.*, quadratic, cubic or trigonometric.

Remarks. 1. Following are some more examples where two variables are uncorrelated but *not* independent.

(*i*) $X \sim N(0, 1)$ and $Y = X^2$

Since $X \sim N(0, 1)$, $E(X) = 0 = E(X^3)$

$$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)\,E(Y)$$

$$= E(X^3) - E(X)\,E(Y) = 0 \qquad (\because Y = X^2)$$

$$\Rightarrow \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\,\sigma_Y} = 0$$

Hence X and Y are uncorrelated but not independent.

(*ii*) Let X be a r.v. with p.d.f.

$$f(x) = \frac{1}{2}, -1 \le x \le 1$$

and let $Y = X^2$. Here we shall get

$$E(X) = 0 \text{ and } E(XY) = E(X^3) = 0, \Rightarrow r(X, Y) = 0$$

2. However, the converse of the theorem holds in the following cases :

(*a*) If X and Y are jointly normally distributed with $\rho = \rho(X, Y) = 0$, then they are independent. If $\rho = 0$, then [*c.f.* § 10·10, Equation (10·25)]

$$f(x, y) = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{X - \mu_X}{\sigma_X}\right)^2\right] \times \frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y - \mu_Y}{\sigma_Y}\right)^2\right]$$

$$\therefore \quad f(x, y) = f_1(x)\,f_2(y)$$

$$\Rightarrow \quad X \text{ and } Y \text{ are independent.}$$

(*b*) *If each of the two variables X and Y takes two values, 0, 1 with positive probabilities, then r (X, Y) = 0 ⇒ X and Y are independent.*

Proof. Let X take the values 1 and 0 with positive probabilities p_1 and q_1 respectively and let Y take the values 1 and 0 with positive probabilities p_2 and q_2 respectively. Then

$$r(X, Y) = 0 \quad \Rightarrow \quad \text{Cov}(X, Y) = 0$$

$$\Rightarrow \quad 0 = E(XY) - E(X)E(Y)$$

$$= 1 \,.\, P(X = 1 \cap Y = 1) - [1 \,.\, P(X) = 1) \times 1 \,.\, P(Y = 1)]$$

$$= P(X = 1 \cap Y = 1) - p_1p_2$$

$$\Rightarrow \quad P(X = 1 \cap Y = 1) = p_1p_2 = P(X = 1) \,.\, P(Y = 1)$$

$$\Rightarrow \quad X \text{ and } Y \text{ are independent.}$$

10·3·2. Assumptions Underlying Karl Pearson's Correlation Coefficient. Pearsonian correlation coefficient r is based on the following assumptions :

(*i*) *The variables X and Y under study are linearly related.* In other words, the scatter diagram of the data will give a straight line curve.

(*ii*) *Each of the variables (series) is being affected by a large number of independent contributory causes of such a nature as to produce normal distribution.* For example, the variables (series) relating to ages, heights, weight, supply, price, etc., conform to this assumption. In the words of Karl Pearson :

"The sizes of the complex of organs (something measurable) are determined by a great variety of independent contributory causes, for example, climate, nourishment, physical training and innumerable other causes which cannot be individually observed or their effects measured." Karl Pearson further observes, "The variations in intensity of the contributory causes are small as compared with their absolute intensity and these variations follow the normal law of distribution."

(*iii*) *The forces so operating on each of the variable series are not independent of each other but are related in a causal fashion.* In other word, cause and effect relationship exists between different forces operating on the items of the two variable series. These forces must be common to both the series. If the operating forces are entirely independent of each other and not related in any fashion, then there cannot be any correlation between the variables under study.

For example, the correlation coefficient between,

(*a*) the series of heights and incomes of individuals over a period of time,

(*b*) the series of marriage rate and the rate of agricultural production in a country over a period of time,

(*c*) the series relating to the size of the shoe and intelligence of a group of individuals,

should be zero, since the forces affecting the two variable series in each of the above cases are entirely independent of each other.

However, if in any of the above cases the value of r for a given set of data is not zero, then such correlation is termed as *chance correlation* or *spurious* or *non-sense correlation.*

Example 10·1. *Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y) :*

X :	65	66	67	67	68	69	70	72
Y :	67	68	65	68	72	72	69	71

Solution.

CALCULATIONS FOR CORRELATION COEFFICIENT

	X	Y	X^2	Y^2	XY
	65	67	4225	4489	4355
	66	68	4356	4624	4488
	67	65	4489	4225	4355
	67	68	4489	4624	4556
	68	72	4624	5184	4896
	69	72	4761	5184	4968
	70	69	4900	4761	4830
	72	71	5184	5041	5112
Total	544	552	37028	38132	37560

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{544}{8} = 68, \bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{8} \times 552 = 69$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma XY - \bar{X}\,\bar{Y}}{\sqrt{\left(\frac{1}{n}\Sigma X^2 - \bar{X}^2\right)\left(\frac{1}{n}\Sigma Y^2 - \bar{Y}^2\right)}}$$

$$= \frac{\frac{1}{8} \times 37560 - 68 \times 69}{\sqrt{\left[\frac{37028}{8} - (68)^2\right]\left[\frac{38132}{8} - (69)^2\right]}}$$

$$= \frac{4695 - 4692}{\sqrt{(4628{\cdot}5 - 4624)\,(4766{\cdot}5 - 4761)}} = \frac{3}{\sqrt{4{\cdot}5 \times 5{\cdot}5}} = 0{\cdot}603$$

Aliter.

(SHORT-CUT METHOD)

X	Y	$U = X - 68$	$V = Y - 69$	U^2	V^2	UV
65	67	−3	− 2	9	4	6
66	68	−2	− 1	4	1	2
67	65	−1	− 4	1	16	4
67	68	−1	− 1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
Total		0	0	36	44	24

$$\bar{U} = \frac{1}{n}\Sigma U = 0, \quad \bar{V} = \frac{1}{n}\Sigma V = 0$$

$$\text{Cov}(U, V) = \frac{1}{n}\Sigma UV - \bar{U}\,\bar{V} = \frac{1}{8} \times 24 = 3$$

$$\sigma_U^2 = \frac{1}{n}\Sigma U^2 - (\bar{U})^2 = \frac{1}{8} \times 36 = 4{\cdot}5$$

$$\sigma_V^2 = \frac{1}{n}\Sigma V^2 - (\bar{V})^2 = \frac{1}{8} \times 44 = 5{\cdot}5$$

$$\therefore \quad r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{3}{\sqrt{4{\cdot}5 \times 5{\cdot}5}} = 0{\cdot}603 = r(X, Y)$$

Remark. The reader is advised to calculate the correlation coefficient by arbitrary origin method rather than by the direct method; since the latter leads to much simpler arithmetical calculations.

Example 10·2. *A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results :*

$$n = 25,\ \Sigma X = 125,\ \Sigma X^2 = 650,\ \Sigma Y = 100,\ \Sigma Y^2 = 460,\ \Sigma XY = 508$$

It was, however, later discovered at the time of checking that he had copied down two pairs as $\begin{array}{c|c} X & Y \\ \hline 6 & 14 \\ 8 & 6 \end{array}$ *while the correct values were* $\begin{array}{c|c} X & Y \\ \hline 8 & 12 \\ 6 & 8 \end{array}$

Obtain the correct value of correlation coefficient.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1988, 1991*]

Solution.

$$\text{Corrected } \Sigma X = 125 - 6 - 8 + 8 + 6 = 125$$
$$\text{Corrected } \Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$$
$$\text{Corrected } \Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$$
$$\text{Corrected } \Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$$
$$\text{Corrected } \Sigma XY = 508 - 6\times14 - 8\times6 + 8\times12 + 6\times8 = 520$$

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{1}{25}\times 125 = 5,\quad \bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{25}\times 100 = 4$$

$$\text{Cov}\,(X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\bar{Y} = \frac{1}{25}\times 520 - 5\times 4 = \frac{4}{5}$$

$$\sigma_X^2 = \frac{1}{n}\Sigma X^2 - \bar{X}^2 = \frac{1}{25}\times 650 - (5)^2 = 1$$

$$\sigma_Y^2 = \frac{1}{n}\Sigma Y^2 - \bar{Y}^2 = \frac{1}{25}\times 436 - 16 = \frac{36}{25}$$

$$\therefore\ \text{Corrected } r(X, Y) = \frac{\text{Cov}\,(X, Y)}{\sigma_X\,\sigma_Y} = \frac{\frac{4}{5}}{1\times\frac{6}{5}} = \frac{2}{3} = 0{\cdot}67$$

Example 10·3. *Show that if X′, Y′ are the deviations of the random variables X and Y from their respective means then*

(*i*) $$r = 1 - \frac{1}{2N}\sum_i \left(\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y}\right)^2$$

(*ii*) $$r = -1 + \frac{1}{2N}\sum_i \left(\frac{X_i'}{\sigma_X} + \frac{Y_i'}{\sigma_Y}\right)^2$$

Deduce that $-1 \le r \le +1$.

[*Delhi Univ. B.Sc. Oct. 1992; Madras Univ. B.Sc., Nov. 1991*]

Solution. (*i*) Here $X_i' = (x_i - \bar{X})$ and $Y_i' = (Y_i - \bar{Y})$

$$\text{R.H.S.} = 1 - \frac{1}{2N}\sum_i \left(\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y}\right)^2$$

$$= 1 - \frac{1}{2N}\sum_i \left[\frac{X_i'^2}{\sigma_X^2} + \frac{Y_i'^2}{\sigma_Y^2} - \frac{2X_i'Y_i'}{\sigma_X\sigma_Y}\right]$$

$$= 1 - \frac{1}{2N}\left[\frac{1}{\sigma_X^2}\sum_i X_i'^2 + \frac{1}{\sigma_Y^2}\sum_i Y_i'^2 - \frac{2}{\sigma_X\sigma_Y}\sum_i X_i'Y_i'\right]$$

$$= 1 - \frac{1}{2N}\left[\frac{1}{\sigma_X^2}\sum_i (X_i - \bar{X})^2 + \frac{1}{\sigma_Y^2}\sum_i (Y_i - \bar{Y})^2 - \frac{2}{\sigma_X\sigma_Y}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})\right]$$

$$= 1 - \frac{1}{2}\left[\frac{1}{\sigma_X^2}\cdot\sigma_X^2 + \frac{1}{\sigma_Y^2}\cdot\sigma_Y^2 - \frac{2}{\sigma_X\sigma_Y}\cdot r\sigma_X\sigma_Y\right]$$

$$= 1 - \frac{1}{2}[1 + 1 - 2r] = r$$

(*ii*) Proceeding similarly, we will get

$$\text{R.H.S.} = -1 + \frac{1}{2}(1 + 1 + 2r) = r$$

Deduction. Since $\left(\frac{X_i'}{\sigma_X} \pm \frac{Y_i'}{\sigma_Y}\right)^2$, being the square of a real quantity is always non-negative, $\sum_i \left(\frac{X_i'}{\sigma_X} \mp \frac{Y_i'}{\sigma_Y}\right)^2$ is also non-negative. From part (*i*), we get

$$r = 1 - (\text{some non-negative quantity}) \quad \Rightarrow \quad r \le 1 \qquad \ldots(*)$$

Also from part (*ii*), we get

$$r = -1 + (\text{some non-negative quantity}) \quad \Rightarrow \quad -1 \le r \qquad \ldots(**)$$

The sign of equality in (*) and (**) holds if and only if

$$\left.\begin{aligned}\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y} &= 0\\ \text{and} \quad \frac{X_i'}{\sigma_X} + \frac{Y_i'}{\sigma_Y} &= 0\end{aligned}\right\} \forall\, i = 1, 2, \ldots, n$$

respectively.

From (*) and (**), we get

$$-1 \le r \le 1$$

Example 10·4. *The variables X and Y are connected by the equation* $aX + bY + c = 0$. *Show that the correlation between them is –1 if the signs of a and b are alike and +1 if they are different.*

[***Nagpur Univ. B.Sc. 1992; Delhi Univ. B.Sc. (Stat. Hons.) 1992***]

Solution. $aX + bY + c = 0 \Rightarrow aE(X) + bE(Y) + c = 0$

$$\therefore \quad a\{X - E(X)\} + b\{Y - E(Y)\} = 0$$

$$\Rightarrow \quad \{X - E(X)\} = -\frac{b}{a}\{Y - E(Y)\}$$

$$\therefore \quad \text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$

$$= -\frac{b}{a} E[\{Y - E(Y)\}^2] = -\frac{b}{a} \cdot \sigma_Y^2$$

$$E\{X - E(X)\}^2 = \frac{b^2}{a^2} E[\{Y - E(Y)\}^2] = \frac{b^2}{a^2} \cdot \sigma_Y^2$$

$$\therefore \quad r = \frac{-\frac{b}{a} \cdot \sigma_Y^2}{\sqrt{\sigma_Y^2}\sqrt{\frac{b^2}{a^2} \cdot \sigma_Y^2}} = \frac{-\frac{b}{a}\sigma_Y^2}{\left|\frac{b}{a}\right| \sigma_Y^2}$$

$$= \begin{cases} +1, \text{ if } b \text{ and } a \text{ are of opposite signs.} \\ -1, \text{ if } b \text{ and } a \text{ are of same sign.} \end{cases}$$

Example 10·5. (a) *If $Z = aX + bY$ and r is the correlation coefficient between X and Y, show that*

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X\sigma_Y$$

(b) *Show that the correlation coefficient r between two random variables X and Y is given by*

$$r = (\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2) / 2\sigma_X\sigma_Y$$

where σ_X, σ_Y and σ_{X-Y} are the standard deviations of X, Y and $X - Y$ respectively.

[***Calcutta Univ. B.Sc., 1992; M.S. Baroda Univ. B.Sc. 1992***]

Solution. Taking expectation of both sides of $Z = aX + bY$, we get

$$E(Z) = aE(X) + bE(Y)$$

$$\therefore \quad Z - E(Z) = a\{X - E(X)\} + b\{Y - E(Y)\}$$

Squaring and taking expectation of both sides, we get

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab \text{ Cov } (X, Y)$$

$$= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X\sigma_Y$$

(*b*) Taking $a = 1$, $b = -1$ in the above case, we have

$$Z = X - Y \text{ and } \sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X\sigma_Y$$

$$\therefore \quad r = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2\sigma_X\sigma_Y}$$

Remark. In the above example, we have obtained

$$V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2ab \text{ Cov } (X, Y)$$

Similarly, we could obtain the result

$$V(aX - bY) = a^2 V(X) + b^2 V(Y) - 2ab \text{ Cov } (X, Y).$$

The above results are useful in solving theoretical problems.

Example 10·6. *X and Y are two random variables with variances σ_X^2 and σ_Y^2 respectively and r is the coefficient of correlation between them. If $U = X + kY$ and $V = X + \frac{\sigma_X}{\sigma_Y} Y$, find the value of k so that U and V are uncorrelated.* [***Delhi Univ. B.Sc. 1992; Andhra Univ. B.Sc. 1993***]

Solution. Taking expectations of $U = X + kY$ and $V = X + \frac{\sigma_X}{\sigma_Y} Y$, we get

$$E(U) = E(X) + kE(Y) \text{ and } E(V) = E(X) + \frac{\sigma_X}{\sigma_Y} E(Y)$$

$$U - E(U) = \{X - E(X)\} + k\{Y - E(Y)\} \text{ and}$$

$$V - E(V) = \{X - E(X)\} + \frac{\sigma_X}{\sigma_Y} \{Y - E(Y)\}$$

$$\text{Cov}(U, V) = E[\{U - E(U)\}\{V - E(V)\}]$$

$$= E\left[\{X - E(X)\} + k(Y - E(Y))\right] \times \left[\{X - E(X)\} + \frac{\sigma_X}{\sigma_Y}\{Y - E(Y)\}\right]$$

$$= \sigma_X^2 + \frac{\sigma_X}{\sigma_Y} \text{Cov}(X, Y) + k \text{ Cov}(X, Y) + k \frac{\sigma_X}{\sigma_Y} \cdot \sigma_Y^2$$

$$= \left[\sigma_X^2 + k\sigma_X\sigma_Y\right] + \left[\frac{\sigma_X}{\sigma_Y} + k\right] \text{Cov}(X, Y)$$

$$= \sigma_X(\sigma_X + k\sigma_Y) + \left[\frac{\sigma_X + k\sigma_Y}{\sigma_Y}\right] \text{Cov}(X, Y)$$

$$= (\sigma_X + k\sigma_Y)\left[\sigma_X + \frac{\text{Cov}(X, Y)}{\sigma_Y}\right] = (\sigma_X + k\sigma_Y)(1 + r)\sigma_X$$

U and V will be uncorrelated if

$$r(U, V) = 0 \Rightarrow \text{Cov}(U, V) = 0$$

i.e., if $(\sigma_X + k\sigma_Y)(1 + r)\sigma_X = 0$

$\Rightarrow$ $\sigma_X + k\sigma_Y = 0$ $(\because \sigma_X \neq 0, r \neq -1)$

$\Rightarrow$ $k = -\frac{\sigma_X}{\sigma_Y}$

Example 10·7. *The random variables X and Y are jointly normally distributed and U and V are defined by*

$$U = X \cos \alpha + Y \sin \alpha,$$

$$V = Y \cos \alpha - X \sin \alpha$$

Show that U and V will be uncorrelated if

$$\tan 2\alpha = \frac{2r\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2},$$

where r = corr. (X, Y), σ_X^2 = Var (X) and σ_Y^2 = Var (Y). Are U and V then independent?

[*Delhi Univ. B.Sc. (Stat. Hons.) 1989; (Maths. Hons.), 1990*]

Solution. We have

$$\text{Cov}(U, V) = E[\{U - E(U)\}\{V - E(V)\}]$$

$$= E\Big[[\{X - E(X)\} \cos \alpha + \{Y - E(Y)\} \sin \alpha]$$

$$\times [\{Y - E(Y)\} \cos \alpha - \{X - E(X)\} \sin \alpha]\Big]$$

$$= \cos^2\alpha \text{ Cov } (X, Y) - \sin\alpha \cos\alpha.\sigma_X^2 + \sin\alpha \cos\alpha.\sigma_Y^2 - \sin^2\alpha \text{ (Cov } (X, Y)$$

$$= (\cos^2\alpha - \sin^2\alpha) \text{ Cov } (X, Y) - \sin\alpha \cos\alpha \ (\sigma_X^2 - \sigma_Y^2)$$

$$= \cos 2\alpha \text{ Cov } (X, Y) - \sin\alpha \cos\alpha \ (\sigma_X^2 - \sigma_Y^2)$$

U and V will be uncorrelated if and only if

$$r(U, V) = 0, \quad i.e., \quad \text{iff Cov } (U, V) = 0$$

i.e., if $\cos 2\alpha \text{ Cov } (X, Y) - \sin\alpha \cos\alpha \ (\sigma_X^2 - \sigma_Y^2) = 0$

or if $$\cos 2\alpha \, r\,\sigma_X\sigma_Y = \frac{\sin 2\alpha}{2} \cdot (\sigma_X^2 - \sigma_Y^2)$$

or if $$\tan 2\alpha = \frac{2r\,\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2}$$

However, $r(U, V) = 0$ does not imply that the variables U and V are independent. [For detailed discussion, see Theorem 10·2, page 10-4.].

Example 10·8. *If X, Y are standardized random variables, and*

$$r(aX + bY, bX + aY) = \frac{1 + 2ab}{a^2 + b^2} \qquad \ldots(*)$$

find $r(X, Y)$, the coefficient of correlation between X and Y.

[***Sardar Patel Univ. B.Sc., 1993; Delhi Univ. B.Sc. (Stat. Hons.), 1989***]

Solution. Since X and Y are standardised random variables, we have

$$\left.\begin{array}{rrcl} & E(X) = E(Y) = 0 & & \\ \text{and} \quad \text{Var } (X) = \text{Var } (Y) = 1 & \Rightarrow & E(X^2) = E(Y^2) = 1 & \\ \text{and} \quad \text{Cov}(X, Y) = E(XY) & \Rightarrow & E(XY) = r(X,Y).\sigma_X\sigma_Y & \\ & & = r(X,Y) & \end{array}\right] \qquad \ldots(**)$$

Also we have

$$r(aX + bY, bX + aY)$$

$$= \frac{E[(aX + bY)(bX + aY)] - E(aX + bY)\,E(bX + aY)}{[\text{Var } (aX + bY) \cdot \text{Var } (bX + aY)]^{1/2}}$$

$$= \frac{E[abX^2 + a^2XY + b^2YX + abY^2] - 0}{\{[a^2 \text{ Var } (X) + b^2 \text{ Var } (Y) + 2ab \text{ Cov } (X,Y)] \times [b^2 \text{ Var } (X) + a^2 \text{ Var } Y + 2ba \text{ Cov } (X,Y)]\}^{1/2}}$$

$$= \frac{ab.1 + a^2 r(X, Y) + b^2 r(X, Y) + ab.1}{\{[a^2 + b^2 + 2ab\, r(X, Y)][b^2 + a^2 + 2ba\, r(X, Y)]\}^{1/2}}$$

[Using (**)]

$$= \frac{2ab + (a^2 + b^2).\, r(X, Y)}{a^2 + b^2 + 2ab.\, r(X, Y)}$$

From (*) and (**), we get

$$\frac{1 + 2ab}{a^2 + b^2} = \frac{(a^2 + b^2).\, r(X, Y) + 2ab}{a^2 + b^2 + 2ab.\, r(X, Y)}$$

Cross multiplying, we get

$$(a^2+b^2)(1+2ab)+2ab.\,r(X,Y)(1+2ab)=(a^2+b^2)^2.\,r(X,Y)+2ab\,(a^2+b^2)$$

$$\Rightarrow \quad (a^4+b^4+2a^2b^2-2ab-4a^2b^2).\,r(X,Y)=(a^2+b^2)$$

$$\Rightarrow \quad [(a^2-b^2)^2-2ab]\,r(X,Y)=a^2+b^2$$

$$\Rightarrow \quad r(X,Y)=\frac{a^2+b^2}{(a^2-b^2)^2-2ab}$$

Example 10·9. *If X and Y are uncorrelated random variables with means zero and variances σ_1^2 and σ_2^2 respectively, show that*

$$U=X\cos\alpha+Y\sin\alpha,\ V=X\sin\alpha-Y\cos\alpha$$

have a correlation coefficient ρ given by

$$\rho=\frac{\sigma_1^2-\sigma_2^2}{[(\sigma_1^2-\sigma_2^2)^2+4\sigma_1^2\sigma_2^2\,\mathrm{cosec}^2\,2\alpha]^{1/2}}$$

Solution. We are given that

$$r(X,Y)=0 \Rightarrow \mathrm{Cov}(X,Y)=0,\ \sigma_X^2=\sigma_1^2 \text{ and } \sigma_Y^2=\sigma_2^2 \qquad \ldots(1)$$

We have

$$\begin{aligned}\sigma_U^2 &= V(X\cos\alpha+Y\sin\alpha)\\ &=\cos^2\alpha\,V(X)+\sin^2\alpha\,V(Y)+2\sin\alpha\cos\alpha\,\mathrm{Cov}(X,Y)\\ &=\cos^2\alpha\,\sigma_1^2+\sin^2\alpha\,\sigma_2^2 \qquad \text{[Using (1)]}\end{aligned}$$

Similarly,

$$\sigma_V^2 = V(X\sin\alpha-Y\cos\alpha)=\sin^2\alpha.\sigma_1^2+\cos^2\alpha.\sigma_2^2$$

$$\begin{aligned}\mathrm{Cov}(U,V) &= E[\{U-E(U)\}\{V-E(V)\}]\\ &= E\Big[\{(X-E(X))\cos\alpha+\{(Y-E(Y))\sin\alpha\}\\ &\qquad\qquad \times\{(X-E(X))\sin\alpha-(Y-E(Y))\cos\alpha\}\Big]\\ &=\sin\alpha\cos\alpha\,V(X)-\cos^2\alpha\,\mathrm{Cov}(X,Y)\\ &\qquad +\sin^2\alpha\,\mathrm{Cov}(X,Y)-\sin\alpha\cos\alpha\,V(Y)\\ &=(\sigma_1^2-\sigma_2^2)\sin\alpha\cos\alpha \qquad \text{[Using (1)]}\end{aligned}$$

Now $$\rho^2=\frac{[\mathrm{Cov}(U,V)]^2}{\sigma_U^2\sigma_V^2}$$

where $$\begin{aligned}\sigma_U^2\sigma_V^2 &= (\cos^2\alpha\,\sigma_1^2+\sin^2\alpha\,\sigma_2^2)(\sin^2\alpha\,\sigma_1^2+\cos^2\alpha\,\sigma_2^2)\\ &=\sin^2\alpha\cos^2\alpha(\sigma_1^4+\sigma_2^4)+\sigma_1^2\sigma_2^2(\cos^4\alpha+\sin^4\alpha)\\ &=\sin^2\alpha\cos^2\alpha(\sigma_1^4+\sigma_2^4)+\sigma_1^2\sigma_2^2[(\sin^2\alpha+\cos^2\alpha)^2-2\sin^2\alpha\cos^2\alpha]\\ &=\sin^2\alpha\cos^2\alpha(\sigma_1^4+\sigma_2^4-2\sigma_1^2\sigma_2^2)+\sigma_1^2\sigma_2^2\\ &=\sin^2\alpha\cos^2\alpha(\sigma_1^2-\sigma_2^2)^2+\sigma_1^2\sigma_2^2\end{aligned}$$

$$\therefore \quad \rho^2=\frac{(\sigma_1^2-\sigma_2^2)^2.\sin^2\alpha\cos^2\alpha}{\sigma_1^2\sigma_2^2+\sin^2\alpha\cos^2\alpha(\sigma_1^2-\sigma_2^2)^2}$$

$$=\frac{\frac{1}{4}(\sigma_1^2-\sigma_2^2)^2\sin^2 2\alpha}{\sigma_1^2\sigma_2^2+\sin^2 2\alpha.\,\frac{1}{4}(\sigma_1^2-\sigma_2^2)^2}$$

$$= \frac{(\sigma_1^2 - \sigma_2^2)^2}{4\sigma_1^2\sigma_2^2 \operatorname{cosec}^2 2\alpha + (\sigma_1^2 - \sigma_2^2)^2}$$

$$\Rightarrow \qquad \rho = \frac{\sigma_1^2 - \sigma_2^2}{[(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_1^2\sigma_2^2 \operatorname{cosec}^2 2\alpha]^{1/2}}$$

Example 10·10. *If $U = aX + bY$ and $V = cX + dY$, where X and Y are measured from their respective means and if r is the correlation coefficient between X and Y, and if U and V are uncorrelated, show that*

$$\sigma_U \sigma_V = (ad - bc)\, \sigma_X \sigma_Y (1 - r^2)^{1/2}$$

[Poona Univ. B.Sc., 1990; Delhi Univ. B.Sc. (Stat. Hons.), 1986]

Solution. We have

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \quad \Rightarrow \quad 1 - r^2 = 1 - \frac{[\text{Cov}(X, Y)]^2}{\sigma_X^2 \sigma_Y^2}$$

$$\Rightarrow \qquad (1 - r^2)\, \sigma_X^2 \sigma_Y^2 = \sigma_X^2 \sigma_Y^2 - [\text{Cov}(X, Y)]^2 \qquad \ldots(*)$$

[This step is suggested by the answer]

$$U = aX + bY\,, \quad V = cX + dY$$

Since X, Y are measured from their means,

$$\left.\begin{aligned} & E(X) = 0 = E(Y) \Rightarrow E(U) = 0 = E(V) \\ \text{and} \quad & \sigma_U^2 = E(U^2)\,; \ \sigma_V^2 = E(V^2) \end{aligned}\right\} \qquad \ldots(**)$$

Also $\quad aX + bY - U = 0 \quad$ and $\quad cX + dY - V = 0$

$$\Rightarrow \qquad \frac{X}{-bV + dU} = \frac{Y}{-cU + aV} = \frac{1}{ad - bc}$$

$$\Rightarrow \qquad \left.\begin{aligned} X &= \frac{1}{ad - bc}(dU - bV) \\ Y &= \frac{1}{ad - bc}(-cU + aV) \end{aligned}\right\} \qquad \ldots(***)$$

$$\therefore \quad \text{Var}(X) = \frac{1}{(ad - bc)^2}[d^2 \sigma_U^2 + b^2 \sigma_V^2 - 2bd\, \text{Cov}(U, V)$$

$$= \frac{1}{(ad - bc)^2}[d^2 \sigma_U^2 + b^2 \sigma_V^2]$$

[Since U, V are uncorrelated $\Rightarrow$ Cov $(U, V) = 0$]

Similarly, we have

$$\text{Var}(Y) = \frac{1}{(ad - bc)^2}(c^2 \sigma_U^2 + a^2 \sigma_V^2)$$

$$\text{Cov}(X, Y) = E(XY) - E(X)\, E(Y) = E(XY) \qquad [\because E(X) = 0 = E(Y)]$$

$$= \frac{1}{(ad - bc)^2} E[(dU - bV)(-cU + aV)] \qquad \text{[From (***)]}$$

$$= \frac{1}{(ad-bc)^2}\,[-cd\,\sigma_U^2 - ab\,\sigma_V^2]$$

[Using (**) and Cov $(U, V) = 0$, given]

$$= \frac{-1}{(ad-bc)^2}\,[cd\,\sigma_U^2 + ab\,\sigma_V^2]$$

Substituting in (*), we get

$$(1-r^2)\,\sigma_X^2\,\sigma_Y^2 = \frac{1}{(ad-bc)^4} \times \left[(d^2\,\sigma_U^2 + b^2\,\sigma_V^2)\,(c^2\,\sigma_U^2 + a^2\,\sigma_V^2) - (cd\sigma_U^2 + ab\,\sigma_V^2)^2\right]$$

$$= \frac{1}{(ad-bc)^4} \times \left[c^2d^2\,\sigma_U^4 + a^2b^2\sigma_V^4 + (a^2d^2 + b^2c^2)\,\sigma_U^2\,\sigma_V^2 - c^2d^2\,\sigma_U^4 - a^2b^2\,\sigma_V^4 - 2abcd\,\sigma_U^2\,\sigma_V^2\right]$$

$$= \frac{1}{(ad-bc)^4}\left[a^2d^2 + b^2c^2 - 2abcd\right]\sigma_U^2\,\sigma_V^2$$

$$= \frac{1}{(ad-bc)^4}\,(ad-bc)^2\,\sigma_U^2\,\sigma_V^2$$

$$= \frac{\sigma_U^2\,\sigma_V^2}{(ad-bc)^2}$$

Cross multiplying and taking square root, we get the required result.

Example 10·11. (a) *Establish the formula* :

$$nr\sigma_X\sigma_Y = n_1r_1\sigma_{X_1}\sigma_{Y_1} + n_2r_2\sigma_{X_2}\,\sigma_{Y_2} + n_1dx_1dy_1 + n_2dx_2dy_2 \qquad ...(10·5)$$

where n_1, n_2 and n are respectively the sizes of the first, second and combined sample : $(\bar{x}_1, \bar{y}_1)$, $(\bar{x}_2, \bar{y}_2)$, $(\bar{x}, \bar{y})$, their means r_1, r_2 and r their coefficients of correlation; $(\sigma_{X_1}, \sigma_{Y_1})$, $(\sigma_{X_2}, \sigma_{Y_2})$, (σ_X, σ_Y) their standard deviations, and

$$dx_1 = \bar{x}_1 - \bar{x} \quad , \quad dy_1 = \bar{y}_1 - \bar{y}$$

$$dx_2 = \bar{x}_2 - \bar{x} \quad , \quad dy_2 = \bar{y}_2 - \bar{y}$$

(b) *Find the correlation co-efficient of combined sample given that*

	Sample I	*Sample II*
Sample size	*100*	*150*
Sample mean $(\bar{x})$	*80*	*72*
Sample mean $(\bar{y})$	*100*	*118*
Sample variance (σ_X^2)	*10*	*12*
Sample variance (σ_Y^2)	*15*	*18*
Correlation coefficient	*0·6*	*0·4*

Solution. (a) Let (x_{1i}, y_{1i}) ; $i = 1, 2, \ldots, n_1$ and (x_{2j}, y_{2j}); $j = 1, 2, \ldots, n_2$, be the two samples of sizes n_1 and n_2 respectively from the bivariate population. Then with the given notations, we have

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}, \quad \bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$$

$$\left.\begin{aligned} n\sigma_X^2 &= n_1(\sigma_{X_1}^2 + dx_1^2) + n_2(\sigma_{X_2}^2 + dx_2^2) \\ n\sigma_Y^2 &= n_1(\sigma_{Y_1}^2 + dy_1^2) + n_2(\sigma_{Y_2}^2 + dy_2^2) \end{aligned}\right\} \quad ...(1)$$

$$r_1 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)(y_{1i} - \bar{y}_1)}{n_1\sigma_{X_1}\sigma_{Y_1}}, \quad r_2 = \frac{\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)(y_{2j} - \bar{y}_2)}{n_2\sigma_{X_2}\sigma_{Y_2}} \quad ...(2)$$

For the pooled sample, we have

$$r = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x})(y_{1i} - \bar{y}) + \sum_{j=1}^{n_2}(x_{2j} - \bar{x})(y_{2j} - \bar{y})}{n\sigma_X\sigma_Y} \quad ...(3)$$

Now

$$\sum_{i=1}^{n_1}(x_{1i} - \bar{x})(y_{1i} - \bar{y}) = \sum_{i=1}^{n_1}\left\{\{(x_{1i} - \bar{x}_1) + (\bar{x}_1 - \bar{x})\}\{(y_{1i} - \bar{y}_1) + (\bar{y}_1 - \bar{y})\}\right\}$$

$$= \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)(y_{1i} - \bar{y}_1) + (\bar{y}_1 - \bar{y})\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)$$

$$+ (\bar{x}_1 - \bar{x})\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1) + n_1(\bar{x}_1 - \bar{x})(\bar{y}_1 - \bar{y})$$

But $\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1) = 0$ and $\sum_{i=1}^{n_1}(y_{1i} - y_1) = 0$,

being the algebraic sum of the deviations from the mean.

$$\therefore \quad \sum_{i=1}^{n_1}(x_{1i} - \bar{x})(y_{1i} - \bar{y}) = n_1 r_1 \sigma_{X_1}\sigma_{Y_1} + n_1\, dx_1 dy_1 \qquad \text{[Using (2)]}$$

Similarly, we will get

$$\sum_{j=1}^{n_2}(x_{2j} - \bar{x})(y_{2j} - \bar{y}) = n_2 r_2 \sigma_{X_2}\sigma_{Y_2} + n_2\, dx_2 dy_2$$

Substituting in (3), we get the required formula.

(*b*) Here we are given :

$$n_1 = 100,\ \bar{x}_1 = 80,\ \bar{y}_1 = 100,\ \sigma_{X_1}^2 = 10,\ \sigma_{Y_1}^2 = 15,\ r_1 = 0{\cdot}6$$

$$n_2 = 150,\ \bar{x}_2 = 72,\ \bar{y}_2 = 118,\ \sigma_{X_2}^2 = 12,\ \sigma_{Y_2}^2 = 18,\ r_2 = 0{\cdot}4$$

$$\therefore \quad \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{100 \times 80 + 150 \times 72}{100 + 150} = 75{\cdot}2$$

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} = \frac{100 \times 100 + 150 \times 118}{100 + 150} = 110{\cdot}8$$

$$dx_1 = \bar{x}_1 - \bar{x} = 4{\cdot}8,\ dy_1 = \bar{y}_1 - \bar{y} = 10{\cdot}8$$

$$dx_2 = \bar{x}_2 - \bar{x} = 3{\cdot}2,\ dy_2 = \bar{y}_2 - \bar{y} = 7{\cdot}2$$

$$n\sigma_X^2 = n_1(\sigma_{X_1}^2 + dx_1^2) + n_2(\sigma_{X_2}^2 + dx_2^2) = 6640$$

$$n\sigma_Y^2 = n_1(\sigma_{Y_1}^2 + dy_1^2) + n_2(\sigma_{Y_2}^2 + dy_2^2) = 23640$$

Substituting these values in the formula and simplifying, we get

$$r = \frac{n_1 r_1 \sigma_{X_1}\sigma_{Y_1} + n_2 r_2 \sigma_{X_2}\sigma_{Y_2} + n_1 dx_1 dy_1 + n_2 dx_2 dy_2}{n\sigma_X\sigma_Y} = 0{\cdot}8186$$

Example 10·12. *The independent variables X and Y are defined by :*

$f(x) = 4ax,\ 0 \le x \le r$; $= 0$, *otherwise* | $f(y) = 4by,\ 0 \le y \le s$; $= 0$, *otherwise*

Show that :

$$Cov\,(U, V) = \frac{b-a}{b+a},$$

where $U = X + Y$ and $V = X - Y$

[*I.I.T. (B. Tech.), Nov. 1992*]

Solution. Since the total area under probability curve is unity (one), we have :

$$\int_0^r f(x)dx = 4a\int_0^r x dx = 1 \quad \Rightarrow \quad 2ar^2 = 1 \Rightarrow \quad a = \frac{1}{2r^2} \qquad ...(i)$$

$$\int_0^r f(y)dy = 4b\int_0^s y dy = 1 \quad \Rightarrow \quad 2bs^2 = 1 \Rightarrow \quad b = \frac{1}{2s^2} \qquad ...(ii)$$

$$\therefore\ f(x) = 4ax = \frac{2x}{r^2},\ 0 \le x \le r\ ;\ \text{and}\ f(y) = 4by = \frac{2y}{s^2},\ 0 \le y \le s \qquad ...(iii)$$

Since X and Y are independent variates,

$$r(X, Y) = 0 \quad \Rightarrow \quad Cov\,(X, Y) = 0 \qquad ...(iv)$$

$$\begin{aligned} Cov\,(U, V) &= Cov\,(X + Y, X - Y) \\ &= Cov\,(X, X) - Cov\,(X, Y) + Cov\,(Y, X) - Cov\,(Y, Y) \\ &= \sigma_X^2 - \sigma_Y^2 \qquad \text{[Using } (iv)\text{]} \end{aligned}$$

$$\begin{aligned} Var\,(U) &= Var\,(X + Y) = Var\,(X) + Var\,(Y) + 2\,Cov\,(X, Y) \\ &= \sigma_X^2 + \sigma_Y^2 \qquad \text{[Using } (iv)\text{]} \end{aligned}$$

$$\begin{aligned} Var\,(Y) &= Var\,(X - Y) = Var\,(X) + Var\,(Y) - 2\,Cov\,(X, Y) \\ &= \sigma_X^2 + \sigma_Y^2 \qquad \text{[Using } (iv)\text{]} \end{aligned}$$

$$\therefore \quad r(U,V) = \frac{\text{Cov}(U,V)}{\sigma_U \sigma_V} = \frac{\sigma_X^2 - \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2} \qquad \ldots (v)$$

We have :

$$E(X) = \int_0^r x f(x)\,dx = \frac{2}{r^2}\int_0^r x^2\,dx = \frac{2r}{3} \qquad \text{[From } (iii)\text{]}$$

$$E(X^2) = \int_0^r x^2 f(x)\,dx = \frac{2}{r^2}\int_0^r x^3 dx = \frac{r^2}{2}$$

$$\therefore \text{ Var } (X) = E(X^2) - [E(X)]^2 = \frac{r^2}{2} - \frac{4r^2}{9} = \frac{r^2}{18} = \frac{1}{36a} \qquad \text{[From } (i)\text{]}$$

Similarly, we shall get

$$E(Y) = \frac{2s}{3},\ E(Y^2) = \frac{s^2}{2} \text{ and Var } (Y) = \frac{s^2}{18} = \frac{1}{36b}$$

Substituting in (v), we get

$$r(U, V) = \frac{1/(36a) - 1/(36b)}{1/(36a) + 1/(36b)} = \frac{b-a}{b+a}.$$

Example 10·13. *Let the random variable X have the marginal density*

$$\left.\begin{aligned} f_1(x) &= 1, -\frac{1}{2} < x < \frac{1}{2} \\ \text{and let the conditional density of } Y \text{ be} \\ f(y \mid x) &= 1,\ x < y < x+1,\ -\frac{1}{2} < x < 0 \\ &= 1,\ -x < y < 1-x,\ 0 < x < \frac{1}{2} \end{aligned}\right\} \qquad (*)$$

Show that the variables X and Y are uncorrelated.

Solution. We have

$$E(X) = \int_{-\frac{1}{2}}^{\frac{1}{2}} x f_1(x)\,dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x.1\,dx = \left|\frac{x^2}{2}\right|_{-1/2}^{1/2} = 0$$

If $f(x, y)$ is the joint p.d.f. of X and Y, then

$$f(x, y) = f(y \mid x) f_1(x) = f(y \mid x). \qquad (**) \qquad [\because f_1(x) = 1]$$

$$E(XY) = \int_{-\frac{1}{2}}^{0}\int_{x}^{x+1} xy f(x, y)\,dx\,dy + \int_{0}^{\frac{1}{2}}\int_{-x}^{1-x} xy f(x, y)\,dx\,dy$$

$$= \int_{-\frac{1}{2}}^{0}\left[x\int_{x}^{x+1} y\,dy\right]dx + \int_{0}^{\frac{1}{2}}\left[x\int_{-x}^{1-x} y\,dy\right]dx \text{ [From } (*) \text{ and}$$

$$= \frac{1}{2}\int_{-\frac{1}{2}}^{0} x(2x+1)dx + \frac{1}{2}\int_{0}^{\frac{1}{2}} x(1-2x)\,dx$$

$$= \frac{1}{2}\left[\frac{2}{3}x^3 + \frac{x^2}{2}\right]_{-\frac{1}{2}}^{0} + \frac{1}{2}\left[\frac{x^2}{2} - \frac{2}{3}x^3\right]_{0}^{\frac{1}{2}}$$

$$= \frac{1}{2}\left[\frac{1}{12} - \frac{1}{8} - \frac{1}{12} + \frac{1}{8}\right] = 0$$

$\therefore$ Cov $(XY) = E(XY) - E(X)\,E(Y) = 0 \Rightarrow r(X, Y) = 0$

Hence the variables X and Y are uncorrelated.

EXERCISE 10(a)

1. (*a*) Show that the co-efficient of correlation r is independent of a change of scale and origin of the variables. Also prove that for two independent variables $r = 0$. Show by an example that the converse is not true. State the limits between which r lies and give its proof.

[Delhi Univ. M.Sc. (O.R.), 1986]

(*b*) Let ρ be the correlation coefficient between two jointly distributed random variables X and Y. Show that $|\rho| \leq 1$ and that $|\rho| = 1$ if and only if X and Y are linearly related. ***[Indan Forest Service, 1991]***

2. (*a*) Calculate the coefficient of correlation between X and Y for the following :

X...	1	3	4	5	7	8	10
Y...	2	6	8	10	14	16	20

Ans. $r(X, Y) = +1$

(*b*) Discuss the statistical validity of the following statements :

(*i*) "High positive coefficient of correlation between increase in the sale of newspapers and increase in the number of crimes leads to the conclusion that newspaper reading may be responsible for the increase in the number of crimes."

(*ii*) "A high positive value of r between the increase in cigarette smoking and increase in lung cancer establishes that cigarette smoking is responsible for lung cancer."

(*c*) (*i*) Do you agree with the statement that "$r = 0{\cdot}8$ implies that 80% of the data are explained."

(*ii*) Comment on the following :

"The closeness of relationship between two variables is proportional to r".

Hint. (*a*) No (*b*) Wrong.

(*d*) By effecting suitable change of origin and scale, compute the product moment correlation coefficient for the following set of 5 observations on (X, Y) :

X :	−10	−5	0	5	10
Y :	5	9	7	11	13

Ans. $r(X, Y) = 0{\cdot}34$

3. The marks obtained by 10 students in Mathematics and Statistics are given below. Find the coefficient of correlation between the two subjects.

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics :	75	30	60	80	53	35	15	40	38	48
Marks in Statistics :	85	45	54	91	58	63	35	43	45	44

4. (*a*) The following table gives the number of blind per lakh of population in different age-groups. Find out the correlation between age and blindness.

Age in years :	0—10	10—20	20—30	30—40	40—50
Number of blind per lakh :	55	67	100	111	150
Age in year :	50—60	60—70	70—80		
Number of blind per lakh :	200	300	500		

Ans. 0·89

(*b*) The following table gives the distribution of items of production and also the relatively defective items among them, according to size-groups. Is there ay correlation between size and defect in quality ?

Size-Group :	15—16	16—17	17—18	18—19	19—20	20—21
No. of Items :	200	270	340	360	400	300
No. of defective items	150	162	170	180	180	120

Hint. Here we have to find the correlation coefficient between the size-group (X) and the percentage of defectives (Y) given below.

X	15·5	16·5	17·5	18·5	19·5	20·5
Y	75	60	50	50	45	40

Ans. $r = 0·94$.

5. Using the formula

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\,r(X, Y)\,\sigma_X\,\sigma_Y$$

obtain the correlation coefficient between the heights of fathers (X) and of the sons (Y) from the following data :

X :	65	66	67	68	69	70	71	67
Y :	67	68	64	72	70	67	70	68

6. (*a*) From the following data, compute the co-efficient of correlation between X and Y.

	X series	*Y series*
No. of items	15	15
Arithmetic mean	25	18
Sum of squares of deviations from mean	136	138

Summation of product of deviations of X and Y series from the respective arithmetic means = 122.

Ans. $r\ (X, Y) = 0{\cdot}891$

(*b*) Coefficient of correlation between two variables X and Y is 0·32. Their covariance is 7·86. The variance of X is 10. Find the standard deviation of Y series.

(*c*) In two sets of variables X and Y with 50 observations each, the following data were observed :

$$\bar{X} = 10,\ \sigma_X = 3,\ \bar{Y} = 6,\ \sigma_Y = 2 \text{ and } r(X, Y) = 0{\cdot}3$$

But on subsequent verification it was found that one value of X (= 10) and one value of Y (= 6) were inaccurate and hence weeded out. With the remaining 49 pairs of values, how is the original value of r affected ?

(Nagpur Univ. B.Sc., 1990)

Hint. $\Sigma X = n\bar{X} = 500,\ \Sigma Y = n\bar{Y} = 300$

$$\Sigma X^2 = n(\sigma_X^2 + \bar{X}^2) = 5450,\ \Sigma Y^2 = 50(4 + 36) = 2000$$

$$r\,\sigma_X\sigma_Y = \text{Cov}\ (X, Y) = \frac{\Sigma XY}{n} - \bar{X}\,\bar{Y}$$

$$\Rightarrow \quad 0{\cdot}3 \times 3 \times 2 = \frac{\Sigma XY}{50} - 10 \times 6$$

$$\Rightarrow \quad \Sigma XY = 50(1{\cdot}8 + 60) = 3090$$

After weeding out the incorrect pair of observation, *viz.*, $(X = 10, Y = 6)$, the corrected values of ΣX, ΣY, ΣX^2, ΣY^2 and ΣXY for the remaining $50 - 1 = 49$ pairs of observations are given below :

Corrected Values :

$\Sigma X = 500 - 10 = 490$; $\Sigma Y = 300 - 6 = 294$

$\Sigma XY = 3090 - 10 \times 6 = 3090 - 60 = 3030$

$\Sigma X^2 = 5450 - 10^2 = 5350,\ \Sigma Y^2 = 2000 - 6^2 = 1964$

$$\therefore r = \frac{\text{Corrected Cov}\ (X, Y)}{(\text{Corrected } \sigma_X) \times (\text{Corrected})\ \sigma_Y} = \frac{90/49}{\sqrt{\frac{450}{49} \times \frac{200}{49}}} = 0{\cdot}3$$

Hence the correlation coefficient is invariant in this case.

(*d*) A prognostic test in Mathematics was given to 10 students who were about to begin a course in Statistics. The scrores (X) in their test were examined in relations to scores (Y) in the final examination in Statistics. The following results were obtained :—

$$\Sigma X = 71,\ \Sigma Y = 70,\ \Sigma X^2 = 555,\ \Sigma Y^2 = 526 \text{ and } \Sigma XY = 527$$

Find the coefficient of correlation between X and Y.

(Kerala Univ. B.Sc., 1990)

7. (*a*) X_1 and X_2 are independent variables with means 5 and 10 and standard deviations 2 and 3 respectively. Obtain $r(U, V)$ where

$$U = 3X_1 + 4X_2 \text{ and } V = 3X_1 - X_2$$

Ans. 0 ***(Delhi Univ. B.Sc. ,1988)***

(*b*) If X and Y are normal and independent with zero means and standard deviations 9 and 12 respectively, and if $X + 2Y$ and $kX - Y$ are non-correlated, find k.

(*c*) X, Y, Z are random variables each with expectation 10 and variances 1, 4 and 9 respectively. The correlation coefficients are

$$r(X, Y) = 0, r(Y, Z) = r(X, Y) = 1/4$$

Obtain the numerical values of :

(*i*) $E(X + Y - 2Z)$, (*ii*) Cov $(X + 3, Y + 3)$, (*iii*) $V(X - 2Z)$ and

(*iv*) Cov $(3X, 5Z)$

Ans. (*i*) = 0, (*ii*) 0, (*iii*) 34, and (*iv*) 45/4.

(*d*) X and Y are discrete random variables. If Var (X) = Var$(Y) = \sigma^2$, Cov $(X, Y) = \frac{\sigma^2}{2}$, find (*i*) Var $(2X - 3Y)$, (*ii*) Corr $(2X + 3, 2Y - 3)$.

8. (*a*) Prove that :

$$V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm 2ab \text{ Cov } (X, Y)$$

Hence deduce that if X and Y are independent

$$V(X \pm Y) = V(X) + V(Y)$$

(*b*) Prove that correlation coefficient between X and Y is positive or negative according as

$$\sigma_{X+Y} > \text{or} < \sigma_{X-Y}$$

9. Show that if X and Y are two random variables each assuming only two values and the correlation co-efficient between them is zero, then they are independent. Indicate with justification whether the result is true in general.

Find the correlation coefficient between X and $a - X$, where X is any random variable and a is constant.

10. (*a*) X_i (i =1, 2, 3) are uncorrelated variables each having the same standard deviation. Obtain the correlation between $X_1 + X_2$ and $X_2 + X_3$.

Ans. 1/2

(*b*) If X_i (i =1, 2, 3) are three uncorrelated variables having standard deviations σ_1, σ_2 and σ_3 respectively, obtain the coefficient of correlation between $(X_1 + X_2)$ and $(X_2 + X_3)$.

Ans. $\sigma_2^2/\sqrt{(\sigma_1^2 + \sigma_2^2)(\sigma_2^2 + \sigma_3^2)}$

(*c*) Two random variables X and Y have zero means, the same variance σ^2 and zero correlation. Show that

$U = X \cos \alpha + Y \sin \alpha$ and $V = X \sin \alpha - Y \cos \alpha$

have the same variance σ^2 and zero correlation.

(Bangalore Univ. B.Sc., 1991)

(*d*) Let X and Y be uncorrelated random variables. If $U = X + Y$ and $V = X - Y$, prove that the coefficient of correlation between U and V is $(\sigma_X^2 - \sigma_Y^2)/(\sigma_X^2 + \sigma_Y^2)$, where σ_X^2 and σ_Y^2 are variances of X and Y respectively.

(*e*) Two independent random variables X and Y have the following variances : $\sigma_X^2 = 36$, $\sigma_Y^2 = 16$. Calculate the coefficient of correlation between

$$U = X + Y \text{ and } V = X - Y$$

(*f*) Random variables X and Y have zero means and non-zero variances σ_X^2 and σ_Y^2. If $Z = Y - X$, then find σ_Z and the correlation coefficient $\rho(X, Z)$ of X and Z in terms of σ_X, σ_Y and the correlation coefficient $\rho\ (X, Y)$ of X and Y.

(*g*) If the independant random variables X_1, X_2 and X_3 have the means 4, 9 and 3 and variances 3, 7, 5, respectively, obtain the mean and variance of

(*i*) $Y = 2X_1 - 3X_2 + 4X_3$, (*ii*) $Z = X_1 + 2X_2 - X_3$, and

(*iii*) Calculate the correlation between Y and Z.

[*Delhi Univ. M.A.(Eco.)., 1989*]

11. (*a*) $X_1, X_2, ..., X_n$ are uncorrelated random variables, all with the same distribution and zero means. Let $\bar{X} = \Sigma X_i/n$

Find the correlation co-efficient between (*i*) X_i and $\bar{X}$ and (*ii*) $X_i - \bar{X}$ and $\bar{X}$.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1993*]

Hint. $$r(X_i, \bar{X}) = \frac{\sigma^2/n}{\sqrt{\sigma^2 . \sigma^2/n}} = \frac{1}{\sqrt{n}}$$

$$\begin{aligned} \text{Cov}(X_i - \bar{X}, \bar{X}) &= \text{Cov}\ (X_i, \bar{X}) - \text{Var}\ (\bar{X}) \\ &= (\sigma^2/n) - (\sigma^2/n) = 0 \end{aligned}$$

$$\therefore \qquad r(X_i - \bar{X}, \bar{X}) = 0$$

(*b*) $X_1, X_2, \ldots \ldots, X_n$ are random variables each with the same expected value μ and s.d. σ. The correlation coefficient between any two X's is ρ. Show that (*i*) $\text{Var}\ (\bar{X}) = \frac{\sigma^2}{n} + \left(1 - \frac{1}{n}\right)\rho\sigma^2$,

(*ii*) $E \sum_1^n (X_i - \bar{X})^2 = (n-1)(1-\rho)\sigma^2$, and (*iii*) $\rho > -\frac{1}{n-1}$

12. (*a*) If X and Y are independent random variables, show that

$$r(X + Y, X - Y) = r^2(X, X + Y) - r^2\ (Y, X + Y),$$

where $r(X + Y, X - Y)$ denotes the co-efficient of correlation between $(X + Y)$ and $(X - Y)$. **(*Meerut Univ. B.Sc., 1991*)**

(*b*) Let X and Y be random variables having mean 0, variance 1 and correlation r. Show that $X - rY$ and Y are uncorrelated and that $X - rY$ has mean zero and variance $1 - r^2$.

13. X_1 and X_2 are two variables with zero means, variances σ_1^2 and σ_2^2 respectively and r is the correlation coefficient between them. Determine the values of the constants a and b which are independent of r such that $X_1 + aX_2$ and $X_1 + bX_2$ are uncorrelated.

14. (*a*) If X_1 and X_2 are two random variables with means μ_1 and μ_2, variances σ_1^2, σ_2^2 and correlation coefficient r, find the correlation co-efficient between

$$U = a_1X_1 + a_2X_2 \text{ and } V = b_1X_1 + b_2X_2,$$

where a_1, a_2 and b_1, b_2 are constants.

(*b*) Let X_1, X_2 be independent random variables with means μ_1, μ_2 and non-zero variances σ_1^2, σ_2^2 respectively. Let $U = X_1 - X_2$ and $V = X_1 X_2$. Find the correlation coefficient between (*i*) X_1 and U, (*ii*) X_1 and V, in terms of μ_1, μ_2, σ_1^2, σ_2^2.

15. (*a*) If $U = aX + bY$ and $V = bX - aY$, where X and Y are measured from their respective means and if U and V are uncorrelated, r the co-efficient of correlation between X and Y is given by the equation.

$$\sigma_U \sigma_V = (a^2 + b^2)\, \sigma_X \sigma_Y (1 - r^2)^{1/2}$$ **(Utkal Univ. B. Sc., 1993)**

(*b*) Let $U = aX + bY$ and $V = aX - bY$ where X, Y represent deviations from the means of two measurements on the same individual. The coefficient of correlation between X and Y is ρ. If U, V are uncorrelated, show that

$$\sigma_U \sigma_V = 2ab\sigma_X \sigma_Y (1 - r^2)^{1/2}$$

16. Show that, if a and b are constants and r is the correlation coefficient between X and Y, then the correlation coefficient between aX and bY is equal to r if the signs of a and b are alike, and to $-r$ if they are different.

Also show that, if constants a, b and c are positive, the correlation coefficient between $(aX + bY)$ and cY is equal to

$$(ar\sigma_X + b\sigma_Y) / \sqrt{(a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X\sigma_Y)}$$

17. If X_1, X_2 and X_3 are three random variables measured from their respective means as origin and of equal variances, find the coefficient of correlation between $X_1 + X_2$ and $X_2 + X_3$ in terms of r_{12}, r_{13} and r_{23} and show that it is equal to

$$(i)\ \frac{r_{12} + 1}{2},\ \text{if } r_{13} = r_{23} = 0, \text{ and } (ii) = \frac{r_{12} + 3}{4},\ \text{if } r_{13} = r_{23} = 1$$

18. (*a*) For a weighted distribution (x_i, w_i), $(i = 1, 2, \ldots, n)$ show that the weighted arithmetic mean $\bar{x}_w = \Sigma\, w_i x_i / \Sigma\, w_i$ > or < the unweighted mean $\bar{x} = \Sigma\, x_i / n$ according as $r_{xw} > \text{or} < 0$.

(*b*) Given N values $x_1, x_2, \ldots, x_N$ of variable X and weights $w_1, w_2, \ldots, w_N$, express the coefficient of correlation between X and W in terms involving the difference between the arithmetic mean and the weighted mean of X.

19. (*a*) A coin is tossed n times. If X and Y denote the (random) number of heads and number of tails turned up respectively, show that $r(X, Y) = -1$.

Hint. Note that $X + Y = n \Rightarrow Y = n - X$

$$\therefore \qquad r(X, Y) = r(X, n - X) = r(X, -X) = -r(X, X) = -1.$$

(*b*) Two dice are thrown, their scores being a and b. The first die is left on the table while the second is picked up and thrown again giving the score c. Suppose the process is repeated a large number of times. What is the correlation coefficient between $X = a + b$ and $Y = a + c$?

Ans. $r(X, Y) = \frac{1}{2}$

20. (*a*) If X and Y are independent random variables with means μ_1 and μ_2 and variances σ_1^2, σ_2^2 respectively, show that the correlation coefficient between $U = X$ and $V = X - Y$ in terms of μ_1, μ_2, σ_1^2 and σ_2^2 is $\sigma_1 / \sqrt{\sigma_1^2 + \sigma_2^2}$.

(*b*) If X and Y are independent random variables with non-zero variances, show that the correlation coefficient between $U = XY$ and $V = X$ in terms of mean and variance of X and Y is given by

$$\mu_2\sigma_1 / \sqrt{\sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2}$$

[*Delhi Univ. B.Sc. (Stat Hons.), 1987*]

21. If X_i, Y_j and Z_k are all independent random variables with mean zero and unit variance, find the correlation coefficient between

$$U = \sum_{i=1}^{m} X_i + \sum_{j=1}^{n} Y_j \text{ and } V = \sum_{i=1}^{m} X_i + \sum_{k=1}^{n} Z_k$$

Ans. $r(U, V) = m/(m + n)$ (*Bombay Univ., B.Sc, 1990*)

22. (*a*) Find the value of l so that the correlation coefficient between $(X - lY)$ and $(X + Y)$ is maximum, where X, Y are independent random variables each with mean zero and variance 1. [**Ans.** $l = -1$]

Hint. $U = X - lY$; $V = X + Y$. Now find l so that $r(U, V) = 1$.

(*b*) If $U = X + kY$ and $V = X + mY$ and r is the correlation coefficient between X and Y, find the correlation coefficient between U and V. Show that U and V are uncorrelated if $k = \dfrac{-\sigma_X(\sigma_X + rm\,\sigma_Y)}{\sigma_Y(r\,\sigma_X + m\sigma_Y)}$

and further if $m = \dfrac{\sigma_X}{\sigma_Y}$, then $k = -\dfrac{\sigma_X}{\sigma_Y}$. (*Gujarat Univ. M.A., 1993*)

23. X_1, X_2, X_3 are three variables, each with variance σ^2 and the correlation coefficient between any two of them is r. If $\bar{X} = (X_1 + X_2 + X_3)/3$, show that

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{3}(1 + 2r)$$

Deduce that $r \geq -1/2$.

24. (*a*) If $U = aX + bY$ and $V = bX - aY$, show that U and V are uncorrelated if $\dfrac{ab}{a^2 - b^2} = \dfrac{\rho\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2}$

where ρ is the coefficient of correlation between X and Y. Show further that, in this case

$$\sigma_U^2 + \sigma_V^2 = (a^2 + b^2)(\sigma_X^2 + \sigma_Y^2) \text{ and } \sigma_U\sigma_V = (a^2 + b^2)\sigma_X\sigma_Y\sqrt{1 - \rho^2}$$

(*b*) If $u = aX + bY$, $v = cX + dY$, show that

$$\begin{vmatrix} \text{var}(u) & \text{cov}(u,v) \\ \text{cov}(u,v) & \text{var}(v) \end{vmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix}^2 \begin{vmatrix} \text{var}(X) & \text{cov}(X, Y) \\ \text{cov}(X, Y) & \text{var}(Y) \end{vmatrix}$$

25. If X is a standard normal variate and $Y = a + bX + cX^2$, where a, b, c are constants, find the correlation coefficient between X and Y. Hence or otherwise obtain the conditions when (*i*) X and Y are uncorrelated and (*ii*) X and Y are perfectly correlated.

26. (*a*) If $X \sim N(0, 1)$, find corr (X, Y) where $Y = a + bX + cX^2$.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1985*]

Ans. $r(X, Y) = \dfrac{b}{\sqrt{b^2 + 2c^2}}$

(*b*) If X has Laplace distribution with parameters $(\lambda, 0)$ and $Y = a + bX + cX^2$, find $\rho(X, Y)$

[*Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1989*]

Hint. $p(x) = \frac{1}{2}\lambda \exp[-\lambda |x|], \quad -\infty < x < \infty.$

$E(X^{2k+1}) = 0 = \mu_{2k+1}$; $E(X^{2k}) = \mu_{2k} = (2k)! / \lambda^{2k}$

$$\rho_{XY} = \frac{\lambda b}{\sqrt{b^2\lambda^2 + 10c^2}}$$

27. In a sample of n random observations from exponential distribution with parameter λ, the number of observations in $(0, 1/\lambda)$ and $(1/\lambda, 2/\lambda)$, denoted by X and Y are noted. Find $\rho(X, Y)$.

Hint. $p_1 = p(0 < X < 1/\lambda) = \displaystyle\int_0^{1/\lambda} \lambda e^{-\lambda x} dx = \frac{e - 1}{e}$

$$p_2 = p(1/\lambda < Y < 2/\lambda) = \int_{1/\lambda}^{2/\lambda} \lambda e^{-\lambda y} dy = \frac{e - 1}{e^2}$$

Then (X, Y) has a trinomial distribution with parameters $(n = 3, p_1, p_2, p_3 = 1 - p_1 - p_2)$.

Hence we have

$$\rho(X, Y) = -\left[\frac{p_1 p_2}{(1 - p_1)(1 - p_2)}\right]^{1/2} = -\frac{e - 1}{\sqrt{e^2 - e + 1}}.$$

28. Prove that :

$$r(X, Y + Z) = \frac{\sigma_Y}{\sigma_{Y+Z}} \cdot r(X, Y) + \frac{\sigma_Z}{\sigma_{Y+Z}} \cdot r(X, Z)$$

29. If X and Y are independent random variables, find $\text{Corr}(X, XY)$. Deduce the value of $\text{Corr}(X, X/Y)$.

Ans. $r(X, XY) = \sigma_X \mu_Y / [\sigma_X^2 \sigma_Y^2 + \mu_X^2 \sigma_Y^2 + \mu_Y^2 \sigma_X^2]^{1/2}$

30. Prove or Disprove :

(*a*) $r(X, Y) = 0 \quad \Rightarrow \quad r(|X|, Y) = 0$

(*b*) $r(X, Y) = 0, r(Y, Z) = 0 \quad \Rightarrow \quad r(X, Z) = 0.$

Ans. (*a*) False, unless X and Y are independent.

(*b*) **Hint.** Let $Z \equiv X$, and X and Y be independent. Then $r(X, Y) = 0 = r(Y, Z)$. But $r(X, Z) = r(X, X) = 1$.

31. Let random variable X have a p.d.f. $f(.)$ with distribution function $F(.)$, mean μ and variance σ^2. Define $Y = \alpha + \beta X$, where α and β are constants satisfying $-\infty < \alpha < \infty$, and $\beta > 0$.

(*a*) Select α and β so that Y has mean 0 and variance 1.

(*b*) What is the correlation coefficient between X and Y ?

32. Let (X, Y) be jointly discrete random variables such that each X and Y have at most two mass points. Prove or disprove : X and Y are independent if and only if they are uncorrelated.

Ans. True.

33. If the variables $X_1, X_2, \ldots, X_{2n}$ all have the same variance σ^2 and the correlation coefficient between X_i and X_j $(i \neq j)$ has the same value, show that the correlation between $\sum_{i=1}^{n} X_i$ and $\sum_{j=n+1}^{2n} X_j$ is given by $[n\rho/\{1 + (n-1)\rho\}]$.

34. The means of independent r.v's $X_1, X_2, \ldots, X_n$ are zero and variances are equal, say unity. The correlation coefficients between the sum of selected t $(< n)$ variables out of these variables and the sum of all n variables are found out. Prove that the sum of squares of all these correlation coefficients is ${}^{n-1}C_{t-1}$.

[Burdwan Univ. B.Sc. (Hons.), 1989]

35. Two variables U and V are made up of the sum of a number of terms as follows :

$$U = X_1 + X_2 + \ldots + X_n + Y_1 + Y_2 + \ldots + Y_a,$$
$$V = X_1 + X_2 + \ldots + X_n + Z_1 + Z_2 + \ldots + Z_b,$$

where a and b are all suffixes and where X's, Y's and Z's are all uncorrelated standardised random variables. Show that the correlation coefficient between U and V is $\dfrac{n}{\sqrt{(n+a)(n+b)}}$. Show further that

$$\left.\begin{aligned} \xi &= \sqrt{(n+b)}\,U + \sqrt{(n+a)}\,V \\ \text{and} \quad \eta &= \sqrt{(n+b)}\,U - \sqrt{(n+a)}\,V \end{aligned}\right\} \quad \ldots(*)$$

are uncorrelated ***[South Gujarat Univ. B.Sc., 1989]***

36. (*a*) Let the random variables X and Y have the joint p.d.f.

$$f(x, y) = 1/3 \; ; \; (x, y) = (0, 0), (1, 1) (2, 0)$$

Compute $E(X)$, $V(X)$, $E(Y)$, $V(Y)$ and $r(X, Y)$. Are X and Y stochastically independent ? Give reasons.

(*b*) Let (X, Y) have the probability distribution :

$$f(0, 0) = 0{\cdot}45, f(0, 1) = 0{\cdot}05, f(1, 0) = 0{\cdot}35, f(1, 1) = 0{\cdot}15.$$

Evaluate $V(X)$, $V(Y)$ and $\rho(X, Y)$.

Show that while X and Y are correlated, X and $X - 5Y$ are uncorrelated. Are X and $X - 5Y$ independent ?

(*c*) Given the bivariate probability distribution :

$$\begin{aligned} f(-1, 0) &= 1/15, \quad f(-1, 1) = 3/15, \quad f(-1, 2) = 2/15 \\ f(0, 0) &= 2/15, \quad f(0, 1) = 2/15, \quad f(0, 2) = 1/15 \\ f(1, 0) &= 1/15, \quad f(1, 1) = 1/15, \quad f(1, 2) = 2/15 \end{aligned}$$

$$f(x, y) = 0, \text{ elsewhere.}$$

Obtain :

(*i*) The marginal distributions of X and Y.

(*ii*) The conditional distributions of Y given $X = 0$.

(*iii*) $E(Y|X = 0)$.

(*iv*) The product moment correlation coefficient between X and Y.

Are X and Y independently distributed ?

37. If X and Y are standardised variates with correlation coefficient ρ, prove that $E[\max(X^2, Y^2)] \leq 1 + \sqrt{1 - \rho^2}$

Hint. $\max(X^2, Y^2) = \frac{1}{2}|X^2 - Y^2| + \frac{1}{2}(X^2 + Y^2)$ ...(*)

$E(X) = E(Y) = 0$; $E(X^2) = E(Y^2) = 1$; $E(XY) = \rho$

and $[E|X - Y| \cdot |X + Y|]^2 \leq E(X - Y)^2 \cdot E(X + Y)^2$

(By Cauchy-Schwartz Inequality)

38. The joint p.d.f. of two variates X and Y is given by

$f(x, y) = k[(x + y) - (x^2 + y^2)]$; $0 < (x, y) < 1$

$= 0$, otherwise.

Show that X and Y are uncorrelated but not independent.

39(*a*). If the random variables X and Y have the joint p.d.f.,

$$f(x, y) = \begin{cases} x + y; & 0 < x < 1,\ 0 < y < 1 \\ 0, & \text{elsewhere} \end{cases}$$

then show that the correlation coefficient between X and Y is $-\frac{1}{11}$.

[*Madras Univ. B.Sc., Oct., 1990*]

(*b*) The density function f of a random variable X is given by

$$f(x) = \begin{cases} kx^2, & \text{if } -1 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

(*i*) What is the value of k ? What is the distribution function of X ?

(*ii*) Obtain the density function of the random variable $Y = X^2$.

(*iii*) Obtain the correlation coefficient between X and Y.

(*iv*) Are X and Y independently distributed ?

40(*a*). If $f(x, y) = \frac{6 - x - y}{8}$; $0 \leq x \leq 2,\ 2 \leq y \leq 4$,

find (*i*) Var (X), (*ii*) Var (Y) (*iii*) $r(X, Y)$.

Ans. (*i*) $\frac{11}{36}$, (*ii*) $\frac{11}{36}$, (*iii*) $-\frac{1}{11}$.

(*b*) Given the joint density of random variables X, Y, Z as :

$f(x, y, z) = k\,x \exp[-(y + z)]$, $0 < x < 2,\ y \geq 0,\ z \geq 0$

$= 0$, elsewhere

Find

(*i*) k,

(*ii*) the marginal density function,

(*iii*) conditional expectation of Y, given X and Z, and

(*iv*) the product moment correlation between X and Y.

[*Madras Univ. B.Sc. (Main Stat.), 1988*]

(*c*) Suppose that the two dimensional random variable (X, Y) has p.d.f. given by $f(x, y) = ke^{-y}$, $0 < x < y < 1$
$= 0$, elsewhere
Find the correlation coefficient r_{XY}. [*Delhi Univ. M.C.A., 1991*]

41. The joint density of (X, Y) is :

$$f(x, y) = \frac{1}{8}(x + y), \quad 0 \le x \le 2, 0 \le y \le 2.$$

Find $\mu'_{rs} = E(X^r Y^s)$ and hence find Corr (X, Y).

Ans. $\mu'_{rs} = 2^{r+s}\left[\frac{1}{(r+2)(s+1)} + \frac{1}{(r+1)(s+2)}\right]$; $r = -\frac{1}{11}$.

(*b*) Find the m.g.f. of the bivariate distribution :
$f(x, y) = 1$, $0 < (x, y) < 1$
$= 0$, otherwise
and hence find $r(X, Y)$.

Ans. $M(t_1, t_2) = (e^{t_1} - 1)(e^{t_2} - 1)/(t_1 t_2)$; $t_1 \neq 0, t_2 \neq 0$, $r(X, Y) = 0$.

42. Let (X, Y) have joint density :

$$f(x, y) = e^{-(x+y)} I_{(0,\infty)}(x) \cdot I_{(0,\infty)}(y)$$

Find Corr (X, Y). Are X and Y independent ?

Ans. Corr $(X, Y) = 0$: X and Y are independent.

43. A bivariate distribution in two discrete random variables X and Y is defined by the probability generating function :

$$\exp[a(u - 1) + b(v - 1) + c(u - 1)(v - 1)],$$

simultaneous probability of $X = r \cap Y = s$, where r and s are integers being the coefficient of $u^r v^s$. Find the correlation coefficient between X and Y.

Hint. Put $u = e^{t_1}$ and $v = e^{t_2}$ in $\exp[a(u - 1) + b(v - 1) + c(u - 1)(v - 1)]$, the result will be the m.g.f. of a bivariate distribution and is given by

$$M(t_1, t_2) = \exp[a(e^{t_1} - 1) + b(e^{t_2} - 1) + c(e^{t_1} - 1)(e^{t_2} - 1)]$$

We have $\left[\frac{\partial M}{\partial t_1}\right]_{t_1 = t_2 = 0} = a$, $\left[\frac{\partial^2 M}{\partial t_1^2}\right]_{t_1 = t_2 = 0} = a(a + 1)$

$\left[\frac{\partial^2 M}{\partial t_1 \partial t_2}\right]_{\substack{t_1 = 0 \\ t_2 = 0}} = ab + c$, $\left[\frac{\partial M}{\partial t_2}\right]_{t_1 = t_2 = 0} = b$, $\left[\frac{\partial^2 M}{\partial t_2^2}\right]_{t_1 = t_2 = 0} = b(b + 1)$

So we have

$E(X) = a, E(X^2) = a(a + 1), E(Y) = b, E(Y^2) = b(b + 1)$ and $E(XY) = ab + c$

$$\therefore \quad r(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{[E(X^2) - \{E(X)\}^2][E(Y^2) - \{E(Y)\}^2]}} = \frac{c}{\sqrt{ab}}$$

44. Let the number X be chosen at random from among the integers 1, 2, 3, 4 and the number Y be chosen from among those at least as large as X. Prove that Cov $(X, Y) = 5/8$. Find also the regression line of Y on X.
[*Delhi Univ. B.Sc. (Maths. Hons.), 1990*]

Hint. $P(X = k) = \frac{1}{4}$; $k = 1, 2, 3, 4$ and $Y \ge X$.

$P(Y = y \mid X = 1) = \frac{1}{4}$; $y = 1, 2, 3, 4$ $(\because y \geq x)$;

$P(Y = y \mid X = 2) = \frac{1}{3}$, $y = 2, 3, 4$

$P(Y = y \mid X = 3) = \frac{1}{2}$, $y = 3, 4$; $P(Y = y \mid X = 4) = 1$, $y = 4$.

The joint probability distribution can be obtained on using :

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y \mid X = x).$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{5/8}{\sqrt{(5/4) \times (41/48)}} = \sqrt{\frac{15}{41}}$$

Regression line of Y on X : $Y - E(Y) = \frac{r \sigma_Y}{\sigma_X}[X - E(X)]$

45. Two ideal dice are thrown. Let X_1 be the score on the first dice ad X_2, the score on the second dice. Let $Y = \max\{X_1, X_2\}$. Obtain the joint distribution of Y and X_1 and show that

$$\text{Corr}((Y, X_1) = \frac{3}{2\sqrt{73}}$$

46. Consider an experiment of tossing two tetrahedra. Let X be the number of the down turned face of first tetrahedron and Y, the larger of the two numbers. Obtain the joint distribution of X and Y and hence $\rho(X, Y)$.

Ans. $\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{5/8}{\sqrt{5/4} \cdot \sqrt{55/64}} = \frac{2}{\sqrt{11}}$

47. Three fair coins are tossed. Let X denote the number of heads on the first two coins and let Y denote the number of tails on the last two coins.

(*a*) Find the joint distribution of X and Y.

(*b*) Find the conditional distribution of Y given that $X = 1$.

(*c*) Find Cov. (X, Y)

Ans. Cov. $(X, Y) = -1/4$.

48. For the trinomial distribution of two random variables X and Y :

$$f(x, y) = \frac{n!}{x!\, y!\, (n - x - y)!} p^x q^y (1 - p - q)^{n - x - y}$$

for $x, y = 0, 1, 2, \ldots, n$ and $x + y \leq n$, $p \geq 0$, $q \geq 0$ and $p + q \leq 1$.

(*a*) Obtain the marginal distribution of Y

(*b*) Obtain $E(X \mid Y = y)$.

(*c*) Find $\rho(X, Y)$.

Ans. (*a*) $X \sim B(n, p)$, $Y \sim B(n, q)$

(*b*) $(X \mid Y = y) \sim B\left(n - y, \frac{p}{1 - q}\right)$

(Note : $p + q \neq 1$)

$\therefore E(X \mid Y = y) = (n - y)\left(\frac{p}{1 - q}\right)$

(*c*) Cov $(X, Y) = -npq$; $\rho(X, Y) = -\left[\frac{pq}{(1-p)(1-q)}\right]^{1/2}$

OBJECTIVE TYPE QUESTIONS

I. Comment on the following :

(*i*) $r_{XY} = 0 \Rightarrow X$ and Y are independent.

(*ii*) If $r_{XY} > 0$ then $r_{X,-Y} > 0$, $r_{-X,Y} > 0$ and $r_{-X,-Y} > 0$

(*iii*) $r_{XY} > 0 \Rightarrow E(XY) > E(X)\,E(Y)$

(*iv*) Pearson's coefficient of correlation is independent of origin but not of scale.

(*v*) The numerical value of product moment correlation coefficient 'r' between two variables X and Y cannot exceed unity.

(*vi*) If the correlation coefficient between the variables X and Y is zero then the correlation coefficient between X^2 and Y^2 is also zero.

(*vii*) If $r > 0$, then as X increases, Y also increases.

(*viii*) "The closeness of relationship between two variables is proportional to r."

(*ix*) r measures every type of relationship between the two variables.

II. Comment on the following values of 'r' (correlation coefficient) :

1, – 0·95, 0, –1·64, 0·87, 0·32, –1, 2·4.

III. (*i*) If $\rho_{XY} = -0{\cdot}9$, then for large values of X, what sort of values do we expect for Y ?

(*ii*) If $\rho_{XY} = 0$, what is the value of cov (X, Y) and how are X and Y related ?

IV. Indicate the correct answer :

(*i*) The coefficient of correlation will have positive sign when

(*a*) X is increasing, Y is decreasing, (*b*) both X and Y are increasing, (*c*) X is decreasing, Y is increasing, (*d*) there is no change in X and Y.

(*ii*) The coefficient of correlation (*a*) can take any value between –1 and +1 (*b*) is always less than –1, (*c*) is always more than +1, (*d*) cannot be zero.

(*iii*) The coefficient of correlation (*a*) cannot be positive, (*b*) cannot be negative, (*c*) is always positive, (*d*) can be both positive as well as negative.

(*iv*) Probable error of r is

(*a*) $0{\cdot}6475\frac{1-r^2}{\sqrt{n}}$, (*b*) $0{\cdot}6754\frac{1+r^2}{\sqrt{n}}$, (*c*) $0{\cdot}6547\frac{1-r^2}{n}$,

(*d*) $0{\cdot}6754\frac{1-r^2}{n}$.

(*v*) The coefficient of correlation between X and Y is 0·6. Their covariance is 4·8. The variance of X is 9. Then the S.D. of Y is

(*a*) $\frac{4{\cdot}8}{3 \times 0{\cdot}6}$, (*b*) $\frac{0{\cdot}6}{4{\cdot}8 \times 3}$, (*c*) $\frac{3}{4{\cdot}8 \times 0{\cdot}6}$, (*d*) $\frac{4{\cdot}8}{9 \times 0{\cdot}6}$.

(*vi*) The coefficient of correlation is independent of (*a*) change of scale only, (*b*) change of origin only, (*c*) both change of scale and origin, (*d*) neither change of scale nor change of origin.

V. Fill in the blanks :

(*i*) The Karl Pearson coefficient of correlation between variables X and Y is ... ...

(*ii*) Two independent variables are ... ...

(*iii*) Limits for correlation coefficient are ... ...

(*iv*) If r be the correlation coefficient between the random variables X and Y then the variance of $X + Y$ is ... ...

(*v*) The absolute value of the product moment correlation coefficient is less than ... ...

(*vi*) Correlation coefficient is invariant under changes of .. and ... ...

VI. How can you use scatter diagram to obtain an idea of extent and nature (direction) of the correlation coefficient ?

10·4. Calculation of the Correlation Coefficient for a Bivariate Frequency Distribution. When the data are considerably large, they may be summarised by using a two-way table. Here, for each variable a suitable number of classes are taken, keeping in view the same considerations as in the univariate case. If there are n classes for X and m classes for Y, there will be in all $m \times n$ cells in the two-way table. By going through the pairs of values of X and Y, we can find the frequency for each cell. The whole set of cell frequencies will then define a *bivariate frequency distribution*. The column totals and row totals will give us the marginal distributions of X and Y. A particular column or row will be called the conditional distribution of Y for given X or of X for given Y respectively.

Suppose that the bivariate data on X and Y are presented in a two-way correlation table (shown on page 10·33) where there are m classes of Y placed along the horizontal line and n classes of X along a vertical line and f_{ij} is the frequency of individuals lying in the (i, j)th cell.

Here $$\sum_x f(x, y) = g(y)$$

is the sum of the frequencies along any row and

$$\sum_y f(x, y) = f(x)$$

is the sum of the frequencies along any column. We observe that

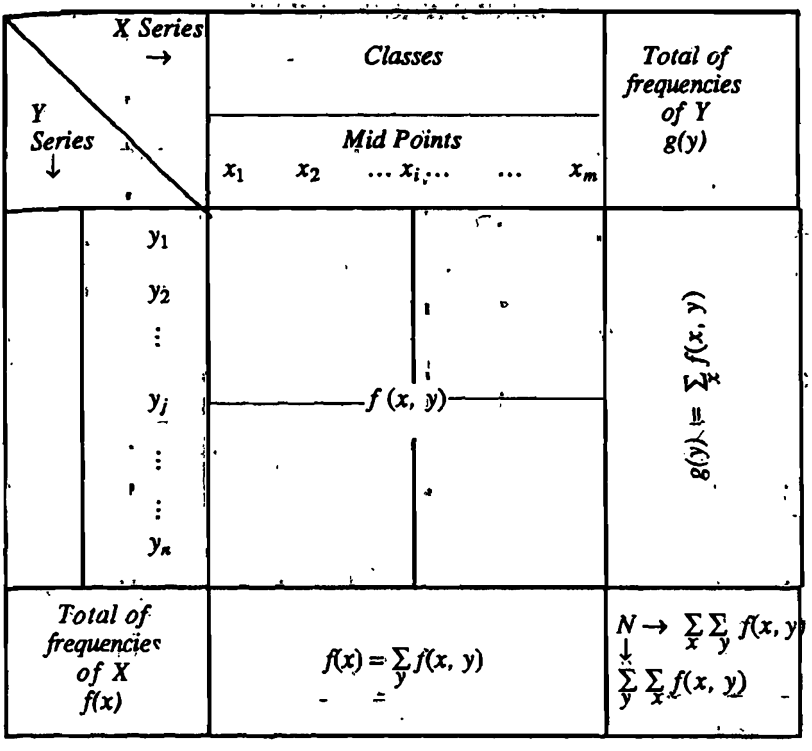
Thus $$\sum_x \sum_y f(x, y) = \sum_y \sum_x f(x, y) = \sum_x f(x) = \sum_y g(y) = N$$

Similarly $$\bar{x} = \frac{1}{N} \sum_x \sum_y x f(x, y) = \frac{1}{N} \left[\sum_x \{ x \sum_y f(x, y) \} \right] = \frac{1}{N} \sum_x x f(x)$$

$$\bar{y} = \frac{1}{N} \sum_x \sum_y y f(x, y) = \frac{1}{N} \sum_y y \, g(y)$$

$$\sigma_x^2 = \frac{1}{N} \sum_x \sum_y x^2 f(x, y) - \bar{x}^2 = \frac{1}{N} \sum_x x^2 f(x) - \bar{x}^2$$

BIVARIATE FREQUENCY TABLE (CORRELATION TABLE)



Example 10·14. *The following table gives, according to age, the frequency of marks obtained by 100 students in an intelligence test.*

Ages in years → / Marks ↓	18	19	20	21	Total
10—20	4	2	2	—	8
20—30	5	4	6	4	19
30—40	6	8	10	11	35
40—50	4	4	6	8	22
50—60	—	2	4	4	10
60—70	—	2	3	1	6
Total	19	22	31	28	100

Calculate the correlation coefficient.

Solution.

CORRELATION TABLE

v	y	u / x Marks	-1	0	1	2	Total $f(v)$	$vf(v)$	$v^2f(v)$	$\sum_u uv\,f(u,v)$
			18	19	20	21				
-2	15	10—20	(8) 4	(0) 2	(4) 2		8	-16	32	4
-1	25	20—30	(5) 5	(0) 4	(6) 6	(8) 4	10	-19	19	-9
0	35	30—40	(0) 6	(0) 8	(0) 10	(0) 11	35	0	0	0
1	45	40—50	(4) 4	(0) 4	(6) 6	(16) 8	22	22	22	18
2	55	50—60		(0) 2	(8) 4	(16) 4	10	20	40	24
3	65	60—70		(0) 2	(9) 3	(6) 1	6	18	54	15
		Total $f(u)$	19	22	31	28	100	25	167	52
		$uf(u)$	-19	0	31	56	68			
		$u^2f(u)$	19	0	31	112	162			
		$u\sum_v vf(u,v)$	9	0	13	30	52			

Let $U = X - 19, V = \{(Y - 35)/10\}$

$$\bar{u} = \frac{1}{N}\sum_u u f(u) = \frac{68}{100} = 0{\cdot}68, \bar{v} = \frac{1}{N}\sum_v v\,g(v) = \frac{25}{100} = 0{\cdot}25$$

$$\text{Cov}(u,v) = \frac{1}{N}\sum_u\sum_v uv\,f(u,v) - \bar{u}\,\bar{v} = \frac{1}{100}\times 52 - 0{\cdot}68 \times 0{\cdot}25 = 0{\cdot}35$$

$$\sigma_U^2 = \frac{1}{N}\sum_u u^2 f(u) - \bar{u}^2 = \frac{162}{100} - (0{\cdot}68)^2 = 1{\cdot}1576$$

$$\sigma_V^2 = \frac{1}{N}\sum_v v^2 g(v) - \bar{v}^2 = \frac{167}{100} - (0{\cdot}25)^2 = 1{\cdot}6075$$

$$\therefore \quad r(U,V) = \frac{\text{Cov}(U,V)}{\sigma_U\,\sigma_V} = \frac{0{\cdot}35}{\sqrt{1{\cdot}1576 \times 1{\cdot}6075}} = 0{\cdot}25$$

Since correlation coefficient is independent of change of origin and scale,

$$r(X, Y) = r(U, V) = 0.25$$

Remark. Figures in circles in the table on page10·34 are the product terms $uvf(u, v)$:

Example 10·15. *The joint probability distribution of X and Y is given below* :

Y \ X	−1	+1
0	$\frac{1}{8}$	$\frac{3}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$

Find the correlation coefficient between X and Y.

Solution.

COMPUTATION OF MARGINAL PROBABILITIES

Y \ X	−1	+1	$g(y)$
0	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{4}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{4}{8}$
$p(x)$	$\frac{3}{8}$	$\frac{5}{8}$	1

We have :

$$E(X) = \Sigma x p(x) = (-1) \times \frac{3}{8} + 1 \times \frac{5}{8} = \frac{1}{4}$$

$$E(X^2) = \Sigma x^2 p(x) = (-1)^2 \times \frac{3}{8} + 1^2 \times \frac{5}{8} = 1$$

$$\therefore \quad \text{Var}(X) = E(X^2) - [E(X)]^2 = 1 - \frac{1}{16} = \frac{15}{16}$$

$$E(Y) = \Sigma y\, g(y) = 0 \times \frac{4}{8} + 1 \times \frac{4}{8} = \frac{1}{2}$$

$$E(Y^2) = \Sigma y^2 g(y) = 0^2 \times \frac{4}{8} + 1^2 \times \frac{4}{8} = \frac{1}{2}$$

$$\therefore \quad \text{Var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

$$E(XY) = 0 \times (-1) \times \frac{1}{8} + 0 \times 1 \times \frac{3}{8} + 1 \times (-1) \times \frac{2}{8} + 1 \times 1 \times \frac{2}{8}$$

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

$$\text{Cov}(X, Y) = E(XY) - E(X)\,E(Y) = 0 - \frac{1}{4} \times \frac{1}{2} = -\frac{1}{8}$$

$$\therefore \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{-\frac{1}{8}}{\sqrt{\frac{15}{16} \times \frac{1}{4}}} = \frac{-1}{\sqrt{15}} = \frac{-1}{3{\cdot}873}$$

$$= -0{\cdot}2582$$

EXERCISE 10(b)

1. Write a brief note on the correlation table :

The following are the marks obtained by 24 students in a class test of Statistics and Mathematics :

Role No. of Students	:	1	2	3	4	5	6	7	8	9	10	11	12
Marks in Statistics	:	15	0	1	3	16	2	18	5	4	17	6	19
Marks in Mathematics	:	13	1	2	7	8	9	12	9	17	16	6	18
Roll No. of Students	:	13	14	15	16	17	18	19	20	21	22	23	24
Marks in Statistics	:	14	9	8	13	10	13	11	11	12	18	9	7
Marks in Mathematics	:	11	3	5	4	10	11	14	7	18	15	15	3

Prepare a correlation table taking the magnitude of each class interval as four marks and the first class interval as "equal to 0 and less than 4". Calculate Karl Pearson's coefficient of correlation between the marks in Statistics and marks in Mathematics from the correlation table.

Ans. 0·5544.

2. An employment bureau asked applicants their weekly wages on jobs last held. The actual wages were obtained for 54 of them; and are recorded in the table below; x represents reported wage, y actual wage, and the entry in the table represents frequency. Find the correlation coefficient and comment on the significance of the computed value. [Four figure log table may be used].

$y \rightarrow$ / $\downarrow x$	15	20	25	30	35	40
40						2
35				3	5	
30			4	15		
25			20			
20		3	1			
15	1					

[Calcutta Univ. B.Sc. (Maths. Hons.), 1986]

3. Calculate the correlation coefficient from the following table :—

$x \rightarrow$ / $y \downarrow$	0—10	10—20	20—30	30—40
0—5	1	3	2	0
5—10	7	10	8	1
10—15	10	13	10	8
15—20	5	8	10	7
20—25	0	1	5	4

4. (*a*) Find the correlation coefficient between age and salary of 50 workers in a factory :

Age (in years) ↓	*Daily pay in rupees*				
	160—169	170—179	180—189	190—199	200—209
20—30	5	3	1	...	...
30—40	2	6	2	1	...
40—50	1	2	4	2	2
50—60	...	1	3	6	2
60—70	...	...	1	1	5

(*b*) Fnd the coefficient of correlation between the ages of 100 mothers and daughters :

Age of mothers in years (X)	*Age of daughters in years (Y)*					*Total*
	5—10	*10—15*	*15—20*	*20—25*	*25—30*	
15—25	6	3				9
25—35	3	16	10			29
35—45		10	15	7		32
45—55			7	10	4	21
55—65				4	5	9
Total	9	29	32	21	9	100

[***Madras Univ. B.Sc. (Main Maths.), 1991***]

5. Given the following frequency distribution of (X, Y) :

$X \rightarrow$ / $Y \downarrow$	5	10	Total
10	30	20	50
20	20	30	50
Total	50	50	100

find the frequency distribution of (U, V), where

$$U = \frac{X - 7{\cdot}5}{2{\cdot}5}, \; V = \frac{Y - 15}{-5}.$$

What shall be the relationship between the correlation coefficients between X, Y, and U, V ?

6. (*a*) Find the coefficient of correlation between X and Y for the following table :

$Y \rightarrow$ / $X \downarrow$	y_1	y_2	Total
x_1	p_{11}	p_{12}	P
x_2	p_{21}	p_{22}	Q
Total	P'	Q'	1

(*b*) Consider the following probability distribution :

$Y \rightarrow$ / $X \downarrow$	0	1	2
0	0·1	0·2	0·1
1	0·2	0·3	0·1

Calculate $E(X)$, Var (X), Cov(X, Y) and $r\,(X, Y)$.

[***Delhi Univ. M.A. (Eco.), 1991***]

(*c*) Let (X, Y) have the p.m.f.

$$p(0, 1) = p(1, 0) = \frac{1}{3};\ p(0, -1) = p(-1, 0) = \frac{1}{6}.$$

Find $r(X, Y)$. Are X and Y independent ? For what values of k, $X + kY$ and $kX + Y$ are uncorrelated ?

10·5. Probable Error of Correlation Coefficient. If r is the correlation coefficient in a sample of n pairs of observations, then its *standard error* is given by

$$\text{S.E}(r) = \frac{1 - r^2}{\sqrt{n}}$$

Probable error of correlation coefficient is given by

$$\text{P.E.}(r) = 0{\cdot}6745 \times \text{S.E.} = 0{\cdot}6745 \frac{(1 - r^2)}{\sqrt{n}} \qquad \ldots(10{\cdot}6)$$

Probable error is an old measure for testing the reliability of an observed correlation coefficient. The reason for taking the factor 0·6745 is that in a normal distribution, the range $\mu \pm 0{\cdot}6745\ \sigma$ covers 50% of the total area. According to Secrist, "*The probable error of the correlation co-efficient is an amount which if added to and substracted from the mean correlation coefficient, produces amounts within which the chances are even that a coefficient of correlation from a series selected at random will fall.*"

If $r <$ P.E.(r), correlation is not at all significant. If $r > 6$ P.E.(r), it is definitely significant. A rigorous method (t-test) of testing the significance of an observed correlation coefficient will be discussed later in "*tests of significance*" in sampling [c.f. § 14·4·11].

Probable error also enables us to find the limits within which the population correlation coefficient can be expected to vary. The limits are $r \pm \text{P.E.}(r)$.

10·6. Rank Correlation. Let us suppose that a group of n individuals is arranged in order of merit or proficiency in possession of two characteristics A and B. These ranks in the two characteristics will, in general, be different. For example, if we consider the relation between intelligence and beauty, it is not necessary that a beautiful individual is intelligent also. Let (x_i, y_i); $i = 1, 2, \ldots, n$ be the ranks of the ith individual in two characteristics A and B respectively. Pearsonian coefficient of correlation between the ranks x_i's and y_i's is called the rank correlation coefficient between A and B for that group of individuals.

Assuming that no two individuals are bracketed equal in either classification, each of the variables X and Y takes the values $1, 2, \ldots, n$.

Hence $$\bar{x} = \bar{y} = \frac{1}{n}(1 + 2 + 3 + \ldots + n) = \frac{n+1}{2}$$

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}(1^2 + 2^2 + \ldots + n^2) - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}$$

$$\therefore \quad \sigma_X^2 = \frac{n^2-1}{12} = \sigma_Y^2$$

In general $x_i \neq y_i$. Let $d_i = x_i - y_i$

$$\therefore \quad d_i = (x_i - \bar{x}) - (y_i - \bar{y}) \qquad (\because \bar{x} = \bar{y})$$

Squaring and summing over i from 1 to n, we get

$$\Sigma d_i^2 = \Sigma\{(x_i - \bar{x}) - (y_i - \bar{y})\}^2$$

$$= \Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2 - 2\Sigma(x_i - \bar{x})(y_i - \bar{y})$$

Dividing both sides by n, we get

$$\frac{1}{n}\Sigma d_i^2 = \sigma_X^2 + \sigma_Y^2 - 2\,\text{Cov}\,(X, Y) = \sigma_X^2 + \sigma_Y^2 - 2\rho\,\sigma_X\sigma_Y$$

where ρ is the rank correlation coefficient between A and B.

$$\therefore \quad \frac{1}{n}\Sigma d_i^2 = 2\sigma_X^2 - 2\rho\sigma_X^2 \quad \Rightarrow \quad 1 - \rho = \frac{\Sigma d_i^2}{2n\sigma_X^2}$$

$$\Rightarrow \quad \rho = 1 - \frac{\sum_{i=1}^{n} d_i^2}{2n\sigma_X^2} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad \ldots(10\cdot7)$$

which is the *Spearman's formula for the rank correlation coefficient.*

Remark. We always have

$$\Sigma d_i = \Sigma(x_i - y_i) = \Sigma x_i - \Sigma y_i = n(\bar{x} - \bar{y}) = 0 \qquad (\because \bar{x} = \bar{y})$$

This serves as a check on the calculations.

10·6·1. Tied Ranks. If some of the individuals receive the same rank in a ranking of merit, they are said to be tied. Let us suppose that m of the individuals, say, $(k+1)^{th}$, $(k+2)^{th}$, ..., $(k+m)^{th}$ are tied. Then each of these m individuals is assigned a common rank, which is the arithmetic mean of the ranks $k+1, k+2, ..., k+m$.

Derivation of $\rho(X, Y)$. We have :

$$\rho(X, Y) = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{[\Sigma (X - \bar{X})^2 . \Sigma (Y - \bar{Y})^2]^{1/2}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} \quad ...(*)$$

where $x = X - \bar{X}, y = Y - \bar{Y}$.

If X and Y each takes the values 1, 2, ..., n, then we have

$$\bar{X} = (n+1)/2 = \bar{Y}$$

and $$n\sigma_X^2 = \Sigma x^2 = \frac{n(n^2-1)}{12} = \text{and } n\sigma_Y^2 = \Sigma y^2 = \frac{n(n^2-1)}{12} \quad ...(**)$$

Also $$\Sigma d^2 = \Sigma (X - Y)^2 = \Sigma [(X - \bar{X}) - (Y - \bar{Y}]^2 = \Sigma (x - y)^2$$

$$\Rightarrow \quad \Sigma d^2 = \Sigma x^2 + \Sigma y^2 - 2\Sigma xy$$

$$\Rightarrow \quad \Sigma xy = \frac{1}{2}[\Sigma x^2 + \Sigma y^2 - \Sigma d^2] \quad ...(***)$$

We shall now investigate the effect of common ranking, (in case of ties), on the sum of squares of the ranks. Let S^2 and S_1^2 denote the sum of the squares of untied and tied ranks respectively.

Then we have :

$$S^2 = (k+1)^2 + (k+2)^2 + ... + (k+m)^2$$
$$= mk^2 + (1^2 + 2^2 + ... + m^2) + 2k.(1 + 2 + ... + m)$$
$$= mk^2 + \frac{m(m+1)(2m+1)}{6} + mk(m+1)$$

$$S_1^2 = m \text{ (Average rank)}^2$$
$$= m\left[\frac{(k+1) + (k+2) + ... + (k+m)}{m}\right]^2$$
$$= m\left(k + \frac{m+1}{2}\right)^2 = mk^2 + \frac{m(m+1)^2}{4} + mk(m+1)$$

$$\therefore \quad S^2 - S_1^2 = \frac{m(m+1)}{12}[2(2m+1) - 3(m+1)] = \frac{m(m^2-1)}{12}$$

Thus the effect of tying m individuals (ranks) is to reduce the sum of the squares by $m(m^2-1)/12$, though the mean value of the ranks remains the same, *viz.*, $(n+1)/2$.

Suppose that there are s such sets of ranks to be tied in the X-series so that the total sum of squares due to them is

$$\frac{1}{12}\sum_{i=1}^{s} m_i (m_i^2 - 1) = \frac{1}{12}\sum_{i=1}^{s} (m_i^3 - m_i) = T_X, \text{ (say)} \quad ... (10·7a)$$

Similarly suppose that there are t such sets of ranks to be tied with respect to the other series Y so that sum of squares due to them is :

$$\frac{1}{12}\sum_{j=1}^{t} m_j' \cdot (m_j'^2 - 1) = \frac{1}{12}\sum_{j=1}^{t} (m_j'^3 - m_j') = T_Y \text{ , (say)} \qquad \ldots(10\cdot7b)$$

Thus, in the case of ties, the new sums of squares are given by :

$$n \text{ Var}'(X) = \Sigma x^2 - T_X = \frac{n(n^2-1)}{12} - T_X$$

$$n \text{ Var}'(Y) = \Sigma y^2 - T_Y = \frac{n(n^2-1)}{12} - T_Y$$

and $n \text{ Cov}'(X, Y) = \frac{1}{2}[\Sigma x^2 - T_X + \Sigma y^2 - T_Y - \Sigma d^2]$ [From (***)]

$$= \frac{1}{2}\left[\frac{n(n^2-1)}{12} - T_X + \frac{n(n^2-1)}{12} - T_Y - \Sigma d^2\right]$$

$$= \frac{n(n^2-1)}{12} - \frac{1}{2}\left[(T_X + T_Y) + \Sigma d^2\right]$$

$$\rho(X, Y) = \frac{\frac{n(n^2-1)}{12} - \frac{1}{2}\left[T_X + T_Y + \Sigma d^2\right]}{\left[\frac{n(n^2-1)}{12} - T_X\right]^{1/2}\left[\frac{n(n^2-1)}{12} - T_Y\right]^{1/2}}$$

$$= \frac{\frac{n(n^2-1)}{6} - \left[\Sigma d^2 + T_X + T_Y\right]}{\left[\frac{n(n^2-1)}{6} - 2T_X\right]^{1/2}\left[\frac{n(n^2-1)}{6} - 2\,T_Y\right]^{1/2}} \qquad \ldots(10\cdot7c)$$

where T_X and T_Y are given by (10·7a) and (10·7b).

Remark. If we adjust only the covariance term *i.e.*, Σxy and not the variances σ_X^2 (or Σx^2) and σ_Y^2 (or Σy^2) for ties, then the formula (10·7c) reduces to :

$$\rho(X, Y) = \frac{\frac{n(n^2-1)}{6} - (\Sigma d^2 + T_X + T_Y)}{n(n^2-1)/6}$$

$$= 1 - \frac{6\,[\Sigma d^2 + T_X + T_Y]}{n(n^2-1)}, \qquad \ldots(10\cdot7d)$$

a formula which is commonly used in practice for numerical problems. For illustration, see Example 10·18.

Example 10·16. *The ranks of same 16 students in Mathematics and Physics are as follows. Two numbers within brackets denote the ranks of the students in Mathematics and Physics.*

(1, 1) (2, 10) (3, 3) (4, 4) (5, 5) (6, 7) (7, 2) (8, 6) (9, 8) (10, 11) (11, 15) (12, 9) (13, 14) (14, 12) (15, 16) (16, 13).

Calculate the rank correlation coefficient for proficiencies of this group in Mathematics and Physics.

Solution.

Ranks in Maths. (X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
Ranks in Physics(Y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13	
$d = X - Y$	0	−8	0	0	0	−1	5	2	1	−1	−4	3	−1	2	−1	3	0
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9	136

Rank correlation coefficient is given by

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 136}{16 \times 255} = 1 - \frac{1}{5} = \frac{4}{5} = 0{\cdot}8$$

Example 10·17. *Ten competitors in a musical test were ranked by the three judges A, B and C in the following order :*

Ranks by A :	*1*	*6*	*5*	*10*	*3*	*2*	*4*	*9*	*7*	*8*
Ranks by B :	*3*	*5*	*8*	*4*	*7*	*10*	*2*	*1*	*6*	*9*
Ranks by C :	*6*	*4*	*9*	*8*	*1*	*2*	*3*	*10*	*5*	*7*

Using rank correlation method, discuss which pair of judges has the nearest approach to common likings in music.

Solution. Here $n = 10$

Ranks by A (X)	*Ranks by B (Y)*	*Ranks by C (Z)*	$d_1 = X - Y$	$d_2 = X - Z$	$d_3 = Y - Z$	d_1^2	d_2^2	d_3^2
1	3	6	− 2	− 5	−3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	− 3	− 4	−1	9	16	1
10	4	8	6	2	−4	36	4	16
3	7	1	− 4	2	6	16	4	36
2	10	2	− 8	0	8	64	0	64
4	2	3	2	1	−1	4	1	1
9	1	10	8	−1	−9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	− 1	1	2	1	1	4
Total			$\Sigma d_1 = 0$	$\Sigma d_2 = 0$	$\Sigma d_3 = 0$	$\Sigma d_1^2 = 200$	$\Sigma d_2^2 = 60$	$\Sigma d_3^2 = 214$

$$\rho(X, Y) = 1 - \frac{6 \Sigma d_1^2}{n(n^2 - 1)} = 1 - \frac{6 \times 200}{10 \times 99} = 1 - \frac{40}{33} = -\frac{7}{33}$$

$$\rho(X, Z) = 1 - \frac{6 \Sigma d_2^2}{n(n^2 - 1)} = 1 - \frac{6 \times 60}{10 \times 99} = 1 - \frac{4}{11} = \frac{7}{11}$$

$$\rho(Y, Z) = 1 - \frac{6\Sigma d_3^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = -\frac{49}{165}$$

Since $\rho(X, Z)$ is maximum, we conclude that the pair of judges A and C has the nearest approach to common likings in music.

10·6·2. Repeated Ranks (Continued). If any two or more individuals are bracketed equal in any classification with respect to characteristics A and B, or if there is more than one item with the same value in the series, then the Spearman's formula (10·7) for calculating the rank correlation coefficient breaks down, since in this case each of the variables X and Y does not assume the values 1, 2, ..., n and consequently, $\bar{x} \neq \bar{y}$.

In this case, common ranks are given to the repeated items. This common rank is the average of the ranks which these items would have assumed if they were sightly different from each other and the next item will get the rank next to the ranks already assumed. As a result of this, following adjustment or correction is made in the rank correlation formula [c.f. (10·7*c*) and (10·7*d*)].

In the formula, we add the factor $\frac{m(m^2-1)}{12}$ to Σd^2, where m is the number of times an item is repeated. This correction factor is to be added for each repeated value in both the X-series and Y-series.

Example 10·18. *Obtain the rank correlation coefficient for the following data*:

X	:	*68*	*64*	*75*	*50*	*64*	*80*	*75*	*40*	*55*	*64*
Y	:	*62*	*58*	*68*	*45*	*81*	*60*	*68*	*48*	*50*	*70*

Solution.

CALCULATIONS FOR RANK CORRELATION

X	*Y*	*Rank X* (*x*)	*Rank Y* (*y*)	$d = x - y$	d^2
68	62	4	5	−1	1
64	58	6	7	−1	1
75	68	2·5	3·5	−1	1
50	45	9	10	−1	1
64	81	6	1	5	25
80	60	1	6	−5	25
75	68	2·5	3·5	−1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
				$\Sigma d = 0$	$\Sigma d^2 = 72$

In the X-series we see that the value 75 occurs 2 times. The common rank given to these values is 2·5 which is the average of 2 and 3, the ranks which these values would have taken if they were different. The next value 68, then gets the next rank which is 4. Again we see that value 64 occurs thrice. The common rank given to it is 6 which is the average of 5, 6 and 7. Similarly in

the Y-series, the value 68 occurs twice and its common rank is 3·5 which is the average of 3 and 4. As a result of these common rankings, the formula for 'ρ' has to be corrected. To Σd^2 we add $\frac{m(m^2-1)}{12}$ for each value repeated, where m is the number of times a value occurs. In the X-series the correction is to be applied twice, once for the value 75 which occurs twice ($m = 2$) and then for the value 64 which occurs thrice ($m = 3$). The total correction for the X-series is

$$\frac{2(4-1)}{12} + \frac{3(9-1)}{12} = \frac{5}{2}$$

Similarly, this correction for the Y-series is $\frac{2(4-1)}{12} = \frac{1}{2}$, as the value 68 occurs twice.

Thus $$\rho = 1 - \frac{6\left[\Sigma d^2 + \frac{5}{2} + \frac{1}{2}\right]}{n(n^2-1)} = 1 - \frac{6(72+3)}{10 \times 99} = 0{\cdot}545$$

10·6·3. Limits for the Rank Correlation Coefficient. Spearman's rank correlation coefficient is given by

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

'ρ' is maximum, if $\sum_{i=1}^{n} d_i^2$ is minimum, *i.e.*, if each of the deviations d_i is minimum. But the minimum value of d_i is zero in the particular case $x_i = y_i$, *i.e.*, if the ranks of the ith individual in the two characteristic are equal. Hence the maximum value of ρ is + 1, *i.e.*, $\rho \leq 1$.

'ρ' is minimum , if $\sum_{i=1}^{n} d_i^2$ is maximum, *i.e.*, if each of the deviations d_i is maximum which is so if the ranks of the n individuals in the two characteristics are in the opposite directions as given below :

x	1	2	3	...	...	$n-1$	n
y	n	$n-1$	$n-2$	...	...	2	1

...(*)

Case 1. Suppose n is odd and equal to $(2m + 1)$ then the values of d are :

$d : 2m, 2m-2, 2m-4, \ldots, 2, 0, -2, -4, \ldots -(2m-2), -2m.$

$$\therefore \sum_{i=1}^{n} d_i^2 = 2\{(2m)^2 + (2m-2)^2 + \ldots + 4^2 + 2^2\}$$

$$= 8\{m^2 + (m-1)^2 + \ldots + 1^2\} = \frac{8m\,(m+1)\,(2m+1)}{6}$$

$$\text{Hence } \rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{8m(m+1)(2m+1)}{(2m+1)\{(2m+1)^2-1\}}$$

$$= \frac{8m(m+1)}{(4m2+4m)} = 1 - \frac{8m(m+1)}{4m(m+1)} = -1$$

Case II. Let n be even and equal to $2m$, (say).

Then the values of d are

$(2m-1), (2m-3), \ldots, 1, -1, -3, \ldots, -(2m-3), -(2m-1)$

$$\therefore \quad \Sigma d_i^2 = 2\{(2m-1)^2 + (2m-3)^2 + \ldots + 1^2\}$$

$$= 2[\{(2m)^2 + (2m-1)^2 + (2m-2)^2 + \ldots + 2^2 + 1^2\}$$

$$-\{(2m)^2 + (2m-2)^2 + \ldots + 4^2 + 2^2\}]$$

$$= 2[1^2 + 2^2 + \ldots + (2m)^2 - \{2^2m^2 + 2^2(m-1)^2 + \ldots + 2^2\}]$$

$$= 2\left[\frac{2m\,(2m+1)(4m+1)}{6} - 4\,\frac{m(m+1)(2m+1)}{6}\right]$$

$$= \frac{2m}{3}\left[(2m+1)(4m+1) - 2(m+1)(2m+1)\right]$$

$$= \frac{2m}{3}\left[(2m+1)(4m+1-2m-2)\right]$$

$$= \frac{2m}{3}(2m+1)(2m-1) = \frac{2m(4m^2-1)}{3}$$

$$\therefore \quad \rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{4m(4m^2-1)}{2m(4m^2-1)} = -1$$

Thus the limits for rank correlation coefficient are given by $-1 \le \rho \le 1$.

Aliter. For an alternate and simpler proof for obtaining the minimum value of ρ, from (*) onward, proceed as in Hint to Question Number 9 of Exercise 10(*c*).

Remarks on Spearman's Rank Correlation Coefficient.

1. $\Sigma d = \Sigma x - \Sigma y = n(\bar{x} - \bar{y}) = 0$, which provides a check for numerical calculations.

2. Since Spearman's rank correlation coefficient ρ is nothing but Pearsonian correlation coefficient between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.

3. Karl Pearson's correlation coefficient assume that the parent population from which sample observations are drawn is normal. If this assumption is violated then we need a measure which is distribution free (or non-parametric). A distribution-free measure is one which does not make any assumptions about the parameters of the population. Spearman's ρ is such a measure (*i.e.*, distribution-free), since no strict assumptions are made about the form of the population from which sample observations are drawn.

4. Spearman' formula is easy to understand and apply as compared with Karl Pearson's formula. The value obtained by the two formulae, *viz.*, Pearsonian r and Spearman's ρ, are generally different. The difference arises due to the fact that when ranking is used instead of full set of observations, there is

always some loss of information. Unless many ties exist, the coefficient of rank correlation should be only slightly lower than the Pearsonian coefficient.

5. Spearman's formula is the only formula to be used for finding correlation coefficient if we are dealing with qualitative characteristics which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of extreme observations, Spearman's formula is preferred to Pearson's formula.

6. Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution (Correlation Table). For $n > 30$, this formula should not be used unless the ranks are given, since in the contrary case the calculations are quite time-consuming.

EXERCISE 10(c)

1. Prove that Spearman's rank correlation coefficient is given by

$1 - \frac{6 \sum d_i^2}{n^3 - n}$, where d_i denotes the difference between the ranks of ith individual.

2. (*a*) Explain the difference between product moment correlation coefficient and rank correlation coefficient.

(*b*) The rankings of ten students in two subjects *A* and *B* are as follows :

A	:	3	5	8	4	7	10	2	1	6	9
B	:	6	4	9	8	1	2	3	10	5	7

Find the correlation coefficient.

3. (*a*) Calculate the coefficient of correlation for ranks from the following data :

(*X*, *Y*) : (5, 8), (10, 3), (6, 2), (3, 9), (19, 12), (5, 3), (6, 17), (12, 18), (8, 22), (2, 12), (10, 17), (19, 20).

[***Calicut Univ. B.Sc. (Subs. Stat.), Oct. 1991***]

(*b*) Ten recruits were subjected to a selection test to ascertain their suitability for a certain course of training. At the end of training they were given a proficiency test.

The marks secured by recruits in the selection test (*X*) and in the proficiency test (*Y*) are given below :—

Serial No.	:	1	2	3	4	5	6	7	8	9	10
X	:	10	15	12	17	13	16	24	14	22	20
Y	:	30	42	45	46	33	34	40	35	39	38

Calculate product moment correlation coefficient and rank correlation coefficient. Why are two coefficients different ?

4. (*a*) The I.Q.'s of a group of 6 persons were measured, and they then sat for a certain examination. Their I.Q.'s and examination marks were as follows :

Person :	*A*	*B*	*C*	*D*	*E*	*F*
I.Q. :	110	100	140	120	80	90
Exam. marks :	70	60	80	60	10	20

Compute the coefficients of correlation and rank correlation. Why are the correlation figures obtained different ?

Ans. 0·882 and 0·9.

The difference arises due to the fact that when ranking is used instead of the full set of observations, there is always some loss of information.

(*b*) The value of ordinary correlation (r) for the following data is 0·636 :—

X :	·05	·14	·24	·30	·47	·52	·57	·61	·67	·72
Y :	1·08	1·15	1·27	1·33	1·41	1·46	1·54	2·72	4·01	9·63

(*i*) Calculate Spearman's rank-correlation (ρ) for this data.

(*ii*) What advantage of ρ was brought out in this example ?

4. Ten competitors in a beauty contest are ranked by three judges as follows :

	Competitors									
Judges :	1	2	3	4	5	6	7	8	9	10
A :	6	5	3	10	2	4	9	7	8	1
B :	5	8	4	7	10	2	1	6	9	3
C :	4	9	8	1	2	3	10	5	7	6

Discuss which pair of judges has the nearest approach to common tastes of beauty.

5. A sample of 12 fathers and their eldest sons gave the following data about their height in inches :

Father :	65	63	67	64	68	62	70	66	68	67	69	71
Son :	68	66	68	65	69	66	68	65	71	67	68	70

Calculate coefficient of rank correlation. **(Ans.** 0·7220)

6. The coefficient of rank correlation between marks in Statistics and marks in Mathematics obtained by a certain group of students is 0·8. If the sum of the squares of the difference in ranks is given to be 33, find the number of student in the group **(Ans.** 10). ***[Madras Univ. B.Sc., 1990]***

7. The coefficient of rank correlation of the marks obtained by 10 students in Maths and Statistics was found to be 0·5. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

Hint. $$0{\cdot}5 = 1 - \frac{6\Sigma d^2}{10 \times 99}$$

$\Rightarrow$ $$\Sigma d^2 = \frac{990}{6 \times 2} = 82{\cdot}5$$

Since one difference was wrongly taken as 3 instead of 7, the correct value of Σd^2 is given by

$$\text{Corrected } \Sigma d^2 = 82{\cdot}5 - (3)^2 + (7)^2 = 122{\cdot}5$$

$\therefore$ $$\text{Corrected } \rho = 1 - \frac{6 \times 122{\cdot}5}{10 \times 99} = 0{\cdot}2576$$

8. If d_i be the difference in the ranks of the ith individual in two different characteristics, then show that the maximum value of $\sum_{i=1}^{n} d_i^2$ is $\frac{1}{3}(n^3 - n)$. Hence or otherwise, show that rank correlation coefficient lies between -1 and $+1$. [*Delhi Univ. B.Sc. (Maths. Hons.), 1986*]

9. Let $x_1, x_2, \ldots, x_n$ be the ranks of n individuals according to a character A and $y_1, y_2, \ldots, y_n$ be the ranks of the same individuals according to other character B. Obviously $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ are permutations of 1, 2, ..., n. It is given that $x_i + y_i = 1 + n$, for $i = 1, 2, \ldots, n$. Show that the value of the rank correlation coefficient ρ between the characters A and B is -1.

Hint. We are given $\quad x_i + y_i = n + 1 \; \forall i = 1, 2, \ldots, n$

Also $\quad x_i - y_i = d_i$

$$\therefore \quad 2x_i = n + 1 + d_i \Rightarrow d_i = 2x_i - (n + 1)$$

$$\therefore \quad \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [4x_i^2 + (n + 1)^2 - 2(n + 1)2x_i]$$

$$= 4\frac{n(n + 1)(2n + 1)}{6} + n(n + 1)^2 - \frac{4(n + 1)n(n + 1)}{2}$$

$$= \frac{n(n^2 - 1)}{3}$$

$$\therefore \quad \rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} = -1$$

Remark. From Spearmans' formula we note that ρ will be minimum if Σd_i^2 is maximum, which will be so if the ranks X and Y are in opposite directions as given below :

x	1	2	3	...	n
y	n	$n-1$	$n-2$	...	1

This gives us

$$x_i + y_i = n + 1, \; i = 1, 2, \ldots, n.$$

Hence the value of $\rho = -1$ obtained above is minimum value of ρ.

10. Show that in a ranked bivariate distribution in which no ties occur and in which the variables are independent

(*a*) $\sum_i d_i^2$ is always even, and

(*b*) there are not more than $\frac{1}{6}(n^3 - n) + 1$ possible values of r.

11. Show that if X, Y be identically distributed with common probability mass function : $P(X = k) = \frac{1}{N}$, for $k = 1, 2, \ldots, N; N > 1$,

then $\rho_{X,Y}$, the correlation coefficient between X and Y, is given by

$$1 - \frac{6E(X - Y)^2}{N^2 - 1}$$

[Delhi Univ. B.Sc. (Maths Hons.), 1992]

10·7. Regression. The term "*regression*" literally means "*stepping back towards the average*". It was first used by a British biometrician Sir Francis Galton (1822—1911), in connection with the inheritance of stature. Galton found that the offsprings of abnormally tall or short parents tend to "regress" or "step back" to the average population height. But the term "regression" as now used in Statistics is only a convenient term without having any reference to biometry.

Definition. *Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.*

In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called *dependent variable* and the variable which influences the values or is used for prediction, is called *independent variable.* In regression analysis independent variable is also known as *regressor or predictor* or *explanatory variable* while the dependent variable is also known as *regressed* or *explained variable.*

10·7·1. Lines of Regression. If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram will cluster round some curve called the "*curve of regression*". If the curve is a straight line, it is called the line of regression and there is said to be *linear regression* between the variables, otherwise regression is said to be *curvilinear.*

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus the line of regression is the line of "*best fit*" and is obtained by the *principles of least squares.*

Let us suppose that in the bivariate distribution (x_i, y_i); $i = 1, 2, \ldots, n$; Y is dependent variable and X is independent variable. Let the line of regression of Y on X be $Y = a + bX$.

According to the principle of least squares, the normal equations for estimating a and b are {*c.f.* (9·2*a*)}

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \qquad \ldots(10\cdot8)$$

and

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \qquad \ldots(10\cdot9)$$

From (10·8) on dividing by n, we get

$$\bar{y} = a + b\bar{x} \qquad \ldots(10\cdot10)$$

Thus the line of regression of Y on X passes through the point $(\bar{x}, \bar{y})$.

Now

$$\mu_{11} = \text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} \implies \frac{1}{n} \sum_{i=1}^{n} x_i y_i = \mu_{11} + \bar{x}\,\bar{y} \qquad \ldots(10\cdot11)$$

Also $$\sigma_X^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \implies \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \sigma_X^2 + \bar{x}^2 \qquad \ldots(10\cdot11a)$$

Dividing (10·9) by n and using (10·11) and (10·11a), we get

$$\mu_{11} + \bar{x}\,\bar{y} = a\bar{x} + b(\sigma_X^2 + \bar{x}^2) \qquad \ldots(10\cdot12)$$

Multiplying (10·10) by $\bar{x}$ and then subtracting from (10·12), we get

$$\mu_{11} = b\sigma_X^2 \implies b = \frac{\mu_{11}}{\sigma_X^2} \qquad \ldots(10\cdot13)$$

Since 'b' is the slope of the line of regression of Y on X and since the line of regression passes through the point $(\bar{x}, \bar{y})$, its equation is

$$Y - \bar{y} = b(X - \bar{x}) = \frac{\mu_{11}}{\sigma_X^2}(X - \bar{x}) \qquad \ldots(10\cdot14)$$

$$\implies \qquad Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \qquad \ldots(10\cdot14a)$$

Starting with the equation $X = A + BY$ and proceeding similarly or by simply interchanging the variables X and Y in (10·14) and (10·14a), the equation of the line of regression of X on Y becomes

$$X - \bar{x} = \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{y}) \qquad \ldots(10\cdot15)$$

$$\implies \qquad X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y}) \qquad \ldots(10\cdot15a)$$

Aliter. The straight line defined by

$$Y = a + bX \qquad \ldots(i)$$

and satisfying the residual (least square) condition

$$S = E[(Y - a - bX)^2] = \text{Minimum} \qquad \ldots(ii)$$

for variations in a and b, is called the line of regression of Y on X.

The necessary and sufficient conditions for a minima of S, subject to variations in a and b are :

$$(i)\ \frac{\partial S}{\partial a}=0,\quad \frac{\partial S}{\partial b}=0 \text{ and} \qquad \ldots(*)$$

$$(ii)\ \Delta=\begin{vmatrix} \frac{\partial^2 S}{\partial a^2} & \frac{\partial^2 S}{\partial a \partial b} \\ \frac{\partial^2 S}{\partial b\,\partial a} & \frac{\partial^2 S}{\partial b^2} \end{vmatrix} > 0 \text{ and } \frac{\partial^2 S}{\partial a^2} > 0 \qquad \ldots(**)$$

Using (*), we get

$$\frac{\partial S}{\partial a} = -2\,E\,[Y - a - bX] = 0 \qquad \ldots(iii)$$

$$\frac{\partial S}{\partial b} = -2\,E\,[X(Y - a - bX)] = 0 \qquad \ldots(iv)$$

$$\Rightarrow \quad E(Y) = a + bE(X) \ldots(v) \text{ and } E(XY) = aE(X) + bE(X^2) \qquad \ldots(vi)$$

Equation (v) implies that the line (i) of regression of Y on X passes through the mean value $[E(X),\ E(Y)]$.

Multiplying (v) by $E(X)$ and substracting from (vi), we get

$$E(XY) - E(X)E(Y) = b[E(X^2) - \{E(X)\}^2]$$

$$\Rightarrow \quad \text{Cov}\,(X, Y) = b\,\sigma_X^2 \quad \Rightarrow \quad b = \frac{\text{Cov}\,(X, Y)}{\sigma_X^2} = \frac{r\sigma_Y}{\sigma_X} \qquad \ldots(vii)$$

Subtracting (v) from (i) and using (vii), we obtain the equation of line of regression of Y on X as :

$$Y - E(Y) = \frac{\text{Cov}\,(X, Y)}{\sigma_X^2}[X - E(X)] \Rightarrow Y - E(Y) = \frac{r\sigma_Y}{\sigma_X}[X - E(X)]$$

Similarly, the straight line defined by $X = A + BY$ and satisfying the residual condition

$$E[X - A - BY]^2 = \text{Minimum},$$

is called the line of regression of X on Y.

Remarks 1. We note that

$$\frac{\partial^2 S}{\partial a^2} = 2 > 0, \text{ and}$$

$$\frac{\partial^2 S}{\partial b^2} = 2E(X^2) \text{ and } \frac{\partial^2 S}{\partial a \partial b} = 2E(X)$$

Substituting in (**), we have

$$\Delta = \frac{\partial^2 S}{\partial a^2}\cdot\frac{\partial^2 S}{\partial b^2} - \left(\frac{\partial^2 S}{\partial a \partial b}\right)^2$$

$$= 4\,[E(X^2) - (E(X))^2] = 4.\,\sigma_X^2 > 0$$

Hence the solution of the least square equations (iii) and (iv), in fact, provides a minima of S.

2. The regression equation (10·14a) implies that the line of regression of Y on X passes through the point $(\bar{x}, \bar{y})$. Similarly (10·15a) implies that the line of regression of X on Y also passes through the point $(\bar{x}, \bar{y})$. Hence both the lines of regression pass through the point $(\bar{x}, \bar{y})$. In other words, the mean

values $(\bar{x}, \bar{y})$ can be obtained as the point of intersection of the two regression lines.

3. *Why two lines of Regression ?* There are always two lines of regression, one of Y on X and the other of X on Y. The line of regression of Y on X (10·14a) is used to estimate or predict the value of Y for any given value of X, *i.e.*, when Y is a dependent variable and X is an independent variable. The estimate so obtained will be best in the sense that it will have the minimum possible error as defined by the principle of least squares. We can also obtain an estimate of X for any given value of Y by using equation (10·14a) but the estimate so obtained will not be best since (10·14a) is obtained on minimising the sum of the squares of errors of estimates in Y and not in X. Hence to estimate or predict X for any given value of Y, we use the regression equation of X on Y (10·15a) which is derived on minimising the sum of the squares of errors of estimates in X. Here X is a dependent variable and Y is an independent variable. The two regression equations are not reversible or interchangeable because of the simple reason that the basis and assumptions for deriving these equations are quite different. The regression equation of Y on X is obtained on minimising the sum of the squares of the errors parallel to the Y-axis while the regression equation of X on Y is obtained on minimising the sum of squares of the errors parallel to the X-axis.

In a particular case of perfect correlation, positive or negative, *i.e.*, $r \pm 1$, the equation of line of regression of Y on X becomes :

$$Y - \bar{y} = \pm \frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

$$\Rightarrow \quad \frac{Y - \bar{y}}{\sigma_Y} = \pm \left(\frac{X - \bar{x}}{\sigma_X}\right) \qquad \ldots(10\cdot16)$$

Similarly, the equation of the line of regression of X on Y becomes :

$$X - \bar{x} = \pm \frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

$$\Rightarrow \quad \frac{Y - \bar{y}}{\sigma_Y} = \pm \left(\frac{X - \bar{x}}{\sigma_X}\right),$$

which is same as (10·16).

Hence in case of perfect correlation, $(r = \pm 1)$, both the lines of regression coincide. Therefore, in general, we always have two lines of regression except in the particular case of perfect correlation when both the lines coincide and we get only one line.

10·7·2. Regression Curves. In modern terminology, the conditional mean $E(Y \mid X = x)$ for a continuous distribution is called the regression function of Y on X, and the graph of this function of x is known as the regression curve of Y on X or sometimes the regression curve for the mean of Y. Geometrically, the regression function represents the y co-ordinate of the centre of mass of the bivariate probability mass in the infinitesimal vertical strip bounded by x and $x + dx$.

Similarly, the regression function of X on Y is $E(X \mid Y = y)$ and the graph of this function of y is called the regression curve (of the mean) of X on Y.

In case a regression curve is a straight line, the corresponding regression is said to be *linear*. If one of the regressions is linear, it does not however follow that the other is also linear. For illustration, See Example 10·21.

Theorem 10·4. *Let (X, Y) be a two-dimensional random variable with $E(X) = \bar{X}$, $E(Y) = \bar{Y}$, $V(X) = \sigma_X^2$, $V(Y) = \sigma_Y^2$ and let $r = r(X, Y)$ be the correlation coefficient between X and Y. If the regression of Y on X is linear then*

$$E(Y \mid X) = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \qquad \ldots(10\cdot16a)$$

Similarly, if the regression of X on Y is linear, then

$$E(X \mid Y) = \bar{X} + r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y}) \qquad \ldots(10\cdot16b)$$

Proof. Let the regression equation of Y on X be

$$E(Y \mid x) = a + bx \qquad \ldots(1)$$

But by definition,

$$E(Y \mid x) = \int_{-\infty}^{\infty} y\, f(y \mid x)\, dy = \int_{-\infty}^{\infty} y\, \frac{f(x, y)}{f_X(x)}\, dy$$

$$\therefore \qquad \frac{1}{f_X(x)} \int_{-\infty}^{\infty} y\, f(x, y)\, dy = a + bx \qquad \ldots(2)$$

Multiplying both sides of (2) by $f_X(x)$ and integrating w.r.t. x, we get

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x, y)\, dy dx = a\int_{-\infty}^{\infty} f_X(x)\, dx + b\int_{-\infty}^{\infty} x\, f_X(x) dx$$

$$\Rightarrow \qquad \int_{-\infty}^{\infty} y\, \left[\int_{-\infty}^{\infty} f(x, y)\, dx\right] dy = a + bE(X)$$

$$\Rightarrow \qquad \int_{-\infty}^{\infty} y\, f_Y(y) dy = a + bE(X)$$

$$i.e., \qquad E(Y) = a + bE(X) \quad \Rightarrow \quad \bar{Y} = a + b\bar{X} \qquad \ldots(3)$$

Multiplying both sides of (2) by $x f_X(x)$ and integrating w.r.t. x, we get

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x, y)\, ay\, dx = a\int_{-\infty}^{\infty} x\, f_X(x)\, dx + b\int_{-\infty}^{\infty} x^2 f_X(x)\, dx$$

$$\Rightarrow \qquad E(XY) = a\,E(X) + b\,E(X^2)$$

$$\Rightarrow \qquad \mu_{11} + \bar{X}\,\bar{Y} = a\bar{X} + b(\sigma_X^2 + \bar{X}^2) \qquad \ldots(4)$$

$[\because \mu_{11} = E(XY) - E(X)E(Y) = E(XY) - \bar{X}\,\bar{Y}$;

$$\sigma_X^2 = E(X^2) - \{E(X)\}^2 = E(X^2) - \bar{X}^2$$

Solving (3) and (4) simultaneously, we get

$$b = \frac{\mu_{11}}{\sigma_X^2} \text{ and } a = \bar{Y} - \frac{\mu_{11}}{\sigma_X^2}\bar{X}$$

Substituting in (1) and simplifying, we get the required equation of the line of regression of Y on X as

$$E(Y \mid x) = \bar{Y} + \frac{\mu_{11}}{\sigma_X^2}(x - \bar{X})$$

$$\Rightarrow \quad E(Y \mid X) = \bar{Y} + \frac{\mu_{11}}{\sigma_X^2}(X - \bar{X})$$

$$\Rightarrow \quad E(Y \mid X) = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

By starting with the line $E(X \mid y) = A + By$ and proceeding similarly we shall obtain the equation of the line of regression of X on Y as

$$E(X \mid Y) = \bar{X} + \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{Y}) = \bar{X} + r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

Example 10·19. *Given*

$$f(x, y) = xe^{-x(y+1)};\ x \geq 0,\ y \geq 0,$$

find the regression curve of Y on X. **[B.H. Univ. M.Sc., 1989]**

Solution. Marginal p.d.f. of X is given by

$$f_1(x) = \int_0^\infty f(x, y)\,dy = \int_0^\infty xe^{-x(y+1)}\,dy$$

$$= xe^{-x}\int_0^\infty e^{-xy}\,dy = xe^{-x}\left|\frac{e^{-xy}}{-x}\right|_0^\infty$$

$$= e^{-x},\ x \geq 0$$

Conditional p.d.f. of Y on X is given by

$$f(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{xe^{-x(y+1)}}{e^{-x}} = xe^{-xy},\ y \geq 0.$$

The regression curve of Y on X is given by

$$y = E(Y \mid X = x) = \int_0^\infty y\,f(y \mid x)\,dy = \int_0^\infty yxe^{-xy}\,dy$$

$$= x\left[\left| \frac{ye^{-xy}}{-x} \right|_0^\infty + \int_0^\infty \frac{e^{-xy}}{x}\, dy\right] = 0 + \left| \frac{e^{-xy}}{-x} \right|_0^\infty = \frac{1}{x}$$

i.e., $$y = \frac{1}{x} \quad \Rightarrow \quad xy = 1.$$

which is the equation of a rectangular hyperbola. Hence the regression of Y on X is not linear.

Example 10·20. *Obtain the regression equation of Y on X for the following distribution* :

$$f(x, y) = \frac{y}{(1+x)^4} \exp\left(-\frac{y}{1+x}\right);\ x,\ y \geq 0$$

Solution. Marginal p.d.f. of X is given by

$$f_1(x) = \int_0^\infty f(x, y)\, dy = \frac{1}{(1+x)^4} \int_0^\infty y\, e^{-y/(1+x)}\, dy$$

$$= \frac{1}{(1+x)^4} \cdot \Gamma 2 \cdot (1+x)^2 \qquad \text{(Using Gamma Integral)}$$

$$= \frac{1}{(1+x)^2};\ x \geq 0$$

The conditional p.d.f. of Y (for given X) is

$$f(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{y}{(1+x)^2} \exp\left(-\frac{y}{1+x}\right);\ y \geq 0$$

Regression equation of Y on X is given by

$$y = E(Y \mid X) = \int_0^\infty y\, f(y \mid x)\, dy = \frac{1}{(1+x)^2} \int_0^\infty y^2\, e^{-y/(1+x)}\, dx$$

$$= \frac{1}{(1+x)^2} \cdot \Gamma 3 \cdot (1+x)^3 \qquad \text{[Using Gamma Integral]}$$

$$\Rightarrow \qquad y = 2(1+x) \qquad [\because \Gamma 3 = 2! = 2]$$

Hence the regression of Y on X is linear.

Example 10·21. *Let (X, Y) have the joint p.d.f. given by*

$$f(x, y) = \begin{cases} 1, & \text{if } |y| < x,\ 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Show that the regression of Y on X is linear but regression of X on Y is not linear.

Solution. $|y| < x \Rightarrow -x < y < x$ and $x > |y|$.

The marginal p.d.f's $f_1(.)$ of X and $f_2(.)$ of Y are given by :

$$f_1(x) = \int_{-x}^{x} f(x, y)\, dy = \int_{-x}^{x} 1.\, dy = 2x;\ 0 < x < 1$$

$$f_2(y) = \int_{|y|}^{1} f(x, y)\, dx = \int_{|y|}^{1} 1.dx = 1 - |y|;\ -1 < y < 1$$

$$\therefore \quad f_1(x \mid y) = \frac{f(x, y)}{f_2(y)} = \frac{1}{1 - |y|} ; -1 \le y < 1, 0 < x < 1$$

$$= \begin{cases} \frac{1}{1 - y}, & 0 < y < 1 ; 0 < x < 1 \\ \frac{1}{1 + y}, & -1 < y < 0 ; 0 < x < 1 \end{cases}$$

$$f_2(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{1}{2x}, 0 < x < 1; |y| < x$$

$$E(Y \mid X = x) = \int_{-x}^{x} y \cdot f_2(y \mid x)\, dy = \int_{-x}^{x} \frac{y}{2x} dy = \frac{1}{4x} \cdot |y^2|_{-x}^{x} = 0$$

Hence the curve of regression of Y on X is $y = 0$, which is a straight line.

$$E(X|Y = y) = \int x f_1(x \mid y) dx$$

$$\therefore \quad E(X \mid Y = y) = \int_0^1 x \left(\frac{1}{1 - y} \right) dx = \frac{1}{2(1 - y)}, 0 < y < 1$$

and $$E(X|Y = y) = \int_0^1 x \left(\frac{1}{1 + y} \right) dx = \frac{1}{2(1 + y)}, -1 < y < 0$$

Hence the curve of regression of X on Y is

$$x = \begin{cases} \frac{1}{2(1 - y)}, & 0 < y < 1 \\ \frac{1}{2(1 + y)}, & -1 < y < 0, \end{cases}$$

which is not a straight line.

Example 10·22. *Variables X and Y have the joint p.d.f.*

$$f(x, y) = \tfrac{1}{3}(x + y), \; 0 \le x \le 1, \; 0 \le y \le 2.$$

Find :

(i) $r(X, Y)$

(ii) The two lines of regression

(iii) The two regression curves for the means.

Solution. The marginal p.d.f.'s of X and Y are given by :

$$f_1(x) = \int_0^2 f(x, y)\, dy = \frac{1}{3} \int_0^2 (x + y)\, dy = \frac{1}{3} \left| xy + \frac{y^2}{2} \right|_0^2$$

$$\Rightarrow \quad f_1(x) = \frac{2}{3}(1 + x) ; 0 \le x \le 1 \qquad \ldots(1)$$

$$f_2(y) = \int_0^1 f(x, y)\, dx = \frac{1}{3} \int_0^1 (x + y)\, dx = \frac{1}{3} \left| \frac{x^2}{2} + xy \right|_0^1$$

$$\Rightarrow \quad f_2(y) = \frac{1}{3} \left(\frac{1}{2} + y \right) ; 0 \le y \le 2 \qquad \ldots(2)$$

The conditional distributions are given by :

$$f_3(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{1}{2}\left(\frac{x + y}{1 + x}\right)$$

$$f_4(x \mid y) = \frac{f(x, y)}{f_2(y)} = \frac{2(x+ y)}{1 + 2y} \qquad \ldots(3)$$

$$E(Y|x) = \int_0^2 y \cdot f_3(y \mid x)\, dy = \frac{1}{2(1 + x)} \int_0^2 y(x + y)dy$$

$$= \frac{1}{2(1 + x)} \left|\frac{xy^2}{2} + \frac{y^3}{3}\right|_{y=0}^{y=2} = \frac{3x + 4}{3(x + 1)}$$

Similarly, we shall get

$$E(X \mid y) = \int_0^1 x f_4(x \mid y)\, dx = \frac{2}{1 + 2y} \int_0^1 (x^2 + xy)\, dx = \frac{2 + 3y}{3(1 + 2y)}$$

(*iii*) Hence the regression curves for means are :

$$y = E(Y|x) = \frac{3x + 4}{3(x + 1)} \quad \text{and} \quad x = E(X|y) = \frac{2 + 3y}{3(1 + 2y)}.$$

From the marginal distributions we shall get

$$E(X) = \int_0^1 x\, f_1(x)\, dx = \frac{5}{9}, \quad E(X^2) = \int_0^1 x^2 f_1(x)dx = \frac{7}{18}$$

$$\Rightarrow \quad \text{Var}(X) = \sigma_X^2 = \frac{7}{18} - \left(\frac{5}{9}\right)^2 = \frac{13}{162}$$

Similarly, we shall get

$$E(Y) = \frac{11}{9}, E(Y^2) = \frac{16}{9}; \sigma_Y^2 = \frac{16}{9} - \left(\frac{11}{9}\right)^2 = \frac{23}{81}$$

Also $$E(XY) = \int_0^1 \int_0^2 xy\, f(x, y)\, dxdy = \frac{1}{3} \int_0^1 \int_0^2 (x^2y + xy^2)\, dxdy$$

$$= \frac{1}{3}\left\{\left(\int_0^1 x^2 dx\right)\left(\int_0^2 y\, dy\right) + \left(\int_0^1 x\, dx\right)\left(\int_0^2 y^2 dy\right)\right\}$$

$$= \frac{1}{3}\left[\frac{1}{3} \times 2 + \frac{1}{2} \times \frac{8}{3}\right] = \frac{2}{3}$$

$$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)\, E(Y) = \frac{2}{3} - \frac{5}{9} \times \frac{11}{9} = -\frac{1}{81}$$

(*i*) $$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{-\frac{1}{81}}{\sqrt{\frac{13}{162} \times \frac{23}{81}}} = -\left(\frac{2}{299}\right)^{1/2}$$

(*ii*) The two lines of regression are :

$$Y - E(Y) = \frac{\text{Cov}(X, Y)}{\sigma_X^2}[X - E(X)] \quad \Rightarrow \quad Y - \frac{11}{9} = -\frac{2}{13}\left(X - \frac{5}{9}\right)$$

and $X - E(X) = \frac{\text{Cov}(X, Y)}{\sigma_Y^2}[Y - E(Y)] \Rightarrow X - \frac{5}{9} = -\frac{1}{23}\left(Y - \frac{11}{9}\right)$

10·7·3. Regression Coefficients. 'b', the slope of the line of regression of Y on X is also called the coefficient of regression of Y on X. It represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable X. More precisely, we write

$$b_{YX} = \text{Regression coefficient of } Y \text{ on } X = \frac{\mu_{11}}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X} \qquad \ldots(10\cdot17)$$

Similarly, the coefficient of regression of X on Y indicates the change in the value of variable X corresponding to a unit change in the value of variable Y and is given by

$$b_{XY} = \text{Regression coefficient of } X \text{ on } Y = \frac{\mu_{11}}{\sigma_Y^2} = r\frac{\sigma_X}{\sigma_Y} \qquad \ldots(10\cdot17a)$$

10·7·4. Properties of Regression Coefficients.

(*a*) *Correlation coefficient is the geometric mean between the regression coefficients.*

Proof. Multiplying (10·17) and (10·17*a*), we get

$$b_{XY} \times b_{YX} = r\frac{\sigma_X}{\sigma_Y} \times r\frac{\sigma_Y}{\sigma_X} = r^2$$

$$\therefore \qquad r = \pm\sqrt{b_{XY} \times b_{YX}} \qquad \ldots(10\cdot18)$$

Remark. We have

$$r = \frac{\mu_{11}}{\sigma_X \cdot \sigma_Y}, \quad b_{YX} = \frac{\mu_{11}}{\sigma_X^2} \text{ and } b_{XY} = \frac{\mu_{11}}{\sigma_Y^2}$$

It may be noted that the sign of correlation coefficient is the same as that of regression coefficients, since the sign of each depends upon the co-variance term μ_{11}. Thus if the regression coefficients are positive, 'r' is positive and if the regression coefficients are negative 'r' is negative.

From (10·18), we have

$$r = \pm\sqrt{b_{XY} \times b_{YX}}$$

the sign to be taken before the square root is that of the regression coefficients.

(*b*) *If one of the regression coefficients is greater than unity, the other must be less than unity.*

Proof. Let one of the regression coefficients (say) b_{YX} be greater than unity, then we have to show that $b_{XY} < 1$.

Now $\quad b_{YX} > 1 \Rightarrow \frac{1}{b_{YX}} < 1 \qquad \ldots(*)$

Also $\quad r^2 \le 1 \Rightarrow b_{YX} \cdot b_{XY} \le 1$

Hence $\quad b_{XY} \le \frac{1}{b_{YX}} < 1 \qquad$ [From (*)]

(*c*) *Arithmetic mean of the regression coefficients is greater than the correlation coefficient r, provided r > 0.*

Proof. We have to prove that $\frac{1}{2}(b_{YX} + b_{XY}) \geq r$

or $\frac{1}{2}\left(r\frac{\sigma_Y}{\sigma_X} + r\frac{\sigma_X}{\sigma_Y}\right) \geq r$ or $\frac{\sigma_Y}{\sigma_X} + \frac{\sigma_X}{\sigma_Y} \geq 2$ $(\because r > 0)$

$\Rightarrow$ $\sigma_Y^2 + \sigma_X^2 - 2\sigma_X\sigma_Y \geq 0$ *i.e.*, $(\sigma_Y - \sigma_X)^2 \geq 0$

which is always true, since the square of a real quantity is ≥ 0.

(d) Regression coefficients are independent of the change of origin but not of scale.

Proof. Let $U = \frac{X - a}{h}$, $V = \frac{Y - b}{k}$ $\Rightarrow$ $X = a + hU, Y = b + kV$,

where $a, b, h\ (> 0)$ and $k\ (> 0)$ are constants.

Then Cov $(X, Y) = hk$ Cov (U, V), $\sigma_X^2 = h^2\sigma_U^2$ and $\sigma_Y^2 = k^2\sigma_V^2$

$$b_{YX} = \frac{\mu_{11}}{\sigma_X^2} = \frac{hk \text{ cov } (U, V)}{h^2\sigma_U^2}$$

$$= \frac{k}{h} \cdot \frac{\text{cov } (U, V)}{\sigma_U^2} = \frac{k}{h} b_{VU}$$

Similarly, we can prove that

$$b_{XY} = (h/k)\, b_{UV}$$

10·7·5. Angle Between Two Lines of Regression. Equations of the lines of regression of Y on X, and X on Y are

$$Y - \bar{y} = r \cdot \frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \text{ and } X - \bar{x} = r \cdot \frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

Slopes of these lines are $r \cdot \frac{\sigma_Y}{\sigma_X}$ and $\frac{\sigma_Y}{r\sigma_X}$ respectively. If θ is the angle between the two lines of regression then

$$\tan\theta = \frac{r\frac{\sigma_Y}{\sigma_X} \sim \frac{\sigma_Y}{r\sigma_X}}{1 + r\frac{\sigma_Y}{\sigma_X} \cdot \frac{\sigma_Y}{r\sigma_X}} = \frac{r^2 \sim 1}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)$$

$$= \frac{1 \sim r^2}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right) \qquad (\because r^2 \leq 1)$$

$$\therefore \quad \theta = \tan^{-1}\left\{\frac{1 - r^2}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)\right\} \qquad \ldots(10\cdot19)$$

Case (i). $(r = 0)$. If $r = 0$, $\tan\theta = \infty$ $\Rightarrow$ $\theta = \frac{\pi}{2}$

Thus if the two variables are uncorrelated, the lines of regression become perpendicular to each other.

Case (ii). $(r = \pm 1)$. If $r = \pm 1$, $\tan\theta = 0$ $\Rightarrow$ $\theta = 0$ or π.

In this case the two lines of regression either coincide or they are parallel to each other. But since both the lines of regression pass through the point

$(\bar{x}, \bar{y})$, they cannot be parallel. Hence in the case of perfect correlation, positive or negative, the two lines of regression coincide.

Remarks 1. Whenever two lines intersect, there are two angles between them, one acute angle and the other obtuse angle. Further $\tan\theta > 0$ if $0 < \theta < \pi/2$, *i.e.*, θ is an acute angle and $\tan\theta < 0$ if $\pi/2 < \theta < \pi$, *i.e.*, θ is an obtuse angle and since $0 < r^2 < 1$, the acute angle (θ_1) and obtuse angle θ_2 between the two lines of regression are given by

$$\theta_1 = \text{Acute angle} = \tan^{-1}\left\{\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2} \cdot \frac{1 - r^2}{r}\right\}, r > 0$$

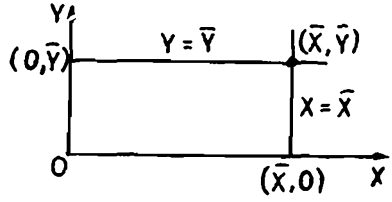
and

$$\theta_2 = \text{Obtuse angle} = \tan^{-1}\left\{\frac{\sigma_X \cdot \sigma_Y}{\sigma_X^2 + \sigma_Y^2} \cdot \frac{r^2 - 1}{r}\right\}, r > 0$$

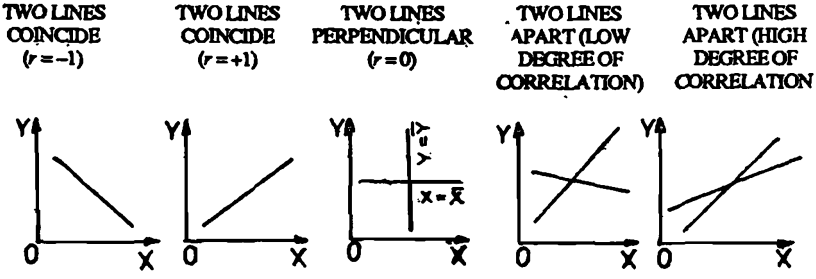
2. When $r = 0$, *i.e.*, variables X and Y are uncorrelated, then the lines of regressions of Y on X and X on Y are given respectively by : [From (10·14*a*) and (10·15*a*)]

$$Y = \bar{Y} \text{ and } X = \bar{X},$$

as shown in the adjoining diagram. Hence, in this case ($r = 0$), the lines of regression are perpendicular to each other and are parallel to X- axis and Y-axis respectively.



3. The fact that if $r = 0$ (variables uncorrelated), the two lines of regression are perpendicular to each and if $r = \pm 1$, $\theta = 0$, *i.e.*, the two lines coincide, leads us to the conclusion that for higher degree of correlation between the variables, the angle between the lines is smaller, *i.e.*,. the two lines of regression are nearer to each other. On the other hand, if the lines of regression make a larger angle, they indicate a poor degree of correlation between the variables and ultimately for $\theta = \pi/2$, $r = 0$, *i.e.*, the lines become perpendicular if no correlation exists between the variables. Thus by plotting the lines of regression on a graph paper, we can have an approximate idea about the degree of correlation between the two variables under study. Consider the following illustrations :



10·7·6. Standard Error of Estimate or Residual Variance. The equation of the line of regression of Y on X is

$$Y = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$\Rightarrow \quad \frac{Y - \bar{Y}}{\sigma_Y} = r\left(\frac{X - \bar{X}}{\sigma_X}\right)$$

The residual variance s_Y^2 is the expected value of the squares of deviations of the observed values of Y from the expected values as given by the line of regression of Y on X. Thus

$$s_Y^2 = E[Y - \{\bar{Y} + (r\sigma_Y (X - \bar{X})/\sigma_X)\}]^2$$

$$= \sigma_Y^2 E\left[\frac{Y - \bar{Y}}{\sigma_Y} - r\left(\frac{X - \bar{X}}{\sigma_X}\right)\right]^2 = \sigma_Y^2 E(Y^* - rX^*)^2$$

where X^* and Y^* are standardised variates so that

$E(X^{*2}) = 1 = E(Y^{*2})$ and $E(X^* Y^*) = r$.

$$\therefore \quad s_Y^2 = \sigma_Y^2[E(Y^{*2}) + r^2 E(X^{*2}) - 2r E(X^*Y^*)] = \sigma_Y^2 (1 - r^2)$$

$$\Rightarrow \quad s_Y = \sigma_Y (1 - r^2)^{1/2}$$

Similarly, the standard error of estimate of X is given by

$$s_X = \sigma_X (1 - r^2)^{1/2}$$

Remarks 1. Since s_X^2 or $s_Y^2 \geq 0$, it follows that

$$(1 - r^2) \geq 0 \quad \Rightarrow \quad |r| \leq 1$$

Hence $$-1 \leq r(X, Y) \leq 1$$

2. If $r = \pm 1$, $s_X = s_Y = 0$ so that each deviation is zero, and the two lines of regression are coincident.

3. Since, as $r^2 \to 1$, s_X and $s_Y \to 0$, it follows that departure of the value r^2 from unity indicates the departure of the relationship between the variables X and Y from linearity.

4. From the definition of linear regression, the minima condition implies that s_Y or s_X is the minimum variance.

10·7·7. Correlation Coefficient between Observed and Estimated Value. Here we will find the correlation between Y and

$$\hat{Y} = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

where $\hat{Y}$ is the estimated value of Y as given by the line of regression of Y on X, which is given by

$$r(Y, \hat{Y}) = \frac{\text{Cov}(\hat{Y}, Y)}{\sigma_Y \sigma_{\hat{Y}}}$$

We have

$$E(\hat{Y}) = E\left[\bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})\right] = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}E(X - \bar{X}) = \bar{Y}$$

$$\therefore \quad \text{Var}(\hat{Y}) = E[\hat{Y} - E(\hat{Y})]^2 = E\left[r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})\right]^2 = r^2\sigma_Y^2$$

$$\Rightarrow \qquad \hat{\sigma}_Y = r\,\sigma_Y$$

Also Cov $(Y, \hat{Y}) = E[\{Y - E(Y)\}\ \{\hat{Y} - E(\hat{Y})\}]$

$$= E\left[\{b(X - E(X))\}\left\{r\frac{\sigma_Y}{\sigma_X}(X - \overline{X})\right\}\right]$$

$$= br\frac{\sigma_Y}{\sigma_X}E[\{X - E(X)\}^2] = \left(r\frac{\sigma_Y}{\sigma_X}\right)^2\sigma_X^2 = r^2\sigma_Y^2$$

$$\therefore \qquad r(Y, \hat{Y}) = \frac{r^2\sigma_Y^2}{\sigma_Y r\,\sigma_Y} = r = r(X, Y)$$

Hence the correlation coefficient between observed and estimated value of Y is the same as the correlation coefficient between X and Y.

Example 10·23. *Obtain the equations of the lines of regression for the data in Example 10·1. Also obtain the estimate of X for Y = 70.*

Solution. Let $U = X - 68$ and $V = Y - 69$, then

$\overline{U} = 0, \overline{V} = 0, \sigma_U^2 = 4{\cdot}5, \sigma_V^2 = 5{\cdot}5$, Cov $(U, V) = 3$ and $r\,(U, V) = 0{\cdot}6$

Since correlation coefficient is independent of change of origin, we get

$$r = r(X, Y) = r(U, V) = 0{\cdot}6$$

We know that if $U = \frac{X - a}{h}$, $V = \frac{Y - b}{k}$, then

$$\overline{X} = a + h\overline{U}, \overline{Y} = b + k\overline{V}, \sigma_X = h\sigma_U \text{ and } \sigma_Y = k\,\sigma_Y$$

In our case $h = k = 1$, $a = 68$ and $b = 69$.

Thus $\overline{X} = 68 + 0 = 68, \overline{Y} = 69 + 0 = 69$

$$\sigma_X = \sigma_U = \sqrt{4{\cdot}5} = 2{\cdot}12 \text{ and } \sigma_Y = \sigma_V = \sqrt{5{\cdot}5} = 2{\cdot}35$$

Equation of line of regression of Y on X is

$$Y - \overline{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \overline{X})$$

i.e., $Y = 69 + 0{\cdot}6 \times \frac{2{\cdot}35}{2{\cdot}12}(X - 68) \Rightarrow Y = 0{\cdot}665\,X + 23{\cdot}78$

Equation of line of regression of X on Y is

$$X - \overline{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \overline{Y})$$

$$\Rightarrow \qquad X = 68 + 0{\cdot}6 \times \frac{2{\cdot}12}{2{\cdot}35}(Y - 69) \text{ } i.e., \text{ } X = 0{\cdot}54Y + 30{\cdot}74$$

To estimate X for given Y, we use the line of regression of X on Y. If $Y = 70$, estimated value of X is given by

$$\hat{X} = 0{\cdot}54 \times 70 + 30{\cdot}74 = 68{\cdot}54,$$

where $\hat{X}$ is estimate of X.

Example 10·24. *In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible :*

Variance of $X = 9$.

Regression equations : $8X - 10Y + 66 = 0$, $40X - 18Y = 214$.

What were (*i*) *the mean values of X and Y,*

(*ii*) *the correlation coefficient between X and Y, and*

(*iii*) *the standard deviation of Y ?*

[Punjab Univ. B.Sc. (Hons.), 1993]

Solution (*i*) Since both the lines of regression pass through the point $(\bar{X}, \bar{Y})$, we have $8\bar{X} - 10\bar{Y} + 66 = 0$, and $40\bar{X} - 18\bar{Y} = 214$.

Solving, we get $\bar{X} = 13, \bar{Y} = 17$.

(*ii*) Let $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$ be the lines of regression of Y on X and X on Y respectively. These equations can be put in the form :

$$Y = \frac{8}{10}X + \frac{66}{10} \text{ and } X = \frac{18}{40}Y + \frac{214}{40}$$

$\therefore$ b_{YX} = Regression coefficient of Y on $X = \frac{8}{10} = \frac{4}{5}$

and b_{XY} = Regression coefficient of X on $Y = \frac{18}{40} = \frac{9}{20}$

Hence $r^2 = b_{YX} \cdot b_{XY} = \frac{4}{5} \cdot \frac{9}{20} = \frac{9}{25}$

$\therefore$ $r = \pm\frac{3}{5} = \pm 0{\cdot}6$

But since both the regression coefficients are positive, we take $r = +0{\cdot}6$

(*iii*) We have $b_{YX} = r \cdot \frac{\sigma_Y}{\sigma_X} \Rightarrow \frac{4}{5} = \frac{3}{5} \times \frac{\sigma_Y}{3}$ $[\because \sigma_X^2 = 9 \text{ (Given)}]$

Hence $\sigma_Y = 4$

Remarks. 1. It can be verified that the values of $\bar{X} = 13$ and $\bar{Y} = 17$ as obtained in part (*i*) satisfy both the regression equations. In numerical problems of this type, this check should invariably be applied to ascertain the correctness of the answer.

2. If we had assumed that $8X - 10Y + 66 = 0$, is the equation of the line of regression of X on Y and $40X - 18Y = 214$ is the equation of line of regression of Y on X, then we get respectively :

$$8X = 10Y - 66 \text{ and } 18Y = 40X - 214$$

$\Rightarrow$ $$X = \frac{10}{8}Y - \frac{66}{8} \text{ and } Y = \frac{40}{18}X - \frac{214}{18}$$

$\Rightarrow$ $$b_{XY} = \frac{18}{8} \text{ and } b_{YX} = \frac{40}{18}$$

$\therefore$ $$r^2 = b_{XY} \cdot b_{YX} = \frac{10}{8} \times \frac{40}{18} = 2{\cdot}78$$

But since r^2 always lies between 0 and 1, our supposition is wrong.

Example 10·25. *Find the most likely price in Bombay corresponding to the price of Rs. 70 at Calcutta from the following :*

	Calcutta	*Bombay*
Average price	*65*	*67*
Standard deviation	*2·5*	*3·5*

Correlation coefficient between the prices of commodities in the two cities is 0·8. [***Nagpur Univ. B.Sc., 1993; Sri Venkateswara Univ. B.Sc. (Oct.) 1990***]

Solution. Let the prices, (in Rupees), in Bombay and Calcutta be denoted by Y and X respectively. Then we are given

$\bar{X} = 65$, $\bar{Y} = 67$, $\sigma_X = 2{\cdot}5$, $\sigma_Y = 3{\cdot}5$ and $r = r(X, Y) = 0{\cdot}8$. We want Y for $X = 70$.

Line of regression of Y on X is

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$\Rightarrow \qquad Y = 67 + 0{\cdot}8 \times \frac{3{\cdot}5}{2{\cdot}5}(X - 65)$$

When $X = 70$, $\hat{Y} = 67 + 0{\cdot}8 \times \frac{3{\cdot}5}{2{\cdot}5}(70 - 65) = 72{\cdot}6$

Example 10·26. *Can $Y = 5 + 2{\cdot}8X$ and $X = 3 - 0{\cdot}5Y$ be the estimated regression equations of Y on X and X on Y respectively ? Explain your answer with suitable theoretical arguments.* [***Delhi Univ. M.A.(Eco.), 1986***]

Solution. Line of regression of Y on X is :

$$Y = 5 + 2{\cdot}8X \quad \Rightarrow \quad b_{YX} = 2{\cdot}8 \qquad \ldots(*)$$

Line of regression of X on Y is :

$$X = 3 - 0{\cdot}5Y \quad \Rightarrow \quad b_{XY} = -0{\cdot}5 \qquad \ldots(**)$$

This is not possible, since each of the regression coefficients b_{YX} and b_{XY} must have the same sign, which is same as that of Cov (X, Y). If Cov (x, y) is positive, then both the regression coefficients are positive and if Cov (X, Y) is negative, then both the regression coefficients are negative. Hence (*) and (**) cannot be the estimated regression equations of Y on X and X on Y respectively.

EXERCISE 10 (d)

1. (*a*) Explain what are regression lines. Why are there two such lines ? Also derive their equations.

(*b*) Define (*i*) Line of regression, (*ii*) Regression coefficient. Find the equations to the lines of regression and show that the coefficient of correlation is the geometric mean of coefficients of regression.

(*c*) What equation is the equivalent mathematical statement for the following words ?

"If the respective deviations in each series, X and Y, from their means were expressed in units of standard deviations, *i.e.*, if each were divided by the

standard deviation of the series; to which it belongs and plotted to a scale of standard deviations, the slope of a straight line best describing the plotted points would be the correlation coefficient r."

2(*a*) Obtain the equation of the line of regression of Y on X and show that the angle θ, between the two lines of regression is given by

$$\tan \theta = \frac{1 - \rho^2}{\rho} \times \frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2}$$

where σ_1, σ_2 are the standard deviations of X and Y respectively, and ρ is the correlation coefficient. ***(Delhi Univ. B.Sc. (Maths. Hons.), 1989)***

Interpret the cases when $\rho = 0$ and $\rho = \pm 1$.

(Bangalore Univ. B.Sc. 1990)

(*b*) If θ is the acute angle between the two regression lines with correlation coefficient r, show that $\sin \theta \leq 1 - r^2$.

3. (*a*) Explain the term "regression" by giving examples. Assuming that the regression of Y on X is linear, outline a method for the estimation of the coefficients in the regression line based on the random paired sample of X and Y, and show that the variance of the error of the estimate for Y for the regression line is $\sigma_Y^2 (1 - \rho^2)$, where σ_Y^2 is the variance of Y and ρ is the correlation coefficient between X and Y.

(*b*) Prove that X and Y are linearly related if and only if $\rho_{XY}^2 = 1$. Further show that the slope of the regression line is positive or negative according as $\rho = +1$ or $\rho = -1$.

(*c*) Let X and Y be two variates. Define $X^* = \frac{X - a}{b}$, $Y^* = \frac{Y - c}{d}$ for some constants a, b, c and d. Show that the regression line (least square) of Y on X can be obtained from that of Y^* on X^*.

(*d*) Show that the coefficient of correlation between the observed and the estimated values of Y obtained from the line of regression of Y on X, is the same as that between X and Y.

4. Two variables X and Y are known to be related to each other by the relation $Y = X/(aX + b)$. How is the theory of linear regression to be employed to estimate the constants a and b from a set of n pairs of observations (x_i, y_i), $i = 1, 2, \ldots, n$?

Hint. $$\frac{1}{Y} = \frac{aX + b}{X} = a + \frac{b}{X}$$

Put $$\frac{1}{X} = U \text{ and } \frac{1}{Y} = V$$

$\therefore$ $$V = a + bU$$

5. Derive the standard error of estimate of Y obtained from the linear regression equation of Y on X. What does this standard error measure ?

6. (*a*) Calculate the coefficient of correlation from the following data :

X :	1	2	3	4	5	6	7	8	9
Y :	9	8	10	12	11	13	14	16	15

Also obtain the equations of the lines of regression and obtain an estimate of Y which should correspond on the average to $X = 6{\cdot}2$.

Ans. $r = 0{\cdot}95$, $Y - 12 = 0{\cdot}95\,(X - 5)$, $X - 5 = 0{\cdot}95\,(Y - 12)$, 13·14

(*b*) Why do we have, in general, two lines of regression ? Obtain the regression of Y on X, and X on Y from the following table and estimate the blood pressure when the age is 45 years :

Age in years (X)	*Blood pressure* (Y)	*Age in years* (X)	*Blood pressure* (Y)
56	147	55	150
42	125	49	145
72	160	38	115
36	118	42	140
63	149	68	152
47	128	60	155

Ans. $Y = 1{\cdot}138X + 80{\cdot}778$, $Y = 131{\cdot}988$ for $X = 45$.

(*c*) Suppose the observations on X and Y are given as :

X :	59	65	45	52	60	62	70	55	45	49
Y :	75	70	55	65	60	69	80	65	59	61

where $N = 10$ students, and Y = Marks in Maths, X = Marks in Economics. Compute the least square regression equations of Y on X and of X on Y.

If a student gets 61 marks in Economics, what would you estimate his marks in Maths to be ?

7. (*a*) In a correlation analysis on the ages of wives and husbands, the following data were obtained. Find

(*i*) the value of the correlation coefficient, and (*ii*) the lines of regression.

Estimate the age of husband whose wife's age is 31 years. Estimate the age of wife whose husband is 40 years old.

Age of wife → / Age of Husband ↓	15—25	25—35	35—45	45—55	55—65
15—30	30	6	3	—	—
30—45	18	32	15	12	8
45—60	2	28	40	16	9
60—75	—	4	9	10	8

(*b*) The following table gives the distribution of total cultivable area (X) and area under cultivation (Y) in a district of 69 villages.

Calculate (*i*) the linear regression of Y on X,

(*ii*) the correlation coefficient $r(X, Y)$, and (*iii*) the average area under wheat corresponding to total area of 1,000 Bighas,

Area under wheat	Total area (in Bighas)				
	0—500	500—1000	1000—1500	1500—2000	2000—2500
0—200	12	6	...	...	...
200—400	2	18	4	2	1
400—600	...	4	7	3	...
600—800	...	1	...	2	1
800—1000	...	...	1	2	3

Ans. (*i*) $Y = 0{\cdot}7641X - 455{\cdot}3854$, (*ii*) $r(X, Y) = 0{\cdot}756$

(*iii*) $Y = 308{\cdot}7146$ for $X = 1000$

8. (*a*) Compare and contrast the roles of correlation and regression in studying the inter-dependence of two variates.

For 10 observations on price (X) and supply (Y) the following data were obtained (in appropriate units).

$\Sigma X = 130$, $\Sigma Y = 220$, $\Sigma X^2 = 2288$, $\Sigma Y^2 = 5506$ and $\Sigma XY = 3467$

Obtain the line of regression of Y on X and estimate the supply when the price is 16 units, and find out the standard error of the estimate.

Ans. $Y = 8{\cdot}8 + 1{\cdot}015X$, 25·04

(*b*) If a number X is chosen at random from among the integers 1, 2, 3, 4 and a number Y is chosen from among those at least as large as X, prove that

$$\text{Cov}\ (X, Y) = \frac{5}{8}$$

Find also the regression line of X on Y.

(*c*) Calculate the correlation coefficient from the following data :—

$N = 100$, $\Sigma X = 12500$ $\Sigma Y = 8000$

$\Sigma X^2 = 1585000$, $\Sigma Y^2 = 648100$ $\Sigma XY = 1007425$.

Also obtain the regression equation of Y on X.

9. (*a*) The means of a bivariate frequency distribution are at (3, 4) and $r = 0{\cdot}4$. The line of regression of Y on X is parallel to the line $Y = X$. Find the two lines of regression and estimate the mean of X when $Y = 1$.

(*b*) For certain data, $Y = 1{\cdot}2X$ and $X = 0{\cdot}6Y$, are the regression lines. Compute $\rho(X, Y)$ and σ_X/σ_Y. Also compute $\rho\ (X, Z)$, if $Z = Y - X$.

(*c*) The equations of two regression lines obtained in a correlation analysis are as follows :

$$3X + 12Y = 19, \quad 3Y + 9X = 46$$

Obtain (*i*) the value of correlation coefficient,

(*ii*) mean values of X and Y, and

(*iii*) the ratio of the coefficient of variability of X to that of Y.

Ans. (*i*) $-\frac{1}{2}\sqrt{3}$, (*ii*) $\bar{X} = 5, \bar{Y} = 1/3$.

(*d*) For an army personnel of strength 25, the regression of weight of kidneys (Y) on weight of heart (X), both measured in ounces is

$$Y - 0{\cdot}399X - 6{\cdot}934 = 0$$

and the regression of weight of heart on weight of kidney is

$$X - 1{\cdot}212Y + 2{\cdot}461 = 0$$

Find the correlation coefficient between X and Y and their mean values. Can you find out the standard deviation of X and Y as well ?

Ans. $r(X, Y) = 0{\cdot}70$, $\bar{X} = 11{\cdot}5086$, $\bar{Y} = 11{\cdot}5261$, No.

(*e*) Find the coefficient of correlation for distribution in which

S.D of $X = 3{\cdot}0$ units

S.D. of $Y = 1{\cdot}4$ units

Coefficient of regression of Y on $X = 0{\cdot}28$.

10. (*a*) Given that $X = 4Y + 5$ and $Y = kX + 4$, are the lines of regression of X on Y and Y on X respectively, show that $0 < 4k < 1$. If $k = \frac{1}{16}$, find the means of the two variables and coefficient of correlation between them.

[*Punjab Univ. B.Sc. (Hons.), 1989*]

Hint. $X = 4Y + 5 \Rightarrow b_{XY} = 4$

$Y = kx + 4 \Rightarrow b_{YX} = k$

$\therefore \quad r^2 = 4k \qquad \ldots(*)$

But $0 \le r^2 \le 1 \Rightarrow 0 \le 4k \le 1$.

If $k = \frac{1}{16}$, then from (*), we get

$r^2 = 4 \times \frac{1}{16} \Rightarrow r = +\frac{1}{2}$ [Since both the regression coefficient are positive]

For $k = \frac{1}{16}$, the two lines of regression become

$$X = 4Y + 5 \text{ and } Y = \frac{1}{16}X + 4$$

Solving the two equations, we get $\bar{Y} = 5{\cdot}75$, $\bar{X} = 28$.

(*b*) For 50 students of a class the regression equation of marks in Statistics (X) on marks in Mathematics (Y) is $3Y - 5X + 180 = 0$. The mean marks in Mathematics is 44 and variance of marks in Statistics is 9/16th of the variance of marks in Mathematics. Find the mean marks in Statistics and the coefficient of correlation between marks in two subjects.

[*Bangalore Univ. B.Sc., 1989*]

Hint. We are given $n = 50$, $\bar{Y} = 44$

and $\sigma_X^2 = \frac{9}{16}\sigma_Y^2 \Rightarrow \frac{\sigma_X}{\sigma_Y} = \frac{3}{4} \qquad \ldots(*)$

The equation of the line of regression of X on Y is given to be

$$3Y - 5X + 180 = 0 \quad \Rightarrow \quad X = \frac{3}{5}Y + \frac{180}{5}$$

$$\therefore \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y} = \frac{3}{5} \quad \Rightarrow \quad r \cdot \frac{3}{4} = \frac{3}{5} \quad \text{or} \quad r = 0{\cdot}8$$

Since the lines of regression pass through the point $(\bar{X}, \bar{Y})$, we get

$$\bar{X} = \frac{3}{5}\bar{Y} + \frac{180}{5} = \frac{3}{5} \times 44 + 36 = 62{\cdot}4$$

(*c*) Out of the two lines of regression given by

$$X + 2Y - 5 = 0 \text{ and } 2X + 3Y - 8 = 0,$$

which one is the regression line of X on Y?

Use the equations to find the mean of X and the mean of Y. If the variance of X is 12, calculate the variance of Y.

Ans. $\bar{X} = 1, \bar{Y} = 2, \sigma_Y^2 = 4$

(*d*) The lines of regression in a bivariate distribution are :

$$X + 9Y = 7 \text{ and } Y + 4X = \frac{49}{3}$$

Find (*i*) the coefficient of correlation, (*iii*) the ratios $\sigma_X^2 : \sigma_Y^2$: Cov (X, Y), (*iii*) the means of the distribution and (*iv*) $E(X \mid Y = 1)$.

(*e*) Estimate X when $Y = 10$, if the two lines of regression are :

$$X = -\frac{1}{18}Y + \lambda \text{ and } Y = -2x + \mu,$$

(λ, μ) being unknown and the mean of the distribution is at $(-1, 2)$. Also compute r, λ and μ. **[*Gujarat Univ. B.Sc., Oct. 1992*]**

11. (*a*) The following results were obtained in the analysis of data on yield of dry bark in ounces (Y) and age in years (X) of 200 cinchona plants :

	X	Y
Average	9·2	16·5
Standard deviation	2·1	4·2

Correlation coefficient = +0·84

Construct the two lines of regression and estimate the yield of dry bark of a plant of age 8 years. **[*Patna Univ. B.Sc., 1991*]**

(*b*) The following data pertain to the marks in subjects A and B in a certain examination :

Mean marks in A = 39·5

Mean marks in B = 47·5

Standard deviation of marks in A = 10·8

Standard deviation of marks in B = 16·8

Coefficient of correlation between marks in A and marks in B = 0·42.

Draw the two lines of regression and explain why there are two regression equations. Give the estimate of marks in B for candidates who secured 50 marks in A.

Ans. $Y = 0{\cdot}65X + 21{\cdot}825, X = 0{\cdot}27Y + 26{\cdot}675$ and $Y = 54{\cdot}342$ for $X = 50$

(*c*) You are given the following information about advertising expenditure and sales :

	Advertising Expenditure (X) (*Rs. lakhs*)	*Sales (Y)* (*Rs. lakhs*)
Mean	10	90
s.d.	3	12

Correlation coefficient = 0·8

What should be the advertising budget if the company wants to attain sales target of Rs. 120 lakhs ? [*Delhi Univ. M.C.A., 1990*]

12. Twenty-five pairs of value of variates X and Y led to the following results :

$N = 25$, $\Sigma X = 127$, $\Sigma Y = 100$, $\Sigma X^2 = 760$, $\Sigma Y^2 = 449$ and $\Sigma XY = 500$

A subsequent scrutiny showed that two pairs of values were copied down as :

X	Y
8	14
8	6

X	Y
8	12
6	8

(*i*) Obtain the correct value of the correlation coefficient.

(*ii*) Hence or otherwise, find the correct equations of the two lines of regression.

(*iii*) Find the angle between the regression lines.

Ans. (*i*) $r(X, Y) = -(0{\cdot}64 \times 0{\cdot}15)^{1/2}$,

(*ii*) $X = -0{\cdot}64Y + 7{\cdot}56$, $Y = -0{\cdot}15X + 4{\cdot}75$.

13. Suppose you have n observations :

$$(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$$

on two variables X and Y, and you have fitted a linear regression $Y = a + bX$ by the method of least squares. Denote the 'expected' value of Y by Y^*, and the residual $Y - Y^*$ by e. Find means and variances of Y^* and e, and the correlation co-efficient between (*i*) X and e, (*ii*) Y and e and (*iii*) Y and Y^*. Use these results to bring out the significance and limitations of the correlation coefficient.

Ans. $r(X, e) = 0$, $r(Y, e) = 0$ and $r(Y, Y^*) = r(X, Y)$.

14. (*a*) The regression lines of Y on X and of X on Y are respectively $Y = aX + b$ and $X = cY + d$. Show that

(*i*) Means are $\bar{X} = (bc + d)/(1 - ac)$ and $\bar{Y} = (ad + b)/(1 - ac)$

(*ii*) Correlation coefficient between X and Y is $\sqrt{a\,c}$

(*iii*) The ratio of the standard deviations of X and Y is $\sqrt{c/a}$.

(*b*) For two random variables X and Y with the same mean, the two regression equations are $Y = aX + b$ and $X = \alpha Y + \beta$. Show that $\frac{b}{\beta} = \frac{1 - a}{1 - \alpha}$. Find also the common mean.

[*Punjab Univ.B.Sc. (Maths Hons.), 1992*]

(*c*) If the lines of regression of Y on X and X on Y are respectively $a_1X + b_1Y + c_1 = 0$ and $a_2X + b_2Y + c_2 = 0$, prove that $a_1b_2 \leq a_2b_1$.

(Delhi Univ. B.Sc. (Stat. Hons.), 1989)

Hint. $r^2 = b_{YX} \cdot b_{XY} \leq 1 \Rightarrow \left(-\frac{a_1}{b_1}\right) \times \left(-\frac{b_2}{a_2}\right) = \frac{a_1b_2}{a_2b_1} \leq 1$

15. (*a*) By minimising $\sum_{i=1}^{n} f_i (x_i \cos \alpha + y_i \sin \alpha - p)^2$ for variations in α and p, show that there are two straight lines passing through the mean of the distribution for which the sum of squares of normal deviations has an extreme value. Prove also that their slopes are given by

$$\tan 2\alpha = \frac{2\mu_{11}}{\sigma_X^2 - \sigma_Y^2}$$

Hint. We have to minimize

$$S = \sum_{i=1}^{n} f_i (x_i \cos \alpha + y_i \sin \alpha - p)^2 \quad \ldots(1)$$

Equating to zero, the partial derivatives of (1) w.r.t. α and p, we have

$$\frac{\partial S}{\partial \alpha} = 0 = 2 \sum_{i=1}^{n} f_i (x_i \cos \alpha + y_i \sin \alpha - p)(-x_i \sin \alpha + y_i \cos \alpha) \quad \ldots(2)$$

$$\frac{\partial S}{\partial p} = 0 = -2 \sum_{i=1}^{n} f_i (x_i \cos \alpha + y_i \sin \alpha - p) \quad \ldots(3)$$

Equation (3) can be written as

$$\sum_{i=1}^{n} f_i (x_i \cos \alpha + y_i \sin \alpha - p) = 0 \Rightarrow \bar{x} \cos \alpha + \bar{y} \sin \alpha - p = 0 \quad \ldots(4)$$

From equation (2), we get a quadratic equation which shows that there are two straight lines for extreme values of E.

From equation (4), it becomes clear that both the straight lines pass through the point $(\bar{x}, \bar{y})$.

Again equation (2) can be written as :

$$\sum_{i=1}^{n} f_i (x_i \cos \alpha + y_i \sin \alpha - p)(y_i \cos \alpha - x_i \sin \alpha) = 0$$

$$\Rightarrow \sum_{i=1}^{n} f_i [\cos \alpha (x_i - \bar{x}) + \sin \alpha (y_i - \bar{y})][y_i \cos \alpha - x_i \sin \alpha] = 0$$

[Using (4)]

$$\Rightarrow \cos^2 \alpha \sum_{i=1}^{n} f_i y_i (x_i - \bar{x}) - \sin \alpha \cos \alpha \sum_{i=1}^{n} f_i x_i (x_i - \bar{x})$$

$$+ \sin \alpha \cos \alpha \sum_{i=1}^{n} f_i y_i (y_i - \bar{y}) - \sin^2 \alpha \sum_{i=1}^{n} f_i x_i (y_i - \bar{y}) = 0 \quad \ldots(5)$$

We have $\mu_{11} = \frac{1}{N} \sum_i f_i (x_i - \bar{x})(y_i - \bar{y})$

$$= \frac{1}{N}\sum_i f_i x_i (y_i - \bar{y}) - \bar{x}.\frac{1}{N}\sum_i f_i (y_i - \bar{y}) = \frac{1}{N}\sum f_i x_i (y_i - \bar{y})$$

Similarly, $\mu_{11} = \frac{1}{N}\sum_i f_i y_i (x_i - \bar{x})$

$$\sigma_X^2 = \frac{1}{N}\sum_i f_i x_i (x_i - \bar{x}) \text{ and } \sigma_Y^2 = \frac{1}{N}\sum_i f_i y_i (y_i - \bar{y})$$

Substituting these values in (5), we get the required result.

(*b*) If the straight line defined by

$$Y = a + bX$$

satisfies the condition $\Sigma[(Y - a - bX)^2]$ = minimum, show that the regression line of the random variable Y on the random variable X is

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}), \text{ where } \bar{X} = E(X), \bar{Y} = E(Y)$$

16. (*a*) Define Curve of regression of Y on X.

The joint density function of X and Y is given by :

$$f(x, y) = x + y, \ 0 < x < 1, 0 < y < 1$$
$$= 0, \text{ otherwise}$$

Find

(*i*) the correlation coefficient between X and Y,

(*ii*) the regression curve of Y on X, and

(*iii*) the regression curve of X on Y.

Ans. $\rho(X, Y) = -\frac{1}{11}$. [***Madras Univ. B.Sc., Stat. (Main), 1992***]

(*b*) Let $f(x_1, x_2) = \frac{2}{a^2}; 0 < x_1 < x_2, 0 < x_2 < a$

$$= 0, \text{ elsewhere}$$

be the joint p.d.f. of X_1 and X_2.

Find conditional means and variances. Also show that $\rho = \frac{1}{2}$.

17. If the joint density of X and Y is given by

$$f(x, y) = \begin{cases} (x + y)/3, \text{ for } 0 < x < 1, \ 0 < y < 2 \\ 0, \text{ otherwise} \end{cases}$$

obtain the regressions (*i*) of Y on X and (*ii*) of X on Y.

Are the regressions linear ? Find the correlation coefficient between X and Y. (***Allahabad Univ. B.Sc. 1992***)

Ans. $y = E(Y|x) = \frac{3x + 4}{3(x + 1)}; x = E(X|y) = \frac{2 + 3y}{3(1 + 2y)}$

Corr. $(X, Y) = -\left(\frac{2}{299}\right)^{1/2}$

18. Let the joint density function of X and Y be given by :

$$f(x, y) = 8xy, \ 0 < x < y < 1$$
$$= 0, \text{ otherwise}$$

Find: (*i*) $E(Y \mid X = x)$, (*ii*) $E[XY \mid X = x]$, (*iii*) $\text{Var}\,[Y \mid X = x]$

[*Delhi Univ. BSc. (Maths Hons.), 1988*]

Ans. (*i*) $E(Y|x) = \frac{2}{3}\left(\frac{1+x+x^2}{1+x}\right)$ $E(XY \mid x) = x.E(Y \mid x)$, (*iii*) $E(Y^2 \mid x) = \frac{1+x^2}{2}$

19. Give an example to show that it is possible to have the regression of Y on X constant (does not depend on X), but the regression of X on Y is not constant (does depend on Y).

Hint. See Example 10·21

20. Prove or disprove :

$$E(Y \mid X = x) = \text{constant} \implies r(X,Y) = 0$$

Ans. True

21. If $f(x,y) = \frac{1}{3}x^2 \exp[-y(1+x)]$, $x \geq 0, y \geq 0$ is the joint p.d.f. of (X,Y), obtain the equation of regression of Y on X.

Ans. $y = E(Y \mid x) = 1/(1+x)$.

22. Variables (X,Y) have joint p.d.f.

$f(x,y) = 6(1-x-y)$, $x > 0, y > 0, x + y < 1$.

$= 0$, otherwise.

Find $f_X(x)$, $f_Y(y)$ and Cov (X,Y). Are X and Y independent ? Obtain the regression curves for the means.

[*Calcutta Univ. B.Sc. (Maths Hons.), 1986*]

Ans. $f_1(x) = 3(1-x)^2$, $0 < x < 1$; $f_2(y) = 3(1-y)^2$, $0 < y < 1$.

X and Y are not independent.

Regression curves for the means are :

$$y = E(Y \mid x) = \frac{1}{3}(1-x) \text{ and } x = E(X \mid y) = \frac{1}{3}(1-y).$$

23. For the joint p.d.f.

$$f(x, y) = 3x^2 - 8xy + 6y^2; 0 \leq (x, y) \leq 1,$$

find the least square regression lines and the regression curves for the means.

[*Calcutta Univ. B.Sc. (Maths, Hons.), 1987*]

Ans. Regression lines :

$$y - \frac{2}{3} = -\frac{10}{67}\left(x - \frac{5}{12}\right) \; ; \quad x - \frac{5}{12} = -\frac{25}{32}\left(y - \frac{2}{3}\right)$$

Regression curves for means are :

$$y = E(Y|x) = \frac{9x^2 - 16x + 9}{6(3x^2 - 4x + 2)} \; ; \quad x = E(X \mid y) = \frac{36y^2 - 32y - 9}{12(6y^2 - 4y + 1)}$$

24. Let (X, Y) be jointly distributed with p.d.f.

$f(x, y) = e^{-y}$, $0 < x < y < \infty$

$= 0$, otherwise

Prove that :

$E(Y \mid X = x) = x + 1$ and $E(X \mid Y = y) = y/2$.

Hence prove that $r(X, Y) = \sqrt{1/2}$.

25. Let $f(x, y) = e^{-y}(1 - e^{-x})$, $0 < x < y$; $0 < y < \infty$
$= e^{-x}(1 - e^{-y})$, $0 < y < x$; $0 < x < \infty$

(*a*) Show that $f(x, y)$ is a p.d.f.

(*b*) Find marginal distributions of X and Y.

(*c*) Find $E(Y \mid X = x)$ for $x > 0$.

(*d*) Find $P(X \leq 2, Y \leq 2)$.

(*e*) Find the correlation coefficient $r(X, Y)$.

(*f*) Find another joint p.d.f. having the same marginals.

Ans. (*b*) $f_1(x) = xe^{-x}$, $0 < x < \infty$; $f_2(y) = ye^{-y}$, $0 < y < \infty$.

(*c*) $E(Y \mid x) = \frac{1 - e^x}{x}[x - 1] + \frac{1}{x}\left(\frac{x^2}{2} + xe^x + e^{-x} - 1\right)$

(*d*) $1 - \frac{1}{e^4} - \frac{4}{e^2}$; (*e*) $r(x, y) = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} = \frac{1}{\sqrt{2}\sqrt{2}} = \frac{1}{2}$

(*f*) **Hint.** $f(x, y, \alpha) = f_1(x) f_2(y)[1 + \alpha(2F(x) - 1)(2F(y) - 1)]$, $|\alpha| < 1$, has the same marginals $f_1(x)$ and $f_2(y)$.

26. Obtain regression equation of Y on X for the distributions :

(*a*) $f(x, y) = \frac{9}{2} \cdot \frac{1 + x + y}{(1 + x)^4 (1 + y)^4}$; $x, y \geq 0$

(*b*) $f(x, y) = \frac{4}{5}(x + 3y)e^{-x-2y}$; $x, y \geq 0$

[*Sardar Patel Univ. M.Sc., 1992*]

Ans. (*a*) **Hint.** See Example 5·25, page 5·55, (*b*) $\frac{x + 3}{2x + 3}$.

27. A ball is drawn at random from an urn containing three white balls numbered 0, 1, 2 ; two red balls numbered 0, 1 and one black ball numbered 0. If the colours white, red and black are again numbered 0, 1 and 2 respectively, find the correlation coefficient between the variates X, the colour number and Y the number of the ball. Write down the equation of regression line of Y on X.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1986*]

OBJECTIVE TYPE QUESTIONS

I. State, giving reasons, whether each of the following statements is true or false.

(*i*) Both regression lines of Y on X and of X on Y do not intersect at all.

(*ii*) In a bivariate regression, $b_{YX} = \frac{1}{5}$, $b_{XY} = 10$

(*iii*) The regression coefficient of Y on X is 3·2 and that of X on Y is 0·8.

(*iv*) There is no relationship between correlation coefficient and regression coefficient.

(*v*) Both the regression coefficients cannot exceed unity.

(*vi*) The greater the value of '*r*', the better are the estimates obtained through regression analysis.

(*vii*) If X and Y are negatively correlated variables, and (0, 0) is on the least squares line of Y on X, and if $X = 1$ is the observed value then predicted value of Y must be negative.

(*viii*) Let the correlation between X and Y be perfect and positive. Suppose the points (3, 5) and (1, 4) are on the regression lines. With this knowledge it is possible to determine the least squares line exactly.

(*ix*) If the lines of regression are $Y = \frac{1}{4}X$ and $X = \frac{1}{9}Y + 1$, then $\rho = \frac{1}{6}$ and $E(X \mid Y = 0) = 1$.

(*x*) In a bivariate distribution, $b_{YX} = 2{\cdot}8$ and $b_{XY} = 0{\cdot}3$.

II. Fill in the blanks :

(*i*) The regression analysis measures ... between X and Y.

(*ii*) Lines of regression are ... if $r_{XY} = 0$ and they are ... if $r_{XY} = \pm 1$.

(*iii*) If the regression coefficients of X on Y and Y on X are $-0{\cdot}4$ and $-0{\cdot}9$ respectively then the correlation coefficient between X and Y is ...

(*iv*) If the two regression lines are $X + 3Y - 5 = 0$ and $4X + 3Y - 8 = 0$, then the correlation coefficient between X and Y is ...

(*v*) If one of the regression coefficients is ... unity, the other must be ... unity.

(*vi*) The farther the two regression lines cut each other, the ... will be the degree of correlation.

(*vii*) When one regression coefficient is positive, the other would be ...

(*viii*) The sign of regression coefficient is ... as that of correlation coefficient.

(*ix*) Correlation coefficient is the ... between regression coefficients.

(*x*) Arithmetic mean of regression coefficients is ... correlation coefficient.

(*xi*) When the correlation coefficient is zero, the two regression lines are ... and when it is ± 1, then the regression lines are ...

III. Indicate the correct answer :

(*i*) The regression line of Y on X (*a*) minimises total of the squares of horizontal deviations, (*b*) total of the squares of the vertical deviations, (*c*) both vertical and horizontal deviations, (*d*) none of these.

(*ii*) The regression coefficients are b_2 and b_1. Then the correlation coefficient r is (*a*) b_1/b_2, (*b*) b_2/b_1, (*c*) $b_1 b_2$ (*d*) $\pm \sqrt{b_1 b_2}$.

(*iii*) The farther the two regression lines cut each other (*a*) the greater will be the degree of correlation, (*b*) the lesser will be the degree of correlation, (*c*) does not matter.

(*iv*) If one regression coefficient is greater than unity, then the other must be (*a*) greater than the first one, (*b*) equal to unity, (*c*) less than unity, (*d*) equal to zero.

(*v*) When the correlation coefficient $r = \pm 1$, then the two regression lines (*a*) are perpendicular to each other, (*b*) coincide, (*c*) are parallel to each other, (*d*) do not exist.

(*vi*) The two lines of regression are given as $X + 2Y - 5 = 0$ and $2X + 3Y = 8$. Then the mean values of X and Y respectively are (*a*) 2, 1, (*b*) 1, 2, (*c*) 2, 5, (*d*) 2, 3.

(*vii*) The tangent of the angle between two regression lines is given as 0·6 and the s.d. of Y is known to be twice that of X. Then the value of correlation coefficient between X and Y is (*a*) $-\frac{1}{2}$, (*b*) $\frac{1}{2}$, (*c*) 0·7, (*d*) 0·3.

IV. σ_X and σ_Y are the standard deviations of two correlated variables X an Y respectively in a large sample, and r is the sample correlation coefficient.

(*i*) State the "Standard Error of Estimate" for linear regression of Y on X.

(*ii*) What is the standard error in estimating Y from X if $r = 0$?

(*iii*) By how much is this error reduced if r is increased to 0·30 ?

(*iv*) How large must r be in order to reduce this standard error to one-half its value for $r = 0$?

(*v*) Give your interpretations for the cases $r = 0$ and $r = 1$.

V. Explain why we have two lines of regression.

10·8. Correlation Ratio. As discused earlier, when variables are linearly related, we have the regression lines of one variable on another variable and correlation coefficient can be computed to tell us about the extent of association between them. However, if the variables are not linearly related but some sort of curvilinear relationship exists between them, the use of r which is a measure of the degree to which the relation approaches a straight line "law" will be misleading. We might come across bivariate distributions where r may be very low or even zero but the regression may be strong, or even perfect. *Correlation ratio* 'η' is the appropriate measure of curvilinear relationship between the two variables. Just as r measures the concentration of points about the straight line of best fit, η measures the concentration of points about the curve of best fit. If regression is linear $\eta = r$, otherwise $\eta > r$ (*c.f.* Remark 2, § 10·8·1).

10·8·1. Measure of Correlation Ratio. In the previous articles we have assumed that there is a single observed value Y corresponding to the given value x_i of X but sometimes there are more than one such value of Y.

Suppose corresponding to the values x_i, $(i = 1, 2, \ldots, m)$ of the variable X, the variable Y takes the values y_{ij} with respective frequencies f_{ij}, $j = 1, 2, \ldots, n$.

Though all the x's in the ith vertical array have the same value, the y's are different. A typical pair of values in the ith array is (x_i, y_{ij}), with frequency f_{ij}.

Thus the first suffix i indicates the vertical array while the second suffix j indicates the positions of y in that array. Let

$$\sum_{j=1}^{n} f_{ij} = n_i \quad \text{and} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} f_{ij}\right) = \sum_{i=1}^{m} n_i = N, \text{ (say).}$$

If $\bar{y}_i$ and $\bar{y}$ denote the means of the ith array and the overall mean respectively, then

$$\bar{y}_i = \frac{\sum_j f_{ij}\, y_{ij}}{\sum_j f_{ij}} = \frac{\sum_j f_{ij}\, y_{ij}}{n_i} = \frac{T_i}{n_i} \quad \text{and} \quad \bar{y} = \frac{\sum_i\sum_j f_{ij}\, y_{ij}}{\sum_i\sum_j f_{ij}} = \frac{\sum_i n_i\, \bar{y}_i}{\sum_i n_i} = \frac{T}{N}$$

In other words $\bar{y}$ is the weighted mean of all the array means, the weights being the array frequencies.

Def. *The correlation ratio of Y on X, usually denoted by* η_{YX} *is given by*

$$\eta_{YX}^2 = 1 - \frac{\sigma_{eY}^2}{\sigma_Y^2} \qquad \ldots(10\cdot21)$$

where σ_{eY}^2 and σ_Y^2 are given by

$$\sigma_{eY}^2 = \frac{1}{N}\sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)^2 \quad \text{and} \quad \sigma_Y^2 = \frac{1}{N}\sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y})^2$$

A convenient expression for η_{YX} can be obtained in terms of standard deviation σ_{mY} of the means of the vertical arrays, each mean being weighted by the array frequency.

We have

$$N\sigma_Y^2 = \sum\sum f_{ij}\,(y_{ij} - \bar{y})^2 = \sum_i\sum_j f_{ij}\,\{(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})\}^2$$

$$= \sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)^2 + \sum_i\sum_j f_{ij}\,(\bar{y}_i - \bar{y})^2 + 2\sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y})$$

The term $2[\sum_i(\bar{y}_i - \bar{y})\{\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)\}]$ vanishes since $\sum_j f_{ij}\,(y_{ij} - \bar{y}_i) = 0$, being the algebraic sum of the deviations from mean.

$$\therefore \quad N\sigma_Y^2 = \sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)^2 + \sum_i n_i(\bar{y}_i - \bar{y})^2$$

$$\Rightarrow \quad N\sigma_Y^2 = N\sigma_{eY}^2 + N\sigma_{mY}^2 \Rightarrow \sigma_Y^2 = \sigma_{eY}^2 + \sigma_{mY}^2$$

$$\Rightarrow \quad 1 - \frac{\sigma_{eY}^2}{\sigma_Y^2} = \frac{\sigma_{mY}^2}{\sigma_Y^2}$$

which on comparison with (10·21) gives

$$\eta_{YX}^2 = \frac{\sigma_{mY}^2}{\sigma_Y^2} = \frac{\sum_i n_i(\bar{y}_i - \bar{y})^2}{\sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y})^2} \qquad \ldots(10\cdot22)$$

We have

$$N\sigma_{m_Y}^2 = \sum_i n_i (\bar{y}_i - \bar{y})^2 = \sum_i n_i \bar{y}_i^2 - N\bar{y}^2 = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

$$\therefore \qquad \eta_{YX}^2 = \left[\sum_i \left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right] \Big/ N\sigma_Y^2, \qquad \ldots(10{\cdot}23)$$

a formula, much more convenient for computational purposes.

Remarks 1. (10·21) implies that

$$\sigma_{e_Y}^2 = \sigma_Y^2 (1 - \eta_{YX}^2)$$

Since $\sigma_{e_Y}^2$ and σ_Y^2 are non-negative, we have

$$1 - \eta_{YX}^2 \geq 0 \quad \Rightarrow \quad \eta_{YX}^2 \leq 1 \quad \Rightarrow \quad |\eta_{YX}| \leq 1$$

2. Since the sum of squares of deviations in any array is minimum when measured from its mean, we have

$$\sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)^2 \leq \sum_i \sum_j f_{ij} (y_{ij} - \hat{y}_{ij})^2 \qquad \ldots(*)$$

where $\hat{y}_{ij}$ is the estimate of y_{ij} for given value of $X = x_i$, say, as given by the line of regression of Y on X *i.e.*, $\hat{y}_{ij} = a + bx_i$, $(j = 1, 2, \ldots, n)$.

But $$\sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)^2 = N\sigma_{e_Y}^2 = N\sigma_Y^2 (1 - \eta_{YX}^2)$$

and $$\sum_i \sum_j f_{ij} (y_{ij} - a - bx_i)^2 = N\sigma_Y^2 (1 - r^2) \qquad (c.f.\ \S\ 10{\cdot}7{\cdot}6)$$

$$\therefore (*) \Rightarrow \qquad 1 - \eta_{YX}^2 \leq 1 - r^2$$

i.e., $$\eta_{YX}^2 \geq r^2 \quad \Rightarrow \quad |\eta_{YX}| \geq |r|$$

Thus the absolute value of the correlation ratio can never be less than the absolute of r, the correlation coefficient.

When the regression of Y on X is linear, straight line of means of arrays coincides with the line of regression and $\eta_{YX}^2 = r^2$. Thus $\eta_{YX}^2 - r^2$ is the departure of regression from linearity. It is also clear (from Remark 1) that the more nearly η_{YX}^2 approaches unity, the smaller is $\sigma_{e_Y}^2$ and, therefore, closer are the points to the curve of means of vertical arrays.

When $\eta_{YX}^2 = 1, \sigma_{e_Y}^2 = 0 \Rightarrow \sum\sum f_{ij} (y_{ij} - \bar{y}_i)^2 = 0$

$\Rightarrow y_{ij} = \bar{y}_i$, $\forall\ j = 1, 2, \ldots, n$, *i.e.*, all the points lie on the curve of means. This implies that there is a functional relationship between X and Y. η_{YX} is, therefore, the measure of the degree to which the association between the variables approaches a functional relationship of the form $Y = F(X)$, where $F(X)$ is a single valued function of X, $[F(X) = a + bX]$.

3. It is worth noting that the value of η_{YX} is not independent of the classification of the data. As the class intervals become narrower η_{YX} approaches unity, since in that case $\sigma_{m_Y}^2$ gets nearer to σ_Y^2. If the grouping is so fine that only one item appears in each row (related to each x-class), that item will constitute the mean of that column and thus in this case $\sigma_{m_Y}^2$ and σ_Y^2 become equal so that $\eta_{YX}^2 = 1$. On the other hand, a very coarse grouping tends to make the value of η_{YX} approach r. "Student" has given a formula for 'the correction'

to be made in the correlation ratio 'for grouping' in Biometrika (Vol IX page 316-320.)

4. It can be easily proved that η_{YX}^2 is independent of change of origin and scale of measurements.

5. η_{XY}^2, the second correlation ratio of X on Y depends upon the scatter of observations about the line of column means.

6. r_{XY} and r_{YX} are same but η_{YX} is, in general, different from η_{XY}.

7. In terms of expectation, correlation ratio is defined as follows :

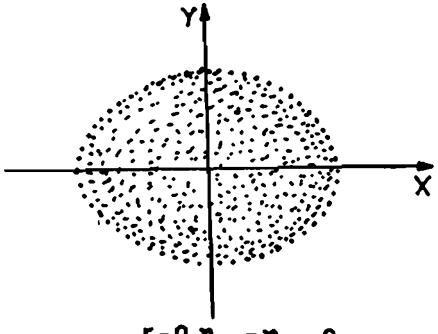
$$\eta_{YX}^2 = \frac{E_X\,[E(Y|X) - E(Y)]^2}{E[Y - E(Y)]^2} = \frac{E[E(Y|X) - E(Y)]^2}{\sigma_Y^2}$$

and

$$\eta_{XY}^2 = \frac{E_Y\,[E\,(X|Y) - E(X)]^2}{E[X - E(X)]^2} = \frac{E[E(X|Y) - E(X)]^2}{\sigma_X^2}$$

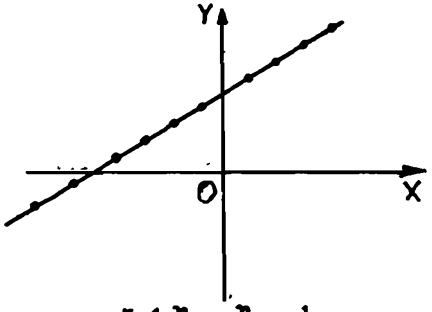
8. We give below some diagrams, exhibiting the relationship between r and η_{YX}.

(i) For completely random scattering of the dots with no trend, both r and η are zero.



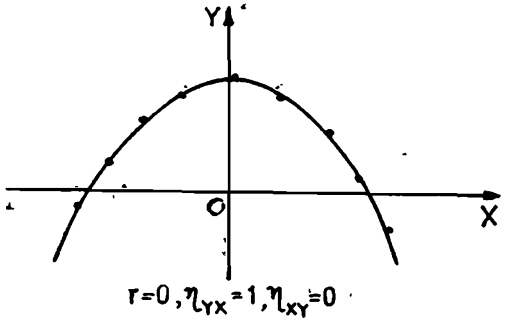
$r = 0, \eta_{YX} = \eta_{XY} = 0$

(ii) If dots lie precisely on a line, $r = 1$ and $\eta = 1$.

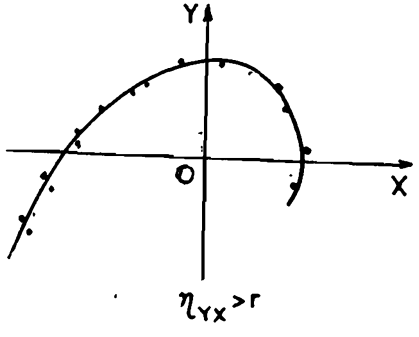


$r = 1, \eta_{YX} = \eta_{XY} = 1$

(iii) If dots lie on a curve, such that no ordinate cuts it more than once, $\eta_{YX} = 1$ and if furthermore, the dots are symmetrically placed about Y-axis, then $\eta_{XY} = 0$, $r = 0$.



(iv) If $\eta_{YX} > r$, the dots are scattered around a definitely curved trend line.



EXERCISE 10(e)

1. (*a*) Define correlation coefficient and correlation ratio. When is the latter a more suitable measure of correlation than the former ? Show that the correlation ratio is never less than the correlation coefficient. What do you infer if the two are equal ? Further, show that none of these can exceed one.

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

(*b*) Show that $1 \geq \eta^2_{YX} \geq r^2_{YX} \geq 0$

Interpret each of the following statements.

(*i*) $r = 0$, (*ii*) $r^2 = 1$, (*iii*) $\eta^2 = 1$, (*iv*) $\eta^2 = r^2$ and (*v*) $\eta = 0$

(*c*) When the correlation coefficient is equal to unity, show that the two correlation ratios are also equal to unity. Is the converse true ?

(*d*) Define correlation ratio η_{XY} and prove that

$$1 \geq \eta^2_{XY} \geq r^2,$$

where r is the coefficient of correlation between X and Y. Show further that $(\eta^2_{XY} - r^2)$ is a measure of non-linearity of regression.

2. For the joint p.d.f.

$$f(x, y) = \tfrac{1}{2}x^3 \exp[-x(y+1)], \; y > 0, x > 0$$
$$= 0 \quad , \text{ otherwise,}$$

find :

(*i*) Two lines of regression.

(*ii*) The regression curves for the means.

(*iii*) $r(X, Y)$.

(*iv*) η^2_{YX} and η^2_{XY}.

[*Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1987*]

Ans. (*i*) $y = -\frac{1}{6}x + 1 \quad ; \quad x = -\frac{2}{3}y + \frac{10}{3}$

(*ii*) $y = E(Y|x) = \frac{1}{x} \quad ; \quad x = E(X|y) = \frac{4}{1+y}$

(*iii*) $r(X, Y) = -\frac{1}{3}$. (*iv*) $\eta^2_{YX} = \frac{1}{3}, \eta^2_{XY} = \frac{1}{5}$

3. Compute $r(X, Y)$ and η_{YX} for the following data :

X :	0·5 — 1·5	1·5 — 2·5	2·5 — 3·5	3·5 — 4·5	4·5 — 5·5
f :	20	30	35	25	15
$\bar{y}_i$:	11·3	12·7	14·7	16·5	19·1

Var $(Y) = 9·61$

Ans. $\eta_{YX} = 0·77$, $r = 0·85$

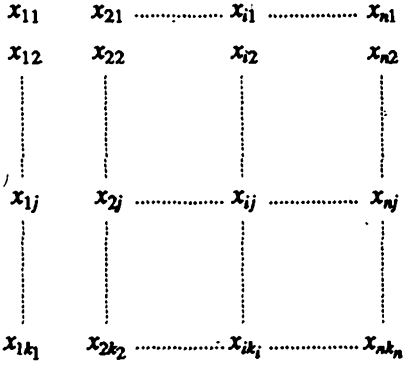
4. Compute η_{XY} for the following table :

$Y \downarrow$ \ $X \rightarrow$	47	52	57	62	67
57	4	4	2	...	...
62	4	8	8	1	...
67	...	7	12	1	4
72	...	3	1	8	5
77	...	...	3	5	6

10·9. Intra-class Correlation. Intra-class correlation means within class correlation. It is distinguishable from product moment correlation in as much as here both the variables measure the same characteristics. Sometimes specially in biological and agricultural study, it is of interest to know how the members of a family or group are correlated among themselves with respect to some one of their common characteristic. For example, we may require the correlation between the heights of brothers of a family or between yields of plots of an experimental block. In such cases both the variables measure the same characteristic, *e.g.*, height and height or weight and weight. There is

nothing to distinguish one from the other so that one may be treated as X-variable and the other as the Y-variable.

Suppose we have $A_1, A_2, \ldots, A_n$ families with $k_1, k_2, \ldots, k_n$ members, each of which may be represented as



and let x_{ij} $(i = 1, 2, \ldots, n; j = 1, 2, \ldots, k_i)$ denote the measurement on the jth member in the ith family.

We shall have $k_i(k_i - 1)$ pairs for the ith family or group like (x_{ij}, x_{il}), $j \neq l$. There will be $\sum_{i=1}^{n} k_i (k_i - 1) = N$ pairs for all the n families or groups. If we prepare a correlation table there will be $k_i (k_i - 1)$ entries for the ith group or family and $\sum_i k_i (k_i - 1) = N$ entries for all the n families or groups. The table is symmetrical about the principal diagonal. Such a table is called an *intra-class correlation table* and the correlation is called *intra-class correlation.*

In the bivariate table x_{i1} occurs $(k_i - 1)$ times, x_{i2} occurs $(k_i - 1)$ times, ..., x_{ik_i} occurs $(k_i - 1)$ times, *i.e.*, from the ith family we have $(k_i - 1) \sum_j x_{ij}$ and hence for all the n families we have $\sum_i (k_i - 1) \sum x_{ij}$ as the marginal frequency, the table being symmetrical about principal diagonal.

$$\therefore \qquad \bar{x} = \bar{y} = \frac{1}{N}\left[\sum_i (k_i - 1) \sum_j x_{ij}\right]$$

Similarly,

$$\sigma_X^2 = \sigma_Y^2 = \frac{1}{N}\left[\sum_i (k_i - 1) \sum_j (x_{ij} - \bar{x})^2\right]$$

Further

$$\text{Cov}\,(X, Y) = \frac{1}{N} \sum_i \left[\sum_{j,\, l} (x_{ij} - \bar{x})(x_{il} - \bar{x})\right], \; j \neq l$$

$$= \frac{1}{N} \sum_i \left[\sum_{j=1}^{k_i} \sum_{l=1}^{k_i} (x_{ij} - \bar{x})(x_{il} - \bar{x}) - \sum_{j=1}^{k_i} (x_{ij} - \bar{x})^2\right]$$

If we write $\bar{x}_i = \sum_j x_{ij} / k_i$, then

$$\sum_i \left[\sum_{j=1}^{k} \sum_{l=1}^{k} (x_{ij} - \bar{x})(x_{il} - \bar{x}) \right] = \sum_i \left[\sum_j (x_{ij} - \bar{x}) \sum_l (x_{il} - \bar{x}) \right]$$

$$= \sum_i [k_i (\bar{x}_i - \bar{x}) k_i (\bar{x}_i - \bar{x})]$$

$$= \sum_i k_i^2 (\bar{x}_i - \bar{x})^2$$

Therefore intra-class correlation coefficient is given by

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)\, V(Y)}} = \frac{\sum_i k_i^2 (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{\sum_i \sum_j (k_i - 1)(x_{ij} - \bar{x})^2} \quad \ldots(10\cdot24)$$

If we put $k_i = k$, *i.e.*, if all families have equal members then

$$r = \frac{k^2 \sum_i (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{(k-1) \sum_i \sum_j (x_{ij} - \bar{x})^2}$$

$$= \frac{nk^2 \sigma_m^2 - nk\sigma^2}{(k-1)\, nk\sigma^2} = \frac{1}{(k-1)} \left\{ \frac{k \sigma_m^2}{\sigma^2} - 1 \right\} \quad \ldots(10\cdot24a)$$

where σ^2 denotes the variance of X and σ_m^2 the variance of means of families.

Limits. We have from (10·24*a*),

$$1 + (k-1) r = \frac{k\sigma_m^2}{\sigma^2} \geq 0 \quad \Rightarrow \quad r \geq -\frac{1}{(k-1)}$$

Also $1 + (k-1) r \leq k$, as the ratio $\frac{\sigma_m^2}{\sigma^2} \leq 1 \quad \Rightarrow \quad r \leq 1$

so that $$-\frac{1}{(k-1)} \leq r \leq 1$$

Interpretation. Intraclass correlation cannot be less than $-1/(k-1)$, though it may attain the value $+1$ on the positive side, so that it is a skew coefficient and a negative value has not the same significance as a departure from independence as an equivalent positive value.

EXERCISE 10 (f)

1. If $x_1, x_2, \ldots, x_k$ be k variates with standard deviation σ and m be any number, prove that

$$k^2\sigma^2 = (k-1) \sum_{r=1}^{k} (x_r - m)^2 - \sum_{r=1}^{k} \sum_{s=1}^{k} (x_r - m)(x_s - m),\ r \neq s$$

Hence deduce that the coefficient of intraclass correlation for n families with varying number of members in each family is

$$1 - \frac{\sum_i k_i \sigma_i^2}{\sigma^2 \sum_i k_i(k_i - 1)}$$

where k_i, σ_i^2 denote the number of members and the variance respectively in the $i$tn family and σ^2 is the general variance.

Given $n = 5$, $\sigma_i = i$, $k_i = i + 1$ $(i \leq 5)$, find the least possible intraclass correlation coefficient.

2. What do you understand by intra-class correlation coefficient.

Calculate its value for the following data :

Family No.	*Height of brothers*			
1	60	62	63	65
2	59	60	61	62
3	62	62	64	63
4	65	66	65	66
5	66	67	67	69

3. In four families each containing eight persons, the chest measurements of persons are given below. Calculate the intraclass correlation co-efficient.

Family	1	2	3	4	5	6	7	8
I	43	46	48	42	50	45	45	49
II	33	34	37	39	82	35	37	41
III	56	52	50	51	54	52	39	52
IV	34	37	38	40	40	41	44	44

10·10. Bivariate Normal Distribution. The bivariate normal distribution is a generalization of a normal distribution for a single variate. Let X and Y be two normally correlated variables with correlation coefficient ρ and $E(X) = \mu_1$, Var $(X) = \sigma_1^2$; $E(Y) = \mu_2$, Var $(Y) = \sigma_2^2$. In deriving the bivariate normal distribution we make the following three assumptions.

(*i*) *The regression of Y on X is linear.* Since the mean of each array is on the line of regression $Y = \rho(\sigma_2/\sigma_1)X$, the mean or expected value of Y is $\rho(\sigma_2/\sigma_1)X$, for different values of X.

(*ii*) *The arrays are homoscedastic, i.e.,* variance in each array is same. The common variance of estimate of Y in each array is then given by $\sigma_2^2 (1 - \rho^2)$, ρ being the correlation coefficient between variables X and Y and is independent of X.

(*iii*) *The distribution of Y in different arrays in normal.* Suppose that one of the variates, say X, is distributed normally with mean 0 and standard deviation σ_1 so that the probability that a random value of X will fall in the small interval dx is

$$g(x)\,dx = \frac{1}{\sigma_1\sqrt{(2\pi)}} . \exp\ (-x^2/2\sigma_1^2)\,dx$$

The probability that a value of Y, taken at random in an assigned vertical array will fall in the interval dy is

$$h(y \mid x)\, dy = \frac{1}{\sigma_2\sqrt{2\pi\,(1-\rho^2)}} \cdot \exp\left\{-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(y - \rho\, x\, \frac{\sigma_2}{\sigma_1}\right)^2\right\}$$

The joint probability differential of X and Y is given by

$$dP(x, y) = g(x)h(y \mid x)dxdy$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \cdot e^{-\frac{1}{2\sigma_1^2}\cdot x^2} \; e\left[-\frac{1}{2\sigma_2^2(1-\rho^2)}\left(y - \rho\,\frac{\sigma_2}{\sigma_1}\,x\right)^2\right]$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)\right\}$$

Shifting the origin to (μ_1, μ_2), we get

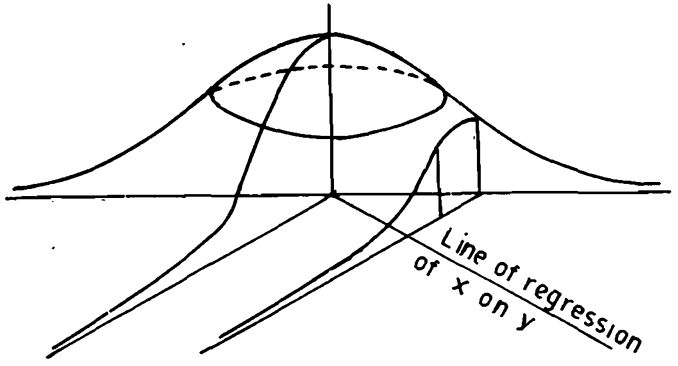
$$f_{XY}(x, y)$$

$$= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\,\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right\}},$$

$$(-\infty < x < \infty, -\infty < y < \infty) \quad \ldots(10\text{-}25)$$

where μ_1, μ_2, σ_1 (>0), σ_2 (>0) and ρ $(-1 < \rho < 1)$ are the five parameters of the distribution.

NORMAL CORRELATION SURFACE



This is the density function of a bivariate normal distribution. The variables X and Y are said to be normally correlated and the surface $z = f(x, y)$ is known as the *normal correlation surface*. The nature of the normal correlation surface is indicated in the above diagram

Remarks 1. The vector $(X, Y)'$ following the joint p.d.f. $f(x, y)$ as given in (10·25), will be abbreviated as $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ or $BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. If in particular $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$ then

$$(X, Y) \sim N(0, 0, 1, 1, \rho) \text{ or } BVN(0, 0, 1, 1, \rho).$$

2. The curve $z = f(x, y)$ which is the equation of a surface in three dimensions, is called the '*Normal Correlation Surface*'.

10·10·1. Moment Generating Function of Bivariate Normal Distribution. Let $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. By def.,

$$M_{XY}(t_1, t_2) = E\left[e^{t_1X + t_2Y}\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(t_1x + t_2y).f(x, y)\,dxdy$$

Put $\frac{x-\mu_1}{\sigma_1} = u, \frac{y-\mu_2}{\sigma_2} = v, \quad -\infty < (u, v) < \infty$

i.e., $x = \sigma_1 u + \mu_1, y = \sigma_2 v + \mu_2 \Rightarrow |J| = \sigma_1\sigma_2$

$$\therefore M_{X,Y}(t_1, t_2) = \frac{\exp(t_1\mu_1 + t_2\mu_2)}{2\pi\sqrt{1-\rho^2}} \times \iint_{u\,v} \exp\left[t_1\sigma_1 u + t_2\sigma_2 v - \frac{1}{2(1-\rho^2)}\left\{u^2 - 2\rho uv + v^2\right\}\right] dudv$$

$$= \frac{\exp(t_1\mu_1 + t_2\mu_2)}{2\pi\sqrt{1-\rho^2}} \times \iint_{u.v} \exp\left[\frac{1}{2(1-\rho^2)}\left\{(u^2 - 2\rho uv + v^2) - 2(1-\rho^2)(t_1\sigma_1 u + t_2\sigma_2 v)\right\}\right] dudv$$

We have

$$(u^2 - 2\rho uv + v^2) - 2(1-\rho^2)(t_1\sigma_1 u + t_2\sigma_2 v)$$
$$= \{(u - \rho v) - (1-\rho^2)t_1\sigma_1\}^2$$
$$+ (1-\rho^2)\{(v - \rho t_1\sigma_1 - t_2\sigma_2)^2 - t_1^2\sigma_1^2 - t_2^2\sigma_2^2 - 2\rho t_1 t_2\sigma_1\sigma_2\} \quad ..(*)$$

By taking

$$\left.\begin{aligned} u - \rho v - (1-\rho^2)\,t_1\sigma_1 &= \omega(1-\rho^2)^{1/2} \\ \text{and} \quad v - \rho\, t_1\sigma_1 - t_2\sigma_2 &= z \end{aligned}\right\} \Rightarrow dudv = \sqrt{1-\rho^2}\,dwdz$$

and using (*), we get

$$M_{X,Y}(t_1, t_2) = \exp[t_1\mu_1 + t_2\mu_2 + \tfrac{1}{2}(t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + 2\rho t_1 t_2\sigma_1\sigma_2)] \times \left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\omega^2/2}d\omega\right] \times \left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-z^2/2}\,dz\right]$$

$$= \exp\left[t_1\mu_1 + t_2\mu_2 + \tfrac{1}{2}(t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + 2\rho t_1 t_2\sigma_1\sigma_2)\right] \quad ...(10·26)$$

In particular if $(X, Y) \sim BVN(0, 0, 1, 1, \rho)$, then

$$M_{X,Y}(t_1, t_2) = \exp\left[\tfrac{1}{2}(t_1^2 + t_2^2 + 2\rho t_1 t_2)\right] \quad ...(10·26a)$$

Theorem 10·5. *Let* $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. *Then X and Y are independent if and only if* $\rho = 0$.

Proof. (*a*) If $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ and $\rho = 0$, then X and Y are independent [*c.f.* Remark 2(*a*) to Theorem 10·2, page 10·5].

Aliter. $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

$$\therefore \quad M_{X,Y}(t_1, t_2) = \exp\{t_1\mu_1 + t_2\mu_2 + \frac{1}{2}(t_1^2\sigma_1^2 + 2\rho t_1 t_2\sigma_1\sigma_2 + t_2^2\sigma_2^2)\}$$

If $\rho = 0$, then

$$M_{X,Y}(t_1, t_2) = \exp\{t_1\mu_1 + \frac{1}{2}t_1^2\sigma_1^2\} \cdot \exp\{t_2\mu_2 + \frac{1}{2}t_2^2\sigma_2^2\}$$

$$= M_X(t_1) \cdot M_Y(t_2). \qquad \ldots(*)$$

[$\because$ If $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then the marginal p.d.f.'s of X and Y are normal *i.e.*, $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$].

$(*) \Rightarrow$ X and Y are independent.

(*b*) Conversely if X and Y are independent, then $\rho = 0$ [*c.f.* Theorem 10·2]

Theorem 10·6. *(X, Y) possesses a bivariate normal distribution if and only if every linear combination of X and Y viz., $aX + bY$, $a \neq 0$, $b \neq 0$, is a normal variate.*

Proof. (*a*) Let $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then we shall prove that $aX + bY$, $a \neq 0, b \neq 0$ is a normal variate.

Since (X, Y) has a bivariate normal distribution, we have

$$M_{X,Y}(t_1, t_2) = E(e^{t_1X + t_2Y})$$

$$= e^{t_1\mu_1 + t_2\mu_2 + \frac{1}{2}(t_1^2\sigma_1^2 + 2\rho t_1 t_2\sigma_1\sigma_2 + t_2^2\sigma_2^2)} \qquad \ldots(*)$$

Then m.g.f. of $Z = aX + bY$, is given by :

$$M_Z(t) = E(e^{tZ}) = E(e^{t(aX + bY)}) = E(e^{atX + btY})$$

$$= \exp\{t(a\mu_1 + b\mu_2) + \frac{t^2}{2}(a^2\sigma_1^2 + 2\rho ab\sigma_1\sigma_2 + b^2\sigma_2^2)\},$$

[Taking $t_1 = at$, $t_2 = bt$ in $(*)$]

which is the m.g.f. of normal distribution with parameters

$$\mu = a\mu_1 + b\mu_2 ; \sigma^2 = a^2\sigma_1^2 + 2\rho ab\sigma_1\sigma_2 + b^2\sigma_2^2. \qquad \ldots(**)$$

Hence by uniqueness theorem of m.g.f.,

$$Z = aX + bY \sim N(\mu, \sigma^2),$$

where μ and σ^2 are given in (**).

(*b*) Conversely, let $Z = aX + bY$, $a \neq 0$, $b \neq 0$ be a normal variate. Then we have to prove that (X, Y) has a bivariate normal distribution.

Let $Z = aX + bY \sim N(\mu, \sigma^2)$,

where $\mu = EZ = E(aX + bY) = a\mu_x + b\mu_y$

and $\sigma^2 = \text{Var } Z = \text{Var}(aX + bY) = a^2\sigma_x^2 + 2ab\rho\sigma_x\sigma_y + b^2\sigma_y^2$

$$\therefore \quad M_Z(t) = \exp[t\mu + t^2\sigma^2/2]$$

$$= \exp[t(a\mu_x + b\mu_y) + \frac{t^2}{2}(a^2\sigma_x^2 + 2ab\rho\sigma_x\sigma_y + b^2\sigma_y^2)]$$

$$= \exp[t_1\mu_x + t_2\mu_y + \frac{1}{2}(t_1^2\sigma_x^2 + 2\rho t_1 t_2\sigma_x\sigma_y + t_2^2\sigma_y^2)] \qquad \ldots(***)$$

where $t_1 = at$ and $t_2 = bt$.

But (***) is the m.g.f. of BVN distribution with parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. Hence by uniqueness theorem of m.g.f.

$$(X, Y) \sim BVN(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$$

10·10·2. Marginal Distribution of Bivariate Normal Distribution.

The marginal distribution of random variable X is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy$$

Put $\dfrac{y - \mu_2}{\sigma_2} = u$, then $dy = \sigma_2\, du$. Therefore,

$$f_X(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \times \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho u\left(\frac{x-\mu_1}{\sigma_1}\right) + u^2\right\}\right] \sigma_2\, du$$

$$= \frac{1}{2\pi\sigma_1\sqrt{(1-\rho^2)}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \times \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{u - \rho\left(\frac{x-\mu_1}{\sigma_1}\right)\right\}^2\right] du$$

Put $\dfrac{1}{\sqrt{(1-\rho^2)}}\left[u - \rho\left(\dfrac{x-\mu_1}{\sigma_1}\right)\right] = t$, then $du = \sqrt{(1-\rho^2)}\, dt$

$$\therefore \quad f_X(x) = \frac{2}{2\pi\sigma_1} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{2}\right) dt$$

$$= \frac{1}{2\pi\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \sqrt{2\pi}$$

$$= \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right], \qquad \ldots(10\cdot27)$$

Similarly, we shall get

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx$$

$$= \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right] \qquad (10\cdot27a)$$

Hence $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ $\qquad \ldots(10\cdot27b)$

Remark. We have proved that if $(X, Y) \sim BVN(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then the marginal p.d.f.'s of X and Y are also normal. However, the converse is not true, *i.e.*, we may have joint p.d.f. $f(X, Y)$ of (X, Y) which is not

normal but the marginal p.d.f.'s may still be normal as discussed in the following illustration.

Consider the joint distribution of X and Y given by :

$$f(x, y) = \frac{1}{2}\left[\frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left\{\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\}\right.$$
$$\left. + \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 + 2\rho xy + y^2)\right\}\right]$$
$$= \frac{1}{2}[f_1(x, y) + f_2(x, y)] \; ; -\infty (x, y) < \infty \qquad \ldots(10\cdot27c)$$

where $f_1(x, y)$ is the p.d.f. of $BVN(0, 0, 1, 1, \rho)$ distribution and $f_2(x, y)$ is the p.d.f. of $BVN(0, 0, 1, 1, -\rho)$ distribution.

It can be easily verified that $f(x, y)$ is the joint p.d.f. of (X, Y) and obviously $f(x, y)$ is not the p.d.f. of bivariate normal distribution.

Marginal distribution of X in (10·27*c*)

$$f_X(x) = \frac{1}{2}\left[\int_{-\infty}^{\infty} f_1(x, y)\, dy + \int_{-\infty}^{\infty} f_2(x, y)\, dy\right]$$

But $\int_{-\infty}^{\infty} f_1(x, y)\, dy$ is the marginal p.d.f. of X, where

$(X, Y) \sim BVN(0, 0, 1, 1, \rho)$ and is given by $X \sim N(0, 1)$.

Similarly $\int_{-\infty}^{\infty} f_2(x, y)\, dy$ is the marginal p.d.f. of X, where

$(X, Y) \sim BVN(0, 0, 1, 1, -\rho)$ and is given by $X \sim N(0, 1)$.

$$\therefore \quad f_X(x) = \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} e^{-x^2/2} + \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\right]$$
$$= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \; ; -\infty < x < \infty \qquad \ldots(i)$$

$\Rightarrow X \sim N(0, 1)$ *i.e.*, the marginal distribution of X i.e. (10·27*c*) is normal.

Similarly, we can show that that the marginal p.d.f. of Y in (10·27*c*) is given by :

$$f_Y(y) = \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} e^{-y^2/2} + \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\right]$$
$$= \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \; ; -\infty < y < \infty \qquad \ldots(ii)$$

$\Rightarrow \qquad Y \sim N(0, 1)$.

Hence *if the marginal distributions of X and Y are normal (Gaussian), it does not necessarily imply that the joint distribution of (X, Y) is bivariate normal.*

For another illustration, see Question Number 17, Exercise 10(*f*).

We further note that for the joint p.d.f. (10·27*c*), on using (*i*) and (*ii*), we have

$$E(X) = 0, \; \sigma_X^2 = 1 \text{ and } E(Y) = 0, \; \sigma_Y^2 = 1.$$

$$\therefore \quad \text{Cov}(X, Y) = \frac{E(XY) - E(X)\,E(Y)}{\sigma_X\,\sigma_Y} = E(XY)$$

$$= \frac{1}{2}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_1(x, y)\, dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_2(x, y)\, dxdy\right]$$

$$= \frac{1}{2}[\rho + (-\rho)] = 0,$$

because, for $f_1(x, y)$, $(X, Y) \sim BVN(0, 0, 1, 1, \rho)$ and for $f_2(x, y)$, $(X, Y) \sim BVN(0, 0, 1, 1, -\rho)$.

$$\therefore \quad \text{Corr.}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

However, we have ; [From (*i*) and (*ii*)]

$$f_X(x) \cdot f_Y(y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2 + y^2)} \neq f(x, y)$$

$\Rightarrow$ X and Y are not independent.

The above example illustrates that we may have a joint density (non-Gaussian) of *rv*'s (X, Y) in which the marginal p.d.f.'s of X and Y are normal and $\rho(X, Y) = 0$ and yet X and Y are not independent.

10·10·3. Conditional Distributions. Conditional distribution of X for a fixed Y is given by

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_1\sqrt{(1-\rho^2)}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\left(1 - 1 - \rho^2\right)\right\}\right]$$

$$= \frac{1}{\sigma_1\sqrt{2\pi}\sqrt{(1-\rho^2)}} \times \exp\left[-\frac{1}{2(1-\rho^2)\sigma_1^2}\left\{(x-\mu_1) - \rho\,\frac{\sigma_1}{\sigma_2}(y-\mu_2)\right\}^2\right]$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_1\sqrt{(1-\rho^2)}} \times \exp\left[-\frac{1}{2(1-\rho^2)\sigma_1^2}\left\{x - \left(\mu_1 + \rho\,\frac{\sigma_1}{\sigma_2}(y-\mu_2)\right)\right\}^2\right]$$

which is the probability function of a univariate normal distribution with mean and variance given by

$$E(X \mid Y = y) = \mu_1 + \rho\,\frac{\sigma_1}{\sigma_2}(y - \mu_2) \text{ and } V(X \mid Y = y) = \sigma_1^2(1 - \rho^2)$$

Hence the conditional distribution of X for fixed Y is given by ;

$$(X \mid Y = y) \sim N\left[\mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2 (1 - \rho^2)\right] \quad \ldots(10\cdot27d)$$

Similarly the conditional distribution of random variables Y for a fixed X is

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_2\sqrt{(1 - \rho^2)}} \times \exp\left[- \frac{1}{2(1 - \rho^2)\,\sigma_2^2} \left\{(y - \mu_2) - \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1)\right\}^2\right],$$

$$-\infty < y < \infty$$

Thus the conditional distribution of Y for fixed X is given by

$$(Y \mid X = x) \sim N\left[\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1)\ ,\ \sigma_2^2\ (1 - \rho^2)\right] \quad \ldots(10\cdot27e)$$

It is apparent from the above results that the array means are collinear, *i.e.*, the regression equations are linear (involving linear functions of the independent variables) and the array variances are constant (*i.e.*, free from independent variable). We express this by saying that *the regression equations of Y on X and X on Y are linear and homoscedastic.*

For $\rho = 0$, the conditional variance $V(Y \mid X)$ is equal to the marginal variance σ_2^2 and the conditional mean $E(Y \mid X)$ is equal to the marginal mean μ_2 and the two variables become independent, which is also apparent from joint distribution function. In between the two extremes when $\rho = \pm 1$, the correlation coefficient ρ provides a measure of degree of association or interdependence between the two variables.

Example 10·27. *Show that for the bivariate normal distribution* :

$$dP = const.\ exp\left[-\frac{1}{2(1 - \rho^2)}(x^2 - 2\rho xy + y^2)\right] dx\, dy,$$

(i) M.G.F. is $M(t_1, t_2) = exp.\left[\frac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2)\right]$

(ii) Moments obey the recurrence relation,

$$\mu_{rs} = (r + s - 1)\,\rho\mu_{r-1,s-1} + (r - 1)(s - 1)(1 - \rho^2)\,\mu_{r-2,s-2}$$

Hence or otherwise, show that

$$\mu_{rs} = 0, \text{ if } r + s \text{ is odd}, \mu_{31} = 3\rho, \mu_{22} = 1 + 2\rho^2$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

Solution. (*i*) From the given probability function, we see that

$$\mu_1 = 0 = \mu_2 \quad \text{and} \quad \sigma_1^2 = 1 = \sigma_2^2.$$

$\therefore$ From (10·26*a*), we get

$$M = M(t_1, t_2) = \exp\left[\frac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2)\right]$$

(*ii*) $\dfrac{\partial M}{\partial t_1} = M(t_1 + \rho t_2)$ and $\dfrac{\partial M}{\partial t_2} = M(t_2 + \rho t_1)$

and $$\frac{\partial^2 M}{\partial t_1 \partial t_2} = \frac{\partial}{\partial t_1}\left(\frac{\partial M}{\partial t_2}\right) = \frac{\partial}{\partial t_1}[M(t_2 + \rho t_1)]$$

$$= M\rho + (t_2 + \rho t_1)(t_1 + \rho t_2)M$$

$$\therefore \quad \frac{\partial^2 M}{\partial t_1 \partial t_2} - \rho t_1 \frac{\partial M}{\rho t_1} - \rho t_2 \frac{\partial M}{\partial t_2}$$

$$= [M\rho + (t_2 + \rho t_1)(t_1 + \rho t_2)M = \rho t_1 (t_1 + \rho t_2)M - \rho t_2(t_2 + \rho t_1)M]$$

$$= M[t_1 t_2 + \rho - \rho^2 t_1 t_2] \qquad (\textit{On simplification})$$

$$= M\rho + (1 - \rho^2)M t_1 t_2$$

$$\therefore \quad \frac{\partial^2 M}{\partial t_1 \partial t_2} = \rho t_1 \frac{\partial M}{\partial t_1} + \rho t_2 \frac{\partial M}{\partial t_2} + M\rho + M\ (1 - \rho^2) t_1 t_2 \qquad \ldots(*)$$

But $$M = \exp\left[\tfrac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2)\right] = \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} \mu_{rs} \cdot \frac{t_1^r t_2^s}{r\,!\,s\,!}$$

$\therefore$ (*) gives

$$\sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \mu_{rs} \cdot \frac{t_1^{r-1} t_2^{s-1}}{(r-1)\,!\,(s-1)\,!}$$

$$= \left[\rho \sum_{r=1}^{\infty}\sum_{s=0}^{\infty} r\mu_{rs} \cdot \frac{t_1^r t_2^s}{r\,!\,s\,!} + \rho \sum_{r=0}^{\infty}\sum_{s=1}^{\infty} s\mu_{rs} \cdot \frac{t_1^r t_2^s}{r\,!\,s\,!} \right.$$

$$\left. + \rho \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} \mu_{rs} \cdot \frac{t_1^r t_2^s}{r\,!\,s\,!} + (1 - \rho^2) \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} \mu_{rs} \cdot \frac{t_1^{r+1} t_2^{s+1}}{r\,!\,s\,!} \right]$$

Equating the coefficients of $\frac{t_1^{r-1}}{(r-1)\,!} \cdot \frac{t_2^{s-1}}{(s-1)\,!}$ on both sides, we get

$$\mu_{rs} = [\rho(r-1)\mu_{r-1,s-1} + \rho(s-1)\mu_{r-1,s-1} + \rho^2 \mu_{r-1,s-1} + (1 - \rho^2)(r-1)(s-1)\mu_{r-2,s-2}]$$

$$\Rightarrow \quad \mu_{rs} = (r + s - 1)\rho\mu_{r-1,s-1} + (r-1)(s-1)(1-\rho^2)\mu_{r-2,s-2}$$

In particular

$$\mu_{31} = 3\rho\mu_{2,0} + 0 = 3\rho\sigma_1^2 = 3\rho \qquad (\because \sigma_1^2 = 1)$$

$$\mu_{22} = 3\rho\mu_{1,1} + (1 - \rho^2)\mu_{0,0} = 3\rho^2 + (1 - \rho^2).1$$

$$= (1 + 2\rho^2) \qquad (\because \mu_{11} = \rho\sigma_1\sigma_2 = \rho)$$

Also $$\mu_{03} = \mu_{30} = 0$$

$$\mu_{12} = 2\rho\mu_{0,1} + 0 = 0 \qquad (\because \mu_{01} = \mu_{10} = 0)$$

$$\mu_{23} = 4\rho\mu_{1,2} + 1{\cdot}2\,(1 - \rho^2)\mu_{0,1} = 0$$

Similarly, we will get $\mu_{21} = 0, \mu_{32} = 0$

If $r + s$ is odd, so is $(r-1) + (s-1)$, $(r-2) + (s-2)$, and so on.

And since $\mu_{30} = 0 = \mu_{03}$, $\mu_{12} = 0 = \mu_{21}$, $\mu_{23} = 0 = \mu_{32}$..., we finally get,

$$\mu_{rs} = 0, \text{ if } r + s \text{ is odd.}$$

Example 10·28. *Show that if X_1 and X_2 are standard normal variates with correlation coefficient ρ between them, then the correlation coefficient between X_1^2 and X_2^2 is given by ρ^2.*

Solution. Since X_1 and X_2, are two standard normal variates, we have

$$E(X_1) = E(X_2) = 0 \text{ and } V(X_1) = E(X_1^2) = 1 = V(X_2) = E(X_2^2)$$

$$\therefore \qquad M_{x_1, x_2}(t_1, t_2) = \exp\left[\tfrac{1}{2}(t_1^2 + 2\rho t_1 t_2 + t_2^2)\right] \qquad \text{[c.f. (10·26)]}$$

$$\text{Now} \quad \rho(X_1^2, X_2^2) = \frac{E(X_1^2 X_2^2) - E(X_1^2)\,E(X_2^2)}{\sqrt{[E(X_1^4) - \{E(X_1^2)\}^2]}\ \sqrt{[E(X_2^4) - \{E(X_2^2)\}^2]}}$$

$$\text{where} \quad E(X_1^2 X_2^2) = \text{Coefficient of } \frac{t_1^2}{2!}\cdot\frac{t_2^2}{2!} \text{ in } M(t_1, t_2) = (2\rho^2 + 1)$$

$$E(X_1^4) = \text{Coefficient of } \frac{t_1^4}{4!} \text{ in } M(t_1, t_2) = 3$$

$$E(X_2^4) = \text{Coefficient of } \frac{t_2^4}{4!} \text{ in } M(t_1, t_2) = 3$$

$$\therefore \quad \rho(X_1^2, X_2^2) = \frac{2\rho^2 + 1 - 1}{\sqrt{(3-1)}\sqrt{(3-1)}} = \rho^2$$

Example 10·29. *The variables X and Y with zero means and standard deviations σ_1 and σ_2 are normally correlated with correlation coefficient ρ. Show that U and V defined as*

$$U = \frac{X}{\sigma_1} + \frac{Y}{\sigma_2} \quad \text{and} \quad V = \frac{X}{\sigma_1} - \frac{Y}{\sigma_2}$$

are independent normal variates with variances $2(1 + \rho)$ and $2(1 - \rho)$ respectively.

Solution. We are given that

$$dF(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right\}\right] dx\,dy$$

$$-\infty < (x, y) < \infty$$

$$\text{Also} \quad u = \frac{x}{\sigma_1} + \frac{y}{\sigma_2},\ v = \frac{x}{\sigma_1} - \frac{y}{\sigma_2}$$

$$\therefore \tfrac{1}{2}(u + v) = \frac{x}{\sigma_1} \text{ and } \tfrac{1}{2}(u - v) = \frac{y}{\sigma_2} \Rightarrow x = \frac{\sigma_1}{2}(u + v) \text{ and } y = \frac{\sigma_2}{2}(u - v)$$

Jacobian of transformation J is given by

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{1}{2}\sigma_1 & \frac{1}{2}\sigma_1 \\ \frac{1}{2}\sigma_2 & -\frac{1}{2}\sigma_2 \end{vmatrix} = -\frac{\sigma_1\sigma_2}{2}$$

$$\frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2} = \frac{1}{4}\left[(u + v)^2 + (u - v)^2\right] = \frac{1}{2}(u^2 + v^2)$$

$$dF_1(u, v) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \times \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{1}{2}(u^2 + v^2) - 2\rho\left(\frac{u^2 - v^2}{4}\right)\right\}\right]\frac{\sigma_1\sigma_2}{2}\,du\,dv$$

$$= \frac{1}{2\pi.\ 2\sqrt{(1-\rho^2)}} \exp\left[-\frac{1}{4(1-\rho^2)}\{(1-\rho)u^2 + (1+\rho)v^2\}\right] du\, dv$$

$$= \frac{1}{2\pi\sqrt{2(1-\rho)}\sqrt{2(1+\rho)}} \exp\left[-\frac{u^2}{2(1+\rho)2} - \frac{v^2}{2(1-\rho)2}\right] du\, dv$$

$$= \left[\frac{1}{\sqrt{2\pi}\sqrt{2(1+\rho)}} \cdot \exp\left\{-\frac{u^2}{2(1+\rho)2}\right\}\right] du$$

$$\times \left[\frac{1}{\sqrt{2\pi}\sqrt{2(1-\rho)}} \cdot \exp\left\{-\frac{v^2}{2(1-\rho)2}\right\}\right] dv$$

$$= [f_1(u)du]\,[f_2(v)dv], \text{ (say)}$$

where $$f_1(u) = \frac{1}{\sqrt{2\pi}\sqrt{2(1+\rho)}} \cdot \exp\left\{-\frac{u^2}{2(1+\rho)2}\right\}$$

and $$f_2(v) = \frac{1}{\sqrt{2\pi}\sqrt{2(1-\rho)}} \cdot \exp\left\{-\frac{v^2}{2(1-\rho)2}\right\}$$

Hence U and V are independently distributed, U as $N[0, 2(1+\rho)]$ and V as $N[0, 2(1-\rho)]$.

Aliter. Find joint m.g.f. of U and V viz.,

$$M(t_1, t_2) = E(e^{t_1U + t_2V}) = E\left[e^{X(t_1+t_2)/\sigma_1 + Y(t_1-t_2)/\sigma_2}\right]$$

and use $E(e^{t_1X + t_2Y}) = \exp[(t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + 2\rho t\, t_2\sigma_1\sigma_2)/2]$

Example 10·30. *If X and Y are standard normal variates with co-efficient of correlation ρ, show that*

(i) Regression of Y on X is linear.

(ii) $X + Y$ and $X - Y$ are independently distributed.

(iii) $Q = \dfrac{X^2 - 2\rho XY + Y^2}{(1-\rho^2)}$ is distributed like a chi-square, i.e., as that of the sum of the squares of standard normal variates.

(Madras Univ. B.E., 1990)

Solution. *(i)* c.f. § 10·10·3.

(ii) Let $u = x + y$ and $v = x - y$

$$dF(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right] dxdy$$

Now $$x = \frac{u+v}{2},\ y = \frac{u-v}{2}$$

$$\therefore \quad J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

$$dG(u, v) = C \exp\left[-\frac{1}{2(1-\rho^2)\,.\,4}\{2(u^2+v^2) - 2\rho(u^2-v^2)\}\right] dudv$$

where $C = \dfrac{1}{4\pi\sqrt{1-\rho^2}}$

$$\therefore dG(u, v) = C \exp\left[-\frac{1}{4(1-\rho^2)}\{(1-\rho)u^2 + (1+\rho)v^2\}\right]du\,dv$$

$$= \left[C_1 \exp\left\{-\frac{u^2}{4(1+\rho)}\right\}du\right] \times \left[C_2 \exp\left\{-\frac{v^2}{4(1-\rho)}\right\}dv\right]$$

$$= [g_1(u)du]\,[g_2(v)dv], \text{ (say).}$$

Hence U and V are independently distributed.

$$(iii)\quad M_Q(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{tQ}\, dF(x, y)$$

$$= \frac{1}{2\pi\sqrt{(1-\rho^2)}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(tQ) \times \exp\left[-\frac{1}{2(1-\rho^2)}\left\{x^2 - 2\rho xy + y^2\right\}\right]dx\,dy$$

$$= \frac{1}{2\pi\sqrt{(1-\rho^2)}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(tQ - \frac{Q}{2}\right)dx\,dy$$

$$= \frac{1}{2\pi\sqrt{(1-\rho^2)}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[-\frac{Q}{2}(1-2t)\right]dx\,dy$$

Put $\sqrt{(1-2t)}\,x = u$ and $\sqrt{(1-2t)}\,y = v$

$$\therefore \quad dx = \frac{du}{\sqrt{(1-2t)}} \text{ and } dy = \frac{dv}{\sqrt{(1-2t)}}$$

$$\text{Also} \quad Q = \frac{1}{(1-\rho^2)}[x^2 - 2\rho xy + y^2] = \frac{1}{(1-\rho^2)}\left[\frac{u^2 - 2\rho uv + v^2}{1-2t}\right]$$

$$\therefore \quad M_Q(t) = \frac{1}{2\pi\sqrt{(1-\rho^2)}\,(1-2t)} \times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-\rho^2)}\,(u^2 - 2\rho uv + v^2)\right]du\,dv$$

$$= \frac{1}{(1-2t)} \cdot 1 = (1-2t)^{-1}$$

which is the m.g.f. of chi-square (χ^2) variate* with n (=2) degrees of freedom.

Example 10·31. *Let X and Y be independent standard normal variates. Obtain the m.g.f. of XY.* [*Gauhati Univ. M.Sc., 1992*]

Solution. We have, by definition :

$$M_{XY}(t) = E(e^{tXY}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{txy} \cdot f(x, y)\, dxdy$$

Since X and Y are independent standard normal variates, their joint p.d.f. $f(x, y)$ is given by :

$$f(x, y) = f_1(x) \cdot f_2(y) = \frac{1}{2\pi} e^{-x^2/2}\, e^{-y^2/2} \; ; -\infty < (x, y) < \infty$$

$$\therefore \quad M_{XY}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2 - 2txy + y^2)}\, dxdy$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2(1-t^2)}\right.$$

$$\left. \times \left\{\frac{x^2}{1/(1-t^2)} - \frac{2t.x.y}{(1/\sqrt{1-t^2})(1/\sqrt{1-t^2})} + \frac{y^2}{1/(1-t^2)}\right\}\right] dxdy \quad \ldots(*)$$

If $(U, V) \sim BVN(0, 0, \sigma_1^2, \sigma_2^2, \rho)$, then we have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right\}} dxdy = 1$$

$$\Rightarrow \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right\}} dxdy = 2\pi\sigma_1\sigma_2\sqrt{1-\rho^2} \quad \ldots(**)$$

Comparing (*) and (**) with

$$\sigma_1^2 = \sigma_2^2 = \frac{1}{(1-t^2)} \quad \text{and} \quad \rho = t, \text{ we get}$$

$$M_{XY}(t) = \frac{1}{2\pi} \cdot 2\pi \frac{1}{\sqrt{1-t^2}} \cdot \frac{1}{\sqrt{1-t^2}} \cdot \sqrt{1-t^2}$$

$$\Rightarrow \quad M_{XY}(t) = (1-t^2)^{1/2} \; ; -1 < t < 1$$

Example 10·32. *Let X and Y have bivariate normal distribution with parameters :*

$\mu_X = 5, \mu_Y = 10, \sigma_X^2 = 1, \sigma_Y^2 = 25$ *and Corr* $(X, Y) = \rho$.

(a) If $\rho > 0$, *find* ρ *when* $P(4 < Y < 16 \mid X = 5) = 0{\cdot}954$

[*Delhi Univ. B.Sc. (Math. Hons.), 1993, '83*]

---

*Chi-square distribution is discussed in Chapter 13

(b) If $\rho = 0$, find $P(X + Y \le 16)$.

Solution. Since $(X, Y) \sim BVN(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, the conditional distribution of Y given $X = x$ is also normal.

$$(Y \mid X = x) \sim N\left[\mu = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X),\ \sigma^2 = \sigma_Y^2(1 - \rho^2)\right]$$

$$\therefore\ (Y \mid X = 5) \sim N\left[\mu = 10 + \rho \times \frac{5}{1}(5 - 5),\ \sigma^2 = 25(1 - \rho^2)\right]$$

$$= N[\mu = 10,\ \sigma^2 = 25(1 - \rho^2)]$$

We want ρ so that

$$P(4 < Y < 16 \mid X = 5) = 0.954$$

where $$Z = \frac{Y - \mu}{\sigma} = \frac{Y - 10}{5\sqrt{(1 - \rho^2)}} \sim N(0, 1)$$

$$\Rightarrow \quad P\left(\frac{4 - 10}{\sigma} < Z < \frac{16 - 10}{\sigma}\right) = 0.954$$

$$\Rightarrow \quad P\left(\frac{-6}{\sigma} < Z < \frac{6}{\sigma}\right) = 0.954 \qquad \ldots(*)$$

But we know that if $Z \sim N(0, 1)$, then

$$P(-2 < Z < 2) = 0.954 \qquad \ldots(**)$$

Comparing (*) and (**), we get

$$\frac{6}{\sigma} = 2 \ \Rightarrow\ \sigma = 3 \ \Rightarrow\ \sigma^2 = 9 = 25(1 - \rho^2)$$

$$\therefore \quad 1 - \rho^2 = \frac{9}{25} \ \Rightarrow\ \rho^2 = \frac{16}{25} \ \Rightarrow\ \rho = \frac{4}{5} = 0.8 \qquad (\because \rho > 0)$$

(b) Since (X, Y) have bivariate normal distribution,

$$\rho = 0 \Rightarrow X \text{ and } Y \text{ are independent } rv\text{'s}$$

and $$X \sim N(\mu_X, \sigma_X^2) \text{ and } Y \sim N(\mu_Y, \sigma_Y^2)$$

$$\therefore \quad X + Y \sim N(\mu = \mu_X + \mu_Y, \sigma^2 = \sigma_X^2 + \sigma_Y^2) = N(15, 26)$$

Hence

$$P(X + Y \le 16) = P\left(Z \le \frac{16 - 15}{\sqrt{26}}\right)$$

where $$Z = \frac{(X + Y) - \mu}{\sigma} \sim N(0, 1).$$

$$\therefore \quad P(X + Y \le 16) = P\left(Z \le \frac{1}{\sqrt{26}}\right) = \Phi\left(1/\sqrt{26}\right),$$

where $\Phi(z) = P(Z \le z)$, is the distribution function of standard normal variate.

Remark . $P(X + Y \le 16) = P\left(Z \le \dfrac{1}{5.099}\right) = P(Z \le 0.196)$

$$= 0.5 + P(0 \le Z \le 0.196)$$
$$= 0.5 + 0.0793 \text{ (approx.)}$$
$$= 0.5793.$$

EXERCISE 10(f)

1. (*a*) Define conditional and marginal distributions. If X and Y follow bivariate normal distribution, find (*i*) the conditional distribution of X given Y and (*ii*) the marginal distribution of X. Show that the conditional mean of X is dependent on the given Y, but the conditional variance is independent of it.

(*b*) Define Bivariate Normal distribution. If (X, Y) has a bivariate normal distribution, find the marginal density function $f_X(x)$ of X.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1988*]

2. (*a*) The marks X and Y scored by candidates in an examination in two subjects Mathematics and Statistics are known to follow a bivariate normal distribution. The mean of X is 52 and its standard deviation is 15, while Y has mean 48 and standard deviation 13. Also the coefficient of correlation between X and Y is 0·6.

Write down the joint distribution of X and Y. If 100 marks in the aggregate are needed for a pass in the examination, show how to calculate the proportion of candidates who pass the examination ?

(*b*) A manufacturer of electric bulbs, in his desire for putting only good bulbs for sale, rejects all bulbs for which a certain quality characteristic X of the filament is less than 65 units. Assume that the quality characteristic X and the life Y, of the bulb in hours are jointly normally distributed with parameters given below :

	X	Y
Mean	80	1100
Standard deviation	10	10

Correlation coefficient $\rho(X, Y) = 0{\cdot}60$

Find (*i*) the proportion of bulbs produced that will burn for less than 1000 hours, (*ii*) the proportion of bulbs produced that will be put for sale, (*iii*) the average life of bulbs put for sale.

3. (*a*) Determine the parameters of the bivariate normal distribution :

$$f(x, y) = k\, exp\left[-\frac{8}{27}\{(x-7)^2 - 2(x-7)(y+5) + 4(y+5)^2\}\right]$$

Also find the value of k.

(*b*) For the bivariate normal distribution :

$$(X, Y) \sim BVN\left(1, 2, 4^2, 5^2, \frac{12}{13}\right)$$

find (*i*) $P(X > 2)$, (*ii*) $P(X > 2 \mid Y = 2)$.

(*c*) The bivariate random variable (X_1, X_2) have a bivariate normal distribution with means 60 and 75 and standard deviations 6 and 12 with a correlation coefficient of 0·55. Find the following probabilities :

(*i*) $P(65 \le X_1 \le 75)$, (*ii*) $P(71 \le X_2 \le 80 \mid X_1 = 55)$ and (*iii*) $P(|X_1 - X_2| \ge 15)$.

4. For a bivariate normal distribution :

$$f_{XY}(x, y) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\},$$

$$-\infty < (x, y) < \infty$$

Find (*i*) marginal distribution of X and Y,

(*ii*) conditional distribution of Y given X,

(*iii*) distribution of $\frac{1}{(1-\rho^2)}[x^2 - 2\rho xy + y^2]$,

and (*iv*) show that in general X and Y are stochastically dependent and will be independent if and only if $\rho = 0$.

5. Let the joint p.d.f. of X and Y be

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)}{\sigma_1}\frac{(y-\mu_2)}{\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

where $-\infty < x < \infty, -\infty < y < \infty, -1 < \rho < 1$.

(*i*) Find the marginal distribution of X.

(*ii*) Find the conditional distribution of Y given $X = x$.

(*iii*) Show that the regression of Y on X is linear and homoscedastic.

(*iv*) Find $P[3 < Y < 8 \mid X = 7)$, given that $\mu_1 = 3$, $\mu_2 = 1$, $\sigma_1^2 = 16$, $\sigma_2^2 = 25$, $\rho = 0{\cdot}6$,

(*v*) Find the probability of the simultaneous materialization of the inequalities, $X > E(X)$ and $Y > E(Y)$

Hint. (*v*) Required probability p is given by

$$p = P[X > E(X), Y > E(Y)] = P[X > \mu_1) \cap (Y > \mu_2)]$$

$$= \int_{\mu_1}^{\infty}\int_{\mu_2}^{\infty} f(x, y)\,dx\,dy$$

$$= \int_0^{\infty}\int_0^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot \exp\left[-\frac{1}{2(1-\rho^2)}(u^2 - 2\rho uv + v^2)\right]du\,dv,$$

$$\left(u = \frac{x-\mu_1}{\sigma_1},\ v = \frac{y-\mu_2}{\sigma_2}\right).$$

Now proceed as in Hint to Question Number 9(*b*).

6. Let the random variables X and Y be assumed to have a joint bivariate normal distribution with

$$\mu_1 = \mu_2 = 0, \sigma_1 = 4, \sigma_2 = 3 \text{ and } r(X, Y) = 0{\cdot}8.$$

(*i*) Write down the joint density function of X and Y.

(*ii*) Write down the regression of Y on X.

(*iii*) Obtain the joint density of $X + Y$ and $X - Y$.

7. For the distribution of random variables X and Y given by

$$dF = k\exp\left[-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right]dx\,dy;\ -\infty \le x \le \infty,\ -\infty \le y \le \infty.$$

Obtain

(*i*) the constant k,

(*ii*) the distributions of X and Y,

(*iii*) the distributions of X for given Y and of Y for given X,

(*iv*) the curves of regression of Y on X and of X on Y,

and (*v*) the distributions of $X + Y$ and $X - Y$.

8. Let (X, Y) be a bivariate normal random variable with $E(X) = E(Y) = 0$, Var (X) = Var $(Y) = 1$ and Cov $(X, Y) = \rho$. Show that the random variable $Z = Y/X$ has a Cauchy distribution.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1989*]

Ans. $f(z) = \frac{1}{\pi}\left[\frac{(1-\rho^2)^{1/2}}{(1-\rho^2)+(z-\rho)^2}\right], -\infty < z < \infty.$

9. (*a*) If $(X, Y) \sim N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, prove that

$$P(X > \mu_x \cap Y > \mu_y) = \frac{1}{4} + \frac{\sin^{-1}\rho}{2\pi}$$

[*Delhi Univ. M.Sc. (Stat.), 1987*]

(*b*) If $(X, Y) \sim N(0, 0, 1, 1, \rho)$ then prove that

$$P(X > 0 \cap Y > 0) = \frac{1}{4} + \frac{\sin^{-1}\rho}{2\pi}.$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1990*]

Hint. $p = P(X > 0 \cap Y > 0)$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \times \int_0^\infty \int_0^\infty \exp\left[-\frac{1}{2(1-\rho^2)}\{x^2 - 2\rho xy + y^2\}\right] dx\,dy$$

Put $x = r\cos\theta$, $y = r\sin\theta \Rightarrow |J| = r$; $0 < r < \infty$, $0 \le \theta \le \pi/2$

$$\therefore\ p = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_0^\infty \int_0^{\pi/2} \exp\left[-\frac{r^2}{2(1-\rho^2)}(1 - \rho\sin 2\theta)\right] r\,dr\,d\theta$$

Now integrate first w.r. to r and then w.r. to θ.

10. (*a*) Let X_1 and X_2 be two independent normally distributed variables with zero means and unit variances. Let Y_1 and Y_2 be the linear functions of X_1 and X_2 defined by

$$Y_1 = m_1 + l_{11}X_1 + l_{12}X_2,\ Y_2 = m_2 + l_{21}X_1 + l_{22}X_2$$

Show that Y_1 and Y_2 are normally distributed with means m_1 and m_2, variances $\mu_{20} = l_{11}^2 + l_{12}^2$, $\mu_{02} = l_{21}^2 + l_{22}^2$, and covariance $\mu_{11} = l_{11}l_{21} + l_{12}l_{22}$.

(*b*) Let X_1 and X_2 be independent standard normal variates. Show that the variates Y_1, Y_2 defined by

$Y_1 = a_1 + b_{11}X_1 + b_{12}X_2$, $Y_2 = a_2 + b_{21}X_1 + b_{22}X_2$ are dependent normal variates and find their mean and variance.

Hint. Y_1 and Y_2, being linear combination of S.N.V's are also normally distributed. To prove that they are dependent, it is sufficient to prove that $r(Y_1, Y_2) \neq 0$. [*c.f.* Remark 2 to Theorem 10·2)

11. (*a*) Show that, if X and Y are independent normal variates with zero means and variances σ_1^2 and σ_2^2 respectively, the point of inflexion of the curve of intersection of the normal correlation surface by planes through the z-axis, lie on the elliptical cylinder,

$$\frac{X^2}{\sigma_1^2} + \frac{Y^2}{\sigma_2^2} = 1$$

(*b*) If X and Y are bivariate normal variates with standard deviations unity and with correlation coefficient ρ, show that the regression of X^2 (Y^2) on Y^2 (X^2) is strictly linear. Also show that the regression of X (Y) on Y^2 (X^2) is not linear.

12. For the bivariate normal distribution :

$$dF = k \exp\left[-\frac{2}{3}(x^2 - xy + y^2 - 3x + 3y + 3)\right] dx\, dy,$$

obtain (*i*) the marginal distribution of Y, and

(*ii*) the conditional distribution of Y given X.

Also obtain the characteristic function of the above bivariate normal ditribution and hence the covariance between X and Y.

13. Let f and g be the p.d.f.'s with corresponding distribution functions F and G. Also let

$$h(x, y) = f(x)\, g(y)\, [1 + \alpha\, (2F(x) - 1)\, (2G(y) - 1)],$$

where $|\alpha| \leq 1$, is a constant and h is a bivariate p.d.f. with marginal p.d.f.'s f and g. Further let f and g be p.d.f.'s of $N(0, 1)$ distribution. Then prove that :

$$\text{Cov}\,(X, Y) = \alpha/\pi$$

14. If $(X, Y) \sim BVN\,(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, compute the correlation coefficient between e^X and e^Y.

Hint. Let $U = e^X$, $V = e^Y$.

$$\mu'_{rs} = E(U^r.V^s) = E\left[e^{rX + sY}\right]$$

$$= \exp\left[r\mu_1 + s\mu_2 + \frac{1}{2}(r^2\sigma_1^2 + s^2\sigma_2^2 + 2\rho rs)\right]$$

[c.f. m.g.f. of $B.V.N.$ distribution : $t_1 = r$, $t_2 = s$]

Now $E(U) = \mu'_{10}$; $E(U^2) = \mu'_{20}$, $E(UV) = \mu_{11}'$ and so on.

Ans. $$\rho(U,V) = \frac{e^{\rho\sigma_1\sigma_2} - 1}{[(e^{\sigma_1^2} - 1)\,(e^{\sigma_2^2} - 1)]^{1/2}}$$

15. If $(X, Y) \sim BVN\,(0, 0, 1, 1, \rho)$, find $E\,[\max\,(X, Y)]$.

Hint. $\max\,(X, Y) = \frac{1}{2}(X + Y) + \frac{1}{2}|X - Y|$

and $Z = X - Y \sim N\,[0, 2\,(1 - \rho)]$ [c.f.Theorem 10·6]

Ans. $E\,[\max\,(X, Y)] = \left(\frac{1 - \rho}{\pi}\right)^{1/2}$

16. If $(X, Y) \sim BVN\,(0, 0, 1, 1, \rho)$ with joint p.d.f. $f(x, y)$ then prove that

(*a*) $$P(XY > 0) = \frac{1}{2} + \frac{1}{\pi} \cdot \sin^{-1}(\rho).$$

Hint. $P(XY > 0) = P(X > 0 \cap Y > 0) + P(X < 0 \cap Y < 0)$
$= 2P(X > 0 \cap Y > 0)$ [By symmetry]

Now proceed as in Hint to Question No. 9(*b*).

(*b*) $$2\pi \int_{-\infty}^{0} \int_{-\infty}^{0} f(x, y)\, dxdy = \pi + \sin^{-1} \rho$$

17. The joint density of *r.v's* (X, Y) is given by :

$$f(x, y) = \frac{1}{2\pi} \cdot \exp\left[-(x^2 + y^2)/2\right] \times \left[1 + xy \exp\left\{-(x^2 + y^2 - 2)/2\right\}\right];$$

$$-\infty < (x, y) < \infty$$

(*i*) Verify that $f(x, y)$ is a p.d.f.

(*ii*) Show that the marginal distribution of each of X and Y is normal.

(*iii*) Are X and Y independent ?

Ans. (*ii*) $X \sim N(0, 1)$, $Y \sim N(0, 1)$

(*ii*) X and Y are not independent.

18. Show by means of an example that the normality of conditional p.d.f.'s does not imply that the bivariate density is normal.

Hint. Consider $f(x, y) =$ constant. $\exp\left[-(1 + x^2)(1 + y^2)\right]$; $-\infty < (x, y) < \infty$

Then $(Y \mid x) \sim N\left(0, \frac{1}{2(1 + x^2)}\right)$ and $(X \mid y) \sim N\left(0, \frac{1}{2(1 + y^2)}\right)$

19. For a bivariate normal *r.v.* (X, Y), does the conditional p.d.f. of (X, Y) given $X + Y = c$, (constant) exist ? If so find it. If not, why not ?

Ans. No, since $P(X + Y = c) = 0$.

20. Let

$$f(x, y) = \frac{1}{2}\left[\frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left\{-\frac{1}{2(1 - \rho^2)}(x^2 - 2\rho xy + y^2)\right\} + \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left\{-\frac{1}{2(1 - \rho^2)}(x^2 + 2\rho xy + x^2)\right\}\right]$$

$$-\infty < x < \infty, \quad -\infty < y < \infty$$

then show that :

(*i*) $f(x, y)$ is a joint p.d.f. such that both marginal densities are normal but $f(x, y)$ is not bivariate normal.

(*ii*) X and Y have zero correlation but X and Y are not independent.

[Delhi Univ. B.Sc. (Stat. Hons.), 1989]

21. Let X, Y be normally correlated variates with zero means and variances σ_1^2, σ_2^2 and if

$$W = \frac{X}{\sigma_1}, \; Z = \frac{1}{\sqrt{(1 - \rho^2)}}\left\{\frac{Y}{\sigma_2} - \frac{\rho X}{\sigma_1}\right\}$$

Show that $$\frac{\partial(w, z)}{\partial(x, y)} = \frac{1}{\sigma_1 \sigma_2 \sqrt{(1 - \rho^2)}}$$

and $$W^2 + Z^2 = \frac{1}{(1-\rho^2)}\left[\frac{X^2}{\sigma_1^2} - \frac{2\rho XY}{\sigma_1\sigma_2} + \frac{Y^2}{\sigma_2^2}\right]$$

Deduce that the joint probability differential of W and Z is

$$dP = \frac{1}{2\pi} \cdot \exp\left[-\frac{1}{2}(w^2 + z^2)\right] dw\, dz$$

and hence that W, Z are independent normal variates with zero means and unit S.D.'s [*Meerut Univ. M.Sc., 1993*]

Hence or otherwise obtain the m.g.f. of the bivariate normal distribution.

22. From a standard bivariate normal population, a random sample of n observations (X_i, Y_i), $(i = 1, 2, \ldots, n)$ is drawn. Show that the distribution of

$$Z_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \text{ and } Z_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

has the moment generating function :

$$\text{Constant}\left[\left(1 - \frac{2t_1}{n}\right)\left(1 - \frac{2t_2}{n}\right) - \frac{4\rho^2 t_1 t_2}{n^2}\right]^{-n/2}$$

Hint. $M_{Z_1, Z_2}(t_1, t_2) = \left[E \exp\left(\frac{t_1 x^2}{n} + \frac{t_2 y^2}{n}\right)\right]^n$

$$= \left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[x^2\left(\frac{t_1}{n} - \frac{1}{2(1-\rho^2)}\right) + \left(\frac{\rho}{1-\rho^2}\right)xy + y^2\left(\frac{t_2}{n} - \frac{1}{2(1-\rho^2)}\right)\right] dx\,dy\right\}^n$$

Now use the result

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp[-(ax^2 + 2hxy + by^2)]\, dx\,dy = \frac{\pi}{\sqrt{ab - h^2}}$$

and simplify.

10·11. Multiple and Partial Correlation. When the values of one variable are associated with or influenced by other variable, *e.g.*, the age of husband and wife, the height of father and son, the supply and demand of a commodity and so on, Karl Pearson's coefficient of correlation can be used as a measure of linear relationship between them. But sometimes there is interrelation between many variables and the value of one variable may be influenced by many others, *e.g.*, the yield of crop per acre say (X_1) depends upon quality of seed (X_2), fertility of soil (X_3), fetilizer used (X_4), irrigation facilities (X_5), weather conditions (X_6) and so on. Whenever we are interested in studying the joint effect of a group of variables upon a variable not included in that group, our study is that of *multiple correlation and multiple regression.*

Suppose in a trivariate or multi-variate distribution we are interested in the relationship between two variables only. The are two alternatives, *viz.*, (*i*) we

consider only those two members of the observed data in which the other members have specified values or (*ii*) we may eliminate mathematically the effect of other variates on two variates. The first method has the disadvantage that it limits the size of the data and also it will be applicable to only the data in which the other variates have assigned values. In the second method it may not be possible to eliminate the entire influence of the variates but the linear effect can be easily eliminated. The correlation and regression between only two variates eliminating the linear effect of other variates in them is called the *partial correlation and partial regression.*

10·11·1. Yule's Notation. Let us consider a distribution involving three random variables X_1, X_2 and X_3. Then the equation of the plane of regression of X_1 on X_2 and X_3 is

$$X_1 = a + b_{12\cdot3}X_2 + b_{13\cdot2}X_3 \qquad \ldots(10\cdot28)$$

Without loss of generality we can assume that the variables X_1, X_2 and X_3 have been measured from their respective means, so that

$$E(X_1) = E(X_2) = E(X_3) = 0$$

Hence on taking expectation of both sides in (10·28), we get $a = 0$.

Thus the plane of regression of X_1 on X_2 and X_3 becomes

$$X_1 = b_{12\cdot3}X_2 + b_{13\cdot2}X_3 \qquad \ldots(10\cdot28a)$$

The coefficients $b_{12\cdot3}$ and $b_{13\cdot2}$ are known as the *partial regression coefficients* of X_1 on X_2 and of X_1 on X_3 respectively.

$$e_{1\cdot23} = b_{12\cdot3}X_2 + b_{13\cdot2}X_3$$

is called the estimate of X_1 as given by the plane of regression (10·28*a*) and the quantity

$$X_{1\cdot23} = X_1 - b_{12\cdot3}X_2 - b_{13\cdot2}X_3,$$

is called the *error of estimate* or *residual.*

In the general case of n variables $X_1, X_2, \ldots, X_n$ the equation of the plane of regression of X_1 on $X_2, X_3, \ldots, X_n$ becomes

$$X_1 = b_{12\cdot34\ldots n}X_2 + b_{13\cdot24\ldots n}X_3 + \ldots + b_{1n\cdot23\ldots(n-1)}X_n$$

The error of estimate or residual is given by

$$X_{1\cdot23\ldots n} = X_1 - (b_{12\cdot34\ldots n}X_2 + b_{13\cdot24\ldots n}X_3 + \ldots + b_{1n\cdot23\ldots(n-1)}X_n)$$

The notations used here are due to Yule. The subscripts before the dot (.) are known as *primary subscripts* and those after the dot are called *secondary subscripts.* The order of a regression coefficient is determined by the number of secondary subscripts, *e.g.*,

$$b_{12\cdot3}, b_{12\cdot34}, \ldots, b_{12\cdot34\ldots n}$$

are the regression coefficients of order 1, 2, ... $(n-2)$ respectively. Thus in general, a regression coefficient with p-secondary subscripts will be called a regression co-efficient of order 'p'. It may be pointed out that the order in which the secondary subscripts are written is immaterial but the order of the primary subscripts is important, *e.g.*, in $b_{12\cdot34\ldots n}$, X_2 is independent while X_1 is dependent variable but in $b_{21\cdot34\ldots n}$, X_1 is independent while X_2 is dependent

Substituting in (**), we get

$$2-\frac{\theta}{2N}-3\times\frac{17}{75}=\frac{78}{75} \quad \Rightarrow \quad \hat{\theta}=\frac{42}{75}N$$

Substituting in (***), we get

$$\frac{\alpha}{2}\left(1-\frac{42}{75}\right)=\frac{17}{75} \quad \Rightarrow \quad \hat{\alpha}=\frac{34}{33}$$

15·14. Method of Least Squares.* The principle of least squares is used to fit a curve of the form

$$y=f(x, a_0, a_1, \ldots, a_n) \qquad \ldots(15\cdot62)$$

where a_i's are unknown parameters, to a set of n sample observations (x_i, y_i); $i = 1, 2, \ldots, n$ from a bivariate population. It consists in minimising the sum of squares of residuals, *viz.*,

$$E=\sum_{i=1}^{n}[y_i-f(x_i, a_0, a_1, \ldots, a_n)]^2 \qquad \ldots(15\cdot63)$$

subject to variations in $a_0, a_1, \ldots, a_n$.

The normal equations for estimating $a_0, a_1, \ldots, a_n$ are given by

$$\frac{\partial E}{\partial a_i}=0; \quad i=1, 2, \ldots, n \qquad \ldots(15\cdot64)$$

Remarks. 1. In chapter 9, we have discussed in detail the method of least squares for fitting linear regression (§ 9·1·1), polynomial regression (§ 9·1·3) and the exponential family of curves reducible to linear regression (§ 9·3). In chapter 10 § 10·12·1, we have discussed the method of fitting multiple linear regression.

2. If we are estimating $f(x, a_0, a_1, \ldots, a_n)$ as a linear function of the parameters $a_0, a_1, \ldots, a_n$, the x's being known given values, the least square estimators obtained as linear functions of the y's will be MVU estimators.

EXERCISE 15(b)

1. (*a*) State and explain the principle of maximum likelihood for estimation of population parameter.

(*b*) (*i*) Describe the M.L. method of estimation and discuss five of its optimal properties.

(*ii*) Examine a situation when M.L. method fails and explain how you tackle such situations.

(*c*) Define the likelihood function for a random sample drawn from (*i*) a discrete population, (*ii*) a continuous population.

Find the likelihood function for a random sample of size n from each of the following populations :

(*a*) Normal (m, σ^2), (*b*) Binomial (n, p), (*c*) Poisson (μ), (*d*) Uniform on (a, b). [*Calcutta Univ. B.Sc. (Maths. Hons.), 1991*]

---

* For detailed discussion see Chapter 9.

2. (*a*) A random variable X takes the values 0 and 1 with respective probabilities p and $1-p$. Obtain on the basis of random sample of size n, the maximum likelihood estimator of p.

(*b*) Obtain the maximum likelihood estimator for the distribution having the probability mass function :

$$f(x, \theta) = \theta^x (1-\theta)^{1-x}, x = 0, 1 ; 0 \le \theta \le 1$$

[***Calcutta Univ. B.Sc. (Maths. Hons.), 1986***]

(*c*) Obtain the maximum likelihood estimator of θ in the following cases :

(*i*) $f(x, \theta) = \frac{1}{\theta} . \exp(-x/\theta) ; x \ge 0, \theta > 0$

(*ii*) $f(x, \theta) = {}^nC_x \theta^x (1-\theta)^{n-x} ; x = 0, 1, 2, \ldots, n$

3. Suppose that X has a distribution $N(\mu, \sigma^2)$, that is, the p.d.f of X is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Using M.L. estimation, determine μ and σ^2. What conclusions do you draw on the nature of the result so obtained ?

4. (*a*) Explain the technique of the method of maximum likelihod and give a formula for the large sample standard error of the maximum-likelihood estimator.

(*b*) For the distribution with p.d.f.

$f(x, \theta) = \theta e^{-\theta x}$, $(x \ge 0 ; \theta > 0)$, find the maximum likelihood estimators of θ and $E(X)$, and obtain their large-sample standard errors.

(*c*) X is a random variable such that

$$P(X \le x) = 0, \text{ for } x < 0$$
$$= 1 - e^{-x\theta}, \text{ for } x \ge 0$$

Based on n independent observations on X, obtain the maximum likelihood estimator of $E(X)$.

5. (*a*) Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution with probability density function :

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}; 0 < x < \infty, \ 0 < \theta < \infty$$

Find the maximum likelihood estimator of θ.

[***Madras Univ. B.Sc. Sept., 1988***]

(*b*) For the distribution :

$$dF(x) = \frac{1}{\theta^p \Gamma(p)} \exp(-x/\theta) x^{p-1} ; 0 \le x < \infty, p > 0, \theta > 0$$

where p is known, find out the maximum likelihood estimate of θ on the basis of a random sample of size n from the distribution. Find the variance of the estimate.

6. (*a*) If x_i $(i = 1, 2, \ldots n)$ is an observed random sample from the distribution having p.d.f.

$$f_\lambda(x) = \frac{\lambda^{k+1} x^k \exp(-\lambda x)}{\Gamma(k+1)}, x > 0$$

where $\lambda > 0$ and k is a known constant, show that the ML estimator $\hat{\lambda}$ for λ is $(k+1)/\bar{x}$ Show that the corresponding estimator is biased but consistent and that its asymptotic distribution for large n is

$$N(\lambda, \lambda^2/[n(k+1)]).$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1986*]

(*b*) Derive the *MLE* of the mean $\frac{\alpha}{\alpha+2}$ of the beta distribution :

$$f(x) = [B(\alpha, 2)]^{-1} x^{\alpha-1}(1-x),\ 0 < x \le 1,\ \alpha > 0.$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1990*]

7. (*a*) From a sample of size n from the population of X, determine the maximum likelihood estimates of the parameters a and b of the probability density

$$f(x) = \text{Constant exp } [-(x-a)/b];\ x \ge a,\ b > 0,\ -\infty < a < \infty$$

[*Calcutta Univ. B.Sc. (Maths Hons.), 1991*]

(*b*) Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution with p.d.f.

$$f(x;\theta_1,\theta_2) = \begin{cases} \frac{1}{\theta_2} e^{-(x-\theta_1)/\theta_2}, & x \ge \theta_1,\ -\infty < \theta_1 < \infty,\ \theta_2 > 0 \\ 0, \text{ elsehwere} \end{cases}$$

Obtain the maximum likelihood estimators for θ_1 and θ_2.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1992*]

(*c*) Given a sample of n independent observations from the distribution with density : $f(x, \theta_1, \theta_2) = \theta_2^{-1} \exp\left[-(x-\theta_1)/\theta_2\right],\ \theta_1 \le x < \infty$

Find the maximum-likelihood estimator of θ_2 when θ_1 is known and the maximum likelihood estimator of θ_1 when θ_2 is known and also the joint maximum likelihood estimators of θ_1 and θ_2. Comment on the estimators you obtain.

8. (*a*) A random variable X has the probability density function :

$$f(x) = (\beta+1)x^{\beta},\ \text{for } (0 < x < 1),\ (\beta > -1).$$
$$= 0, \qquad \text{otherwise.}$$

Based on n-independent observations on X, obtain the maximum likelihood estimator of β and an unbiased estimator of $(\beta+1)/(\beta+2)$, when $\beta \ne -2$.

(*b*) A random variable X has a distribution with density function

$$f(x) = (\alpha+1)x^{\alpha},\ (0 \le x \le 1,\ \alpha > -1)$$
$$= 0, \qquad \text{otherwise}$$

and a random sample of size 8 produces the data :

$$0{\cdot}2,\ 0{\cdot}4,\ 0{\cdot}8,\ 0{\cdot}5,\ 0{\cdot}7,\ 0{\cdot}9,\ 0{\cdot}8,\ 0{\cdot}9.$$

Find the maximum likelihood estimate of the unknown parameter α, it being given that *ln* $(0{\cdot}0145152 \simeq -4{\cdot}2326$ (*ln* denotes natural logarithm).

[*Burdwan Univ. B.Sc. (Hons.), 1989*]

(*c*) Find the MLE of θ for a random sample of size n from the distribution :

$$f(x, \theta) = (\theta + 1)\, x^{\theta}, \quad 0 \le x \le 1$$
$$= 0, \qquad \text{otherwise}$$

Show that it is also sufficient statistic for θ.

Ans. $MLE\ (\hat{\theta}) = \left[- \frac{n}{\sum_{i=1}^{n} \log x_i} - 1 \right]$ ...(*)

$$T = \prod_{i=1}^{n} x_i, \text{ is sufficient estimator for } \theta$$

$\Rightarrow$ $\hat{\theta} = \left[\frac{-n}{\log (\prod_i x_i)} - 1 \right]$, being a one to one function of sufficient statistic, is also a sufficient statistic for θ.

9. (*a*) Obtain the MLE for the parameter θ in a random sample of size n from the uniform population $U[0, \theta]$.

Ans. $\hat{\theta} = x_{(n)}$, the largest sample observation.

(*b*) Show by means of an example, that MLE are not, in general unique.

Ans. See Example 15·34.

(*c*) Show that in a random sample from a distribution with p.d.f.

$$f(x, \theta) = \theta e^{-\theta x}, x \ge 0$$

$1/\bar{X}$ is the MLE for θ and has greater variance than the unbiased estimator $(n - 1)/(n\bar{X})$.

Hint. MLE $\hat{\theta} = \frac{1}{\bar{X}} = \frac{n}{T}$, $T = \sum_{i=1}^{n} X_i \Rightarrow n\bar{X} = T$

X_i, $(i = 1, 2, \ldots, n)$ are *i.i.d.* $\gamma(\theta, 1)$

$\Rightarrow$ $T = \sum_i X_i \sim \gamma(\theta, n)$

$$E\left[\frac{n-1}{n\bar{X}}\right] = E\left[\frac{n-1}{T}\right] = (n-1)\, E(1/T) = \theta.$$

$$\text{Var}\left(\frac{n-1}{n\bar{X}}\right) = \left(\frac{n-1}{n}\right)^2 \text{Var}\left(\frac{1}{\bar{X}}\right) < \text{Var}\left(\frac{1}{\bar{X}}\right) = \text{Var}\,\hat{\theta}$$

10. (*a*) Let $x_1, x_2, \ldots, x_n$ be a random sample from a population with density :

$$f(x, \theta) = \frac{1}{2} \exp\left[- | x - \theta | \right], - \infty < x < \infty.$$

Find the estimator for θ based on the method of maximum likelihood.

[Madras Univ. B.Sc., 1989]

Hint. $L = \left(\frac{1}{2}\right)^n \exp\left[- \sum_{i=1}^{n} | x_i - \theta | \right]$ is maximum, if $\sum_{i=1}^{n} |x_i - \theta|$ is minimum. $\Rightarrow$ $\hat{\theta}$ = Median of $(x_1, x_2, \ldots, x_n)$.

(*b*) Obtain the maximum likelihood estimator of θ based on a random sample of size n from the population with p.d.f.

(*i*) $f(x, \theta) = e^{-(x-\theta)}$; $\theta \le x < \infty$, $-\infty < \theta < \infty$

(*ii*) $f(x, \theta) = \theta x^{\theta-1}$; $0 < x < 1$, $0 < \theta < \infty$.

Examine in each case, whether θ is unbiased.

Hint. (*i*) L is maximum if $\sum_{i=1}^{n} (x_i - \theta)$ is minimum.

$\Rightarrow$ Each deviation $(x_i - \theta)$, $i = 1, 2, \ldots, n$ is minimum $\Rightarrow$ $\hat{\theta} = x_{(1)}$.

11. (*a*) Explain what is meant by an estimate of a population parameter. Find the maximum likelihood estimate of the parameter θ of a population having density function :

$$2(\theta - x)/\theta^2, (0 < x < \theta)$$

for a sample of unit size and examine whether the estimate so obtained is biased or not. [***Calcutta Univ. B.Sc. (Maths. Hons.), 1987***]

Ans. $\hat{\theta} = 2x$; biased.

(*b*) Obtain Maximum Likelihood Estimator of θ for the distribution :

$$f(x, \theta) = \frac{C_0 \theta^x}{x!}; x = 0, 1, 2, \ldots; \theta > 0,$$

C_0 is a constant. Also write the Maximum Likelihood Estimator of $3\theta^2 + 4\theta + 5$. (***Agra Univ. B.Sc., 1988***)

Hint. For MLE of $3\theta^2 + 4\theta + 5$, use Invariance Property of MLE (*c.f.* Theorem 15·17)

(*c*) A population has a density function given by :

$$f(x) = 2\nu \sqrt{\frac{\nu}{\pi}}\, x^2 e^{-\nu x^2}; -\infty < x < \infty$$

Find the maximum likelihood estimate for ν.

[***Calcutta Univ. B.Sc. (Maths. Hons.), 1988***]

12. (*a*) Consider a population made up of 3 different types of individuals occurring in the population with probabilities θ^2, $2\theta(1 - \theta)$ and $(1 - \theta)^2$, respectively where $0 < \theta < 1$. Let n_1, n_2 and n_3 denote the respective random sample sizes of the above three types of individuals. Determine the maximum likelihood estimator for θ. [***Rajasthan PCS, 1989***]

(*b*) Obtain the maximum likelihood estimate of θ, if the variable takes the values 1, 2, 3 and 4 with probabilities $(1 - \theta)/2$, $(1 - \theta)/2$, $\theta(1-\theta)$ and θ^2 respectively and the observed frequencies are n_1, n_2, n_3 and n_4 respectively.

13. In life-testing it is sometimes assumed that the life-time of an item is a random variable which is greater than or equal to x with probability

$$\exp\left[-\left(\frac{x}{\theta}\right)^m\right],$$

$x \ge 0$, $m > 0$ is known and $\theta > 0$ is unknown. Suppose n such items are tested and yield $X_1, X_2, \ldots, X_n$ as their times of "death".

Find the maximum likelihood estimate of θ.

14. X_1, X_2, X_3, X_4 are independent normal random variables with means $\alpha + \beta$, $\alpha - \beta$, $\alpha + 2\beta$, $\alpha - \beta$ respectively and a common variance σ^2, on the basis of one observation on each X_i ; obtain the maximum likelihood estimators of α, β and σ^2. What is the asymptotic variance of α^2 ?

[Bharatiyan Univ. M.Sc. (Maths), 1993]

15. (*a*) For the bivariate normal distribution $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ find the maximum likelihood estimators

(*i*) of σ_1^2, σ_2^2 and ρ when μ_1 and μ_2 are known,

(*ii*) of all five parameters of the distribution.

(*b*) Describe clearly the important properties to be possessed by a good estimator.

If (x_i, y_i), $(i = 1, 2, \ldots, n)$ come from a bivariate normal population with zero means, unit variances and co-efficient of correlation ρ, obtain the maximum likelihood estimator of ρ.

16. (*a*) Show that the most general continuous distribution for which the M.L.E. of a parameter θ is the sample harmonic mean is :

$$f(x, \theta) = \exp\left[\frac{1}{x}\left\{\theta \frac{\partial \psi}{\partial \theta} - \psi(\theta)\right\} - \frac{\partial \psi}{\partial \theta} + \xi(x)\right]$$

where $\psi(\theta)$ and $\xi(x)$ are arbitrary functions of θ and x respectively.

(*b*) Explain the principle of maximum likelihood estimation. Give examples to show that MLE need not be unique and also not necessarily unbiased.

Show that the most general form of the distribution for which the sample arithmetic mean $\bar{X}$ is the MLE of θ has the p.d.f.

$$f(x, \theta) = \exp\left[(x - \theta) A'(\theta) + A(\theta) + B(x)\right]$$

[Delhi Univ. B.Sc. (Stat. Hons.), 1988]

17. (*a*) Suppose that distribution of X is represented by the function :

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x\,!}; \; x = 0, 1, 2, \ldots$$

where $\lambda > 0$. Given a random sample of size n, show that the sample mean is the maximum likelihood estimate of λ. Show further that this estimate is (*i*) best unbiased, and (*ii*) consistent. *[Delhi Univ. M.A. (Eco.), 1986]*

(*b*) Consider the estimation of the Poisson parameter from a random sample.

(*i*) Work out the maximum likelihood estimator and its variance.

(*ii*) Work out the Cramer – Rao Lower bound and show that it is equal to the variance worked out in (*i*). Comment on the significance of this result.

[Delhi Univ. M.A. (Eco.), 1990]

18. X is a discrete random variable and

$$P(X = r) = (1 - p)\, p^{r-1}; \; r = 1, 2, 3, \ldots$$

Find the MLE of p based on a random sample of n observations and its variance in large samples.

Show that the variance attains the lower bound of C.R. inequality.

19. Explain the terms : (*i*) sufficient estimator, (*ii*) efficient estimator, (*iii*) Cramer-Rao lower bound to the variance of an estimator, (*iv*) maximum likelihood estimator; and describe the relations amongst these four concepts.

20. (*a*) Describe the method of moments for estimating the parameters. What are the properties of the estimates obtained by this method ?

(*b*) Let $(X_1, X_2, \ldots, X_n)$ be a random sample from the p.d.f.

$$f(x, \theta) = \theta e^{-\theta x}, \; 0 < x < \infty, \; \theta > 0;$$
$$= 0, \quad \text{elsewhere}$$

Estimate θ using the method of moments.

(***Madras Univ. B.Sc., 1988***)

21. $X_1, X_2, \ldots, X_n$ is a random sample from

$$f(x\,; a, b) = \frac{1}{b - a}; \; a < x < b$$
$$= 0, \quad \text{elsewhere}$$

Find estimates of a and b by the method of moments.

[***Gujarat Univ. B.Sc. Oct., 1993***]

22. Explain the methods of estimation-method of moments and maximum likelihood. Do these lead to the same estimates in respect of the standard deviation of a normal population ? Examine the properties of the estimates from the point of view of consistency and unbiasedness.

23. (*a*) Estimate θ in the density function

$$f(x, \theta) = (1 + \theta)\, x^{\theta}; \; 0 < x < 1$$

by the method of moments and obtain the standard error of the estimator.

(*b*) The sample values from population with p.d.f.

$$f(x) = (1 + \theta)\, x^{\theta}, \; 0 < x < 1, \; \theta > 0,$$

are given below :

0·46, 0·38, 0·61, 0·82, 0·59, 0·53, 0·72, 0·44, 0·59, 0·60

Find the estimate of θ by (*i*) method of moments and (*ii*) maximum likelihood estimation.

24. (*a*) For the distribution with probability function :

$$f(x, \theta) = \frac{e^{-\theta}\, \theta^{x}}{x\,!\,(1 - e^{-\theta})}; \; x = 1, 2, 3, \ldots$$

obtain the estimate of θ by the method of moments.

(*b*) For the following probability function :

$$f(x, p) = \binom{3}{x} \frac{p^{x}\,(1 - p)^{3-x}}{1 - (1 - p)^{3}}, \; [x = 1, 2, 3]$$

obtain the estimator of p by the method of moments, if the frequencies at $x = 1, 2$ and 3 are respectively 22, 20 and 18.

25. Let $x_1, x_2, \ldots, x_n$ be a sample from a distribution with density function :

$$f_{\theta}(x) = \theta(\theta + 1)\, x^{\theta - 1}\,(1 - x), \; 0 < x < 1, \; \theta > 0$$

Determine the estimate of θ by the method of moments.

[*Indian Civil Services, 1981*]

26. Explain the method of minimum chi-square in estimation, with a suitable example. [*Madras Univ. B.Sc., March 1989*]

27. Describe the method of moments and discuss when the estimates obtained by the method of moments are identical with those of maximum likelihood estimates.

Estimate α and β by the method of moments for the distribution :

$$f(x;\ \alpha,\ \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1}\, e^{-\beta x}\ ,\ 0 \le x < \infty.$$

[*Delhi Univ. B.Sc. (Stat. Hons.), 1987, 1983*]

28. State the conditions under which Maximum Likelihood Estimators of the parameters are identical with those given by the method of moments.

Examine if the MLEs of the parameter(s) are identical with those obtained by the method of moments in random sampling from the following distributions :

(*i*) $f(x, \theta) = \frac{1}{\theta} \cdot \exp\left(-\frac{x}{\theta}\right)\ ;\ 0 < x < \infty$

(*ii*) $f(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(x-\mu)^2/2\sigma^2\right]\ ;\ -\infty < x < \infty.$

Ans. (*i*) *MLE* $(\hat{\theta}) = \bar{x} = \hat{\theta}$ (Method of Moments)

(*ii*) MLE $(\hat{\mu}) = \bar{X} = \hat{\mu}$ (Method of Moments)

MLE $(\hat{\sigma^2}) = s^2$ (sample variance) $= \hat{\sigma^2}$ (Method of Moments).

29. Independent samples of sizes n_1 and n_2 are taken from two normal populations with equal means μ and variances respectively equal to $\lambda\sigma^2$, σ^2. Find the maximum likelihood estimator of μ based on $(n_1 + n_2)$ sample observations and show that its large sample variance is

$$Var\,(\hat{\mu}) = \sigma^2 \Big/ \left(\frac{n_1}{\lambda} + n_2\right)$$

Hence show that the unbiased estimator, $t = (n_1\bar{x}_1 + n_2\bar{x})/(n_1 + n_2)$ has efficiency, $\frac{\lambda(n_1 + n_2)^2}{(n_1\lambda + n_2)(n_1 + n_2\lambda)}$ which attains the value 1 if and only if $\lambda = 1$.

Ans. MLE $(\hat{\mu}) = \left(\frac{n_1\bar{x}_1}{\lambda} + n_2\bar{x}_2\right) \Big/ \left(\frac{n_1}{\lambda} + n_2\right)$

OBJECTIVE TYPE QUESTIONS

1. Comment on the following statements :

(*i*) In case of the Poisson distribution with parameter λ, $\bar{x}$ is sufficient for λ.

(*ii*) If $(X_1, X_2, \ldots, X_n)$ be a sample of independent observations from the

uniform distribution on $(\theta, \theta + 1)$, then the maximum likelihood estimator of θ is unique.

(*iii*) A maximum likelihood estimator is always unbiased.

(*iv*) Unbiased estimator is necessarily consistent

(*v*) A consistent estimator is also unbiased.

(*vi*) An unbiased estimator whose variance tends to zero as sample size increases is consistent.

(*vii*) If t is a sufficient statistic for θ then $f(t)$ is a sufficient statistic for $f(\theta)$.

(*viii*) If t_1 and t_2 are two independent estimators of θ, then $t_1 + t_2$ is less efficient then both t_1 and t_2.

(*ix*) If T is consistent estimator of a parameter θ, then $aT + b$ is a consistent estimator of $a\theta + b$, where a and b are constants.

(*x*) If x is the number of successes in n independent trials with a constant probability p of success in each trial, then x/n is a consistent estimator of p.

II. Fill in the blanks :

(*i*) In a random sample of size n from a population with mean μ, the sample mean ($\bar{x}$) is ... estimate of ...

(*ii*) The sample median is ... estimate for the mean of normal population.

(*iii*) An estimator $\hat{\theta}$ of a parameter θ is said to be unbiased if ...

(*iv*) The variance s^2 of a sample of size n is a ... estimator of population variance σ^2.

(*v*) If a sufficient estimator exists, it is a function of the ... estimator.

(*vi*) ... estimate may not be unique.

III. (*a*) Give example of a statistic t which is unbiased for a parameter θ but t^2 is not unbiased for θ^2.

(*b*) Give example of an M.L. estimator which is not unbiased.

IV. What is the relationship between a sufficient estimator and a maximum likelihood estimator ?

V. (*i*) If $\bar{x}$ is an unbiased estimator for the population mean μ, state which of the following are unbiased estimators for μ^2 :

(*a*) $\bar{x}^2$, (*b*) $\bar{x}^2 - \frac{\sigma^2}{n}$ (σ^2 is known/unknown).

(*ii*) If t is the maximum likelihood estimator for θ, state the condition under which $f(t)$ will be the maximum likelihood estimator for $f(\theta)$.

(*iii*) Write down the condition for the Cramer-Rao lower bound for the variance of an unbiased estimator to be attained.

(*iv*) Write down the general form of the distribution admitting sufficient statistic.

VI. A random variable X takes the values 1, 2, 3 and 4, each with probability $\frac{1}{4}$. A random sample of three values of x is taken, $\bar{x}$ is the mean and m is the median of this sample. Show that both $\bar{x}$ and m are unbiased estimators